How Not To Drown in Data

Review

How Not To Drown in Data: A Guide for Biomaterial Engineers

Aliaksei S. Vasilevich,[1] Aurélie Carlier,[1] Jan de Boer,[1] and Shantanu Singh[2,*]

High-throughput assays that produce hundreds of measurements per sample are powerful tools for quantifying cell–material interactions. With advances in automation and miniaturization in material fabrication, hundreds of biomaterial samples can be rapidly produced, which can then be characterized using these assays. However, the resulting deluge of data can be overwhelming. To the rescue are computational methods that are well suited to these problems. Machine learning techniques provide a vast array of tools to make predictions about cell–material interactions and to find patterns in cellular responses. Computational simulations allow researchers to pose and test hypotheses and perform experiments *in silico*. This review describes approaches from these two domains that can be brought to bear on the problem of analyzing biomaterial screening data.

Deciphering Cell–Material Interactions

Tissue microenvironments respond to and interact with various biological, chemical, and physical cues. Understanding how materials modulate this environment is essential from a clinical standpoint because the compatibility and functionality of medical implants are influenced by this interaction [1–3]. Traditionally, the material properties of implants that were primarily considered were **biofunctionality** (see Glossary) and **biocompatibility** [4]. However, there is more to consider beyond these two attributes. Materials can affect various aspects of cell behavior, many of which can have crucial physiological effects [5]. To name a few: surface topography, as well as the chemical composition of the material, can influence bone-bonding [6], substrate stiffness can modulate stem cell differentiation [7], and substrate stress relaxation is observed to control cell spreading [8]. The large number of material properties, which continue to increase as new discoveries are made, makes the expanding properties space far too vast [2,9] to perform comprehensive experimental screening across all salient material attributes. Therefore, we need methods that, given a biomaterial as input, can predict the biological effect of materials on cells and tissues, or, given the biological response, can create signatures of materials to find relationships between them.

Creating such predictive models poses several challenges: to build predictive models that relate the biomaterial properties (input) to the cell response (output), both the input and output need to be appropriately described (i.e., parameterized). How do we parameterize biomaterial properties? How do we parameterize biological properties (of cells and tissues)? Given the high dimensionality of these data, which corresponds to the number of measurements, what are appropriate models to consider? More broadly, given that the datasets from these experiments are large and complex, how do we avoid drowning in the data? We believe that recent advances in computing and data

Trends

Biomaterials can be rapidly produced at the rate of hundreds of unique samples in a day owing to advances in automation and miniaturization.

Cell–material interactions can be quantified across hundreds of parameters using high-throughput gene expression and imaging assays at relatively low cost.

The biomaterials field is thus faced with a deluge of data.

Machine learning algorithms and mathematical modeling are powerful tools that are well suited for working with these data. There are already a few examples of their use in biomaterials research.

High-performance cloud computing is now becoming much more affordable and requires no hardware investment, making it easy for labs to employ these computational methods.

[1] Laboratory for Cell Biology-Inspired Tissue Engineering, MERLN Institute for Technology-Inspired Regenerative Medicine, Maastricht University, Maastricht, The Netherlands
[2] Imaging Platform, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA

*Correspondence:
shsingh@broadinstitute.org (S. Singh).

http://dx.doi.org/10.1016/j.tibtech.2017.05.007

analysis can be brought to bear on these problems. This review addresses these questions and provides examples from material science and related fields.

Parameterization of Cellular Responses

Defining and measuring parameters – parameterization – of cell responses is important for understanding the mechanism of cell–material interactions. A parameter can be defined as a set of physical properties whose values determine the characteristics of an entity. Applied to cell responses, possible parameters include cell shape, cell structure, gene expression, and cell differentiation potential, among several others. In this review we focus on methods that can be performed with high throughput as well as produce multiple distinct measurements at the same time, resulting in a rich description of the cellular state at large scale. Gene expression analysis and high-content imaging, performed with high throughput, are two techniques that provide this level of detail and scale.

Gene Expression Analysis

A classical output of parameterization of cellular responses is gene expression analysis to measure changes in cellular mRNA levels induced by the material. Biologists have extensively used microarray gene expression analysis for a couple of decades, and analysis pipelines are well-established. Recent advances enable expression profiling at a lower cost, making it feasible to assay large sample sizes of different biomaterials. L1000 profiling [10] measures the expression levels of 978 'landmark' genes with high throughput (using multi-well plates); the levels of an additional 21 290 genes are then inferred using a linear model. These technologies are increasingly used to profile genetic and chemical perturbations [11–13]; a list of software for gene expression analysis is given in the supplemental information online.

Gene expression analysis has been previously used in biomaterial research. For example, Groen and colleagues explored the transcriptomic landscape induced by 23 different materials related to bone regeneration, identifying genes that are responsive to biophysical and chemical cues [14]. Gene expression analysis has been used to identify hub genes which, upon activation, trigger ectopic bone formation on calcium scaffolds *in vivo* [15]. In another study, MG63 cells were cultured on a variety of osteoinductive and non-osteoinductive biomaterials, and gene pathway and network exploration tools were employed to construct a signaling map of osteogenesis on biomaterials [16].

High-Content Imaging

Another method to parameterize the cellular state at relatively low cost is high-content imaging using fluorescent staining and immunofluorescence, performed at high throughput [13,17]. Proteins of interest or subcellular constituents can be labeled with unique fluorescent markers. To distinguish single cells, in addition to particular proteins of interest, the nucleus and cytoplasm can be labeled using stains such as diamidino phenylindole (DAPI) or phalloidin, respectively. Similarly, other organelles can be identified using appropriate stains. For instance, in the Cell Painting assay [18], six markers have been used to identify eight different subcellular constituents, and nearly 1500 measurements are extracted from each cell in the experiment (e.g., shapes, textures, intensity, neighborhood information, and the correlation between channels). Recent progress in multiplexed ion-beam imaging can help to increase the number of markers that can be imaged, enabling the measurement of up to nearly 100 targets in a single experiment [19]. Software tools for high-content imaging are discussed in [20] and listed in the supplemental information online.

Challenges

Applying these approaches can be challenging in biomaterials research, depending on the materials and cellular models being used. For example, autofluorescent and non-transparent

Glossary

Biocompatibility: the property of a material of not being toxic or having injurious effects on biological function.

Biofunctionality: the ability of a material to sufficiently replace tissue function. For example, in the case of hip implants, the material should have the strength to replace the bone tissue.

Mathematical model: an abstract representation of a complex system in terms of equations or rules, and a description of the region (spatial and/or temporal) in which they are valid. For example, Newton's second law accurately describes the trajectory of a falling apple with a second-order equation, but is not valid to describe the movement of electrons around the nucleus of an atom because the theory does not hold at this (subatomic) scale.

Molecular dynamics (MD): algorithms that start from an initial set of molecular positions; next, the forces acting on each molecule or atom are calculated with a heavily parameterized classical mechanics model, which are in turn used to update the molecular positions. The algorithm repeats the previous step until a suitable energy potential function is minimized given some constraints. MD approaches are frequently used to study the motions of proteins and nucleic acids, as well as their interaction with other molecules (e.g., ligand docking). MD simulations can also be used to predict protein structures and to investigate the properties of (biological) membranes or molecular assemblies.

Monte Carlo (MC) simulations: computational algorithms that rely on repeated random sampling to obtain numerical results. For example, a computer code that draws a large number of random variables from the interval [0,1], and that assigns values less than or equal to 0.5 as heads, and greater than 0.5 as tails, is an MC simulation of repeatedly tossing a coin. Notably, MC simulations investigate the same type of problems as MD simulations but with a different method and results. MD looks at the temporal evolution of the system given known physical parameters, while MC generates random possible scenarios for the molecular assemblies and then biomaterials preclude the use of standard microscopy imaging techniques. In addition, cells can be grown in spheroid cultures which have 3D structure [21]. Further, microarray analysis requires the difficult task of extracting cell material from 3D constructs.

Parameterization of Biomaterials

To parameterize biomaterials, two fundamental properties should be assessed: their chemical composition and their spatial organization, properties that underlie the majority of biomaterial parameters. For example, the chemistry of materials determines the adhesivity, hydrophobicity, mechanical properties, and degradability of the biomaterials, whereas the spatial organization may influence the hardness and electrical properties of the biomaterials. Canonical examples of the spatial organization effect are diamond and graphite: both comprise the same building block, carbon atoms, but the different spatial organizations (tetrahedral and hexagonal arrangements for diamond and graphite, respectively) result in a very hard insulator (diamond) and a soft conductor (graphite). A similar example in the biomaterials field is that of calcium phosphate ceramics: a different spatial organization of molecules in calcium phosphate ceramics can create biomaterials with different osteoinductive properties [22], related to an altered topography at the macro-, micro-, and nano-levels. Finally, parameters such as stiffness are controlled by both the chemistry and spatial organization of the materials. Recent reviews [2,5] discuss an expansive list of biomaterial properties.

Some chemical and spatial attributes of biomaterials can be represented by their design parameters – for example, the elementary chemical composition of the polymers [23,24] or spatial organization [25]. Others should be physically measured; reviewing routine techniques for measuring the physicochemical properties of materials is beyond the scope of this manuscript and is surveyed in [26].

Mathematical Representation

Once the parameters denoting the cellular state have been selected, each biological sample can be represented by a feature vector, which is simply the set of parameter values for that sample. The size of this set is the dimensionality of the feature space. The set of all the feature vectors, corresponding to each of the samples in an experiment, is the data matrix, or, more specifically, the output data matrix, given that the cellular response is typically the output that is predicted or modeled. Similarly, given the biomaterial parameterization, the input data matrix can be constructed: the set of biomaterial parameter values is the feature vector denoting that material sample, and the set of all these feature vectors from an experiment constitute the input data matrix.

Machine Learning: Finding Patterns and Making Predictions

What is Machine Learning?

Machine learning is among the fastest-growing technical fields and is being increasingly adopted across several domains of science, technology, and commerce [27]. The underlying techniques in machine learning span a wide range, from traditional statistical approaches (e.g., linear regression) with simple input–output relationships to modern **neural networks** [28] that capture a hierarchy of nonlinear relationships. However, the general concept is straightforward: machine learning is a field that studies computer programs that improve automatically through experience [29]. In this sense, computers can be trained to perform tasks, similarly to humans, making the field tightly linked to the study of artificial intelligence [30]. We instead take a more procedural view of learning in the case of computers. We consider the formal, but simple, definition: 'a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.' [29].

calculates the energy of such a scenario.

Neural networks: a class of machine learning algorithms that learn the mapping from input to output by estimating the weights of the connections between several intermediate 'nodes', each of which are models similar to regression models. Nodes are arranged in layers, and each node in one layer is connected to some or all nodes in the next layer (layers may also be skipped). The first layer is the input, the last layer is the output. Information propagates from input to output, and is transformed at each layer by the nodes based on their 'weights'. After each training pass, the difference between the predicted output and the true output is used to update the weights of the nodes. In 'deep learning', neural network models can have several tens of layers and hundreds of nodes in each layer. Modern computer architectures, large amounts of data, and progress in algorithms that update the weights have led to the success of this approach.

Random forest (RF): a classifier that creates several 'decision trees' to predict the class of a datapoint. Each decision tree comprises a set of rules that splits the feature space into compartments; each compartment has a class associated with it. Not all features are used in each tree – features are randomly selected and a decision tree is estimated from each set, given the data.

Support vector machines (SVM): a classifier that finds a 'separating hyperplane' with 'maximum margin' that separates two classes of datapoints. The separating hyperplane is a line in a high-dimensional space that splits the space into two parts, corresponding to the two classes. The margin is the distance of the closest point in either class to this hyperplane. The greater the margin, the better the classifier is likely to be. The 'kernel trick' is central to the performance of SVMs – it is the nonlinear mapping of data into a high-dimensional space where it is easier to find a separating hyperplane with a large margin.

How is this relevant to biomaterials research? Consider an example of designing a new biomaterial that induces a particular cellular phenotype determined by three parameters of interest (the output): cell proliferation, extracellular matrix (ECM) deposition, and expression level of a protein of interest – all of which can be measured using microcopy imaging. The design of the biomaterial (the input) can be represented by different parameters related its chemistry (C) and structure (S): elasticity (C + S), crystallinity (S), grain size (S), and phase composition (C). How can machine learning be used to create the desired biomaterial?

First, the problem can be simplified to consider only one output variable (e.g., cell proliferation) and only one input variable (e.g., grain size). In this example, several (n = 30) experiments have been conducted: biomaterials of different grain sizes are created, cells are grown on them, and cell proliferation is measured for each. This data can be visualized as a scatter plot with cell proliferation on the y axis and grain size on the x axis, and the goal is to find a relationship between proliferation and grain size. One solution is linear regression, discussed in Box 1, which estimates a 'best-fit' line that captures the relationship between the input and the output. This approach can be extended to predict all four output variables using multiple regression, also discussed in Box 1.

How does this example fit into the definition of machine learning presented above? The computer program (linear regression) learns from experience (the 30 datapoints) to perform a task (predict the relationship between cell proliferation and grain size), evaluated based on a performance measure (sum of squared error, SSE, described in Box 1). Given more experience, the program typically performs better at the task. Importantly, it is not only the repetition, but also the variation, of experience that determine how well the program will learn to perform the task.

This regression example shows how a relatively simple machine learning approach and experimental data can guide the creation of new biomaterials. However, the example required several simplifying assumptions. First, the variables considered here were continuous-value

Box 1. Regression

Linear regression is used to estimate a best-fit line (shown in blue in Figure IA; each dot corresponds to an experiment): it minimizes the difference between the true and predicted values of Y, summed across all the datapoints. The differences are squared before summing, hence the term 'sum of squared error' (SSE); the best-fit line for the data minimizes the SSE. The goodness-of-fit measures of how well the regression model fits the set of observations and is given by the coefficient of determination, denoted R^2. It ranges from 0 to 1 and indicates the proportion of the variance in Y that is predictable from X. It is computed as the square of the correlation between predicted values of Y and the actual value of Y.

This line estimates the grain size that induces the desired rate [red broken lines; a target rate of 29 arbitrary units (a.u.) is assumed]. A grain size of 6 a.u. is the most likely to induce the target rate. The prediction is not perfect: the grey region denotes the error (the 95% confidence region) in prediction.

Multiple linear regression, an extension of the simple linear regression described above, predicts values for all four biomaterial design parameters that induce the desired cell proliferation rate. The cellular response for each biomaterial sample is denoted by a 1D feature vector (cell proliferation). The set of all 30 feature vectors – corresponding to the 30 biomaterial samples – is the output data matrix. Similarly, a biomaterial sample is denoted by a 4D feature (elasticity, crystallinity, grain size, and phase composition) and the set of all 30 feature vectors is the input data matrix. A multiple regression model that predicts cell proliferation using all four biomaterial design features can be estimated using these data (Figure IB). In this case, 'estimating the model' refers to finding the values of coefficients a, b_1, b_2, b_3, b_4.

This model can be used to predict the design parameter values X_1, X_2, X_3, X_4 (the independent variables) that induce the desired cell proliferation rate Y, the dependent variable. This is achieved by assigning, the desired value of Y in the equation, and using the estimated values of coefficients a, b_1, b_2, b_3, and b_4 to find the values of X_1, X_2, X_3, X_4 that would satisfy the equation. A range of values for the independent variables would satisfy this requirement. Other external criteria – such as ease of fabrication – can be used to further filter down this solution space.

The multiple regression formulation can be extended to solve the original problem posed in the main text – that of finding the combination of design parameter values that induces the desired cellular response as measured by all three parameters: cell proliferation, ECM deposition, and expression level of the protein of interest. For each of the cellular response parameters (denoted by Y_1, Y_2, Y_3), it is possible to (i) estimate a multiple regression model, (ii) find the range of solutions H_1, H_2, H_3, H_4 for each, and finally (iii) find the intersection of H_1, H_2, H_3, H_4, thereby providing a range of design parameter values that would induce the desired cellular response. If there are more dependent variables (the output) than independent variables (the input), then this intersection could be void, in which case it might be appropriate to instead find a cellular response phenotype that is closest to the target that has a non-empty solution space. Description of these techniques is beyond the scope of this paper, the interested reader can refer to convex optimization methods [73] for details, and to [74] for texts on regression methods.

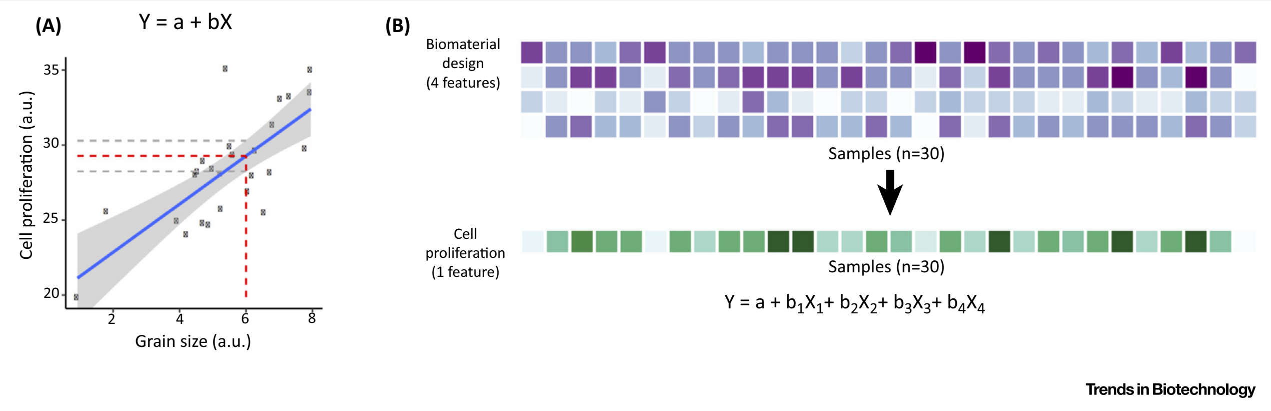


Figure I. Regression. (A) Simple linear regression. The dependent variable (*Y*) is predicted using the independent variable (*X*). The output (cell proliferation) is the independent variable and the input (grain size) is the dependent variable, but it could be the other way around depending on the problem at hand. We estimate a 'linear model' – a straight line – to fit the data; the result of the estimation is the value of *a* (*y* intercept) and *b* (slope). The blue line represents the estimated model using all 30 points. The coefficient of determination (R^2) of this model fit is 0.57. (B) Multivariate regression. There are four 'independent' variables – elasticity, crystallinity, grain size, and phase composition, denoted by X_1, X_2, X_3, X_4 comprising the input, and one 'dependent' variable – cell proliferation – denoted by *Y*, comprising the output. Multivariate linear regression is used to find the values of the 'coefficients' *a*, b_1, b_2, b_3, b_4 that will best estimate the dependent variable given all the independent variables.

input and output variables, but they could instead be discrete (e.g., whether levels of a protein are high, medium, or low compared to some reference) or categorical (e.g., polymerization mechanism); this problem can be posed a 'classification' problem. Second, the relationship between input and output was assumed to be linear, which is often not the case (the 'goodness-of-fit' may indicate this). Third, in this case the goal is to predict a parameter, and the experiments provide data from direct observations of this parameter (termed 'supervised learning'), but there are situations where the goal is instead to find patterns in the data, or the parameter being predicted cannot be observed directly (termed 'unsupervised learning'). We discuss methods that address these cases in Boxes 2 and 3. Some challenges faced when using machine learning techniques are discussed in Box 4.

We note that unsupervised learning methods, including clustering (Box 3) and dimensionality reduction (Box 4) allow the data to be explored in a relatively unbiased manner. Used with caution (Box 4), this is an important application of machine learning – which is to discover relationships in the data without the constraints of preconceived biases that domain knowledge may introduce.

Software and Tools

Software tools for machine learning abound; most require some knowledge of programming. Fortunately, the dominant programming languages in machine learning, such as R and Python, have excellent introductory texts [31–33] as well as online courses such as Codecademy, Udacity, and Software Carpentry. In Python, scikit-learn [34] is a comprehensive library of functions for machine learning. Similarly, the caret package [35] in R provides a standardized interface to a large number of machine learning libraries. Visual tools such as Weka [36] and KNIME [37] enable users to perform data analysis without the need for much programming experience.

Examples of Machine Learning in Biomaterials Research

There are several examples of the application of machine learning in biomaterials research. For example, a regression technique was used to predict bacterial attachment to hundreds of

Box 2. Classification

What if the goal is to find edge-case biomaterials that would induce either a very high or very low cell proliferation rate? Extending the example described in the main text, in Figure IA cell proliferation values have been replaced with three 'classes' or 'levels' corresponding to high (green), low (orange), or medium (grey). These data can teach a classifier that can predict cell proliferation level given the biomaterial design parameters.

Among the simplest classifiers is *k*-nearest neighbors (k-NN). As with the regression example, each biomaterial sample is represented here by its 4D feature. It is possible to imagine a 4D space in which these 30 biomaterial samples are datapoints. Each datapoint is either green, orange, or grey. Given a new biomaterial, which class it will fall into? It is possible to conceptually 'plot' a datapoint in the 4D space corresponding to this new biomaterial and assign it a color based on a simple rule: find its *k*-nearest datapoints (say, $k = 5$) and assign it the color to which the majority of these five datapoints belong. For this it is necessary to specify a measure of distance. 'Euclidean' distance is frequently used: extending the notion of distance in physical space, the differences in each dimension are squared and summed across all dimensions, and then the square root of the sum is computed. This concept is illustrated in Figure IB in a 2D space.

Thus, using a relatively simple concept of distance in a high-dimensional space (Euclidean distance), the new biomaterial can be classified into one of the three classes. This example also addresses the case where the relationship between input and output may be nonlinear: the k-NN classifier does not require a linear relationship. k-NN classification is surprisingly effective in many scenarios and is often used to establish baseline performance when comparing classifiers. There are several classifiers proposed and used in practice. Excellent texts are available on this topic [29,39]. Some of the salient classifiers – SVM, **random forests** (RFs) [39], and neural networks – are defined in the Glossary.

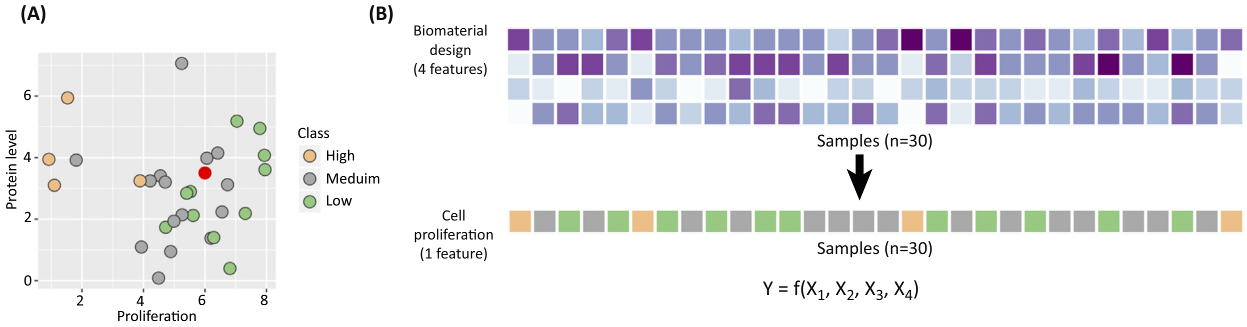
polymeric materials in a high-throughput microarray format [38]. **Support vector machines** (SVMs) [39] have been used to predict metal organic framework materials with enhanced CO_2 adsorption, by training the classifier on data comprising several tens of thousands of materials with known adsorption rates [40]. Biomaterials that support the adhesion of human embryonic stem cell embryoid bodies (EBs) were identified by training a neural network classifier on a microarray library of hundreds of polymers for which EB adhesion was measured [41]. Examining the global response of bone marrow-derived human mesenchymal stem cells allowed investigation of how strontium-doped biomaterials can improve clinical outcomes [42]: the global response – characterized using gene expression microarrays and Raman

Box 3. Clustering

The examples discussed in Boxes 1 and 2 addressed predicting outputs, but what if the goal is instead to find novel patterns in the data? For instance, given the 30 biomaterial samples, the samples can be grouped based on the cellular responses they induce. The formulation is as follows. The cellular response for each biomaterial sample is denoted by a 3D feature vector, comprising cell proliferation, ECM deposition, and expression level of the protein of interest. The set of all 30 feature vectors – corresponding to the 30 biomaterial samples – constitutes the output data matrix (Figure IA). Each feature vector in this matrix can be conceptually plotted in a 3D space. The problem of finding groups of biomaterials inducing a similar cellular response can be viewed as finding clusters of points in this space that are close together.

k-Means is a clustering algorithm that solves this problem (Figure IB). First, the number of clusters (*k*) to group the points into is specified; the algorithm then randomly picks *k* points as initial centers of the *k* clusters in the 3D space, assigns each point to the cluster that it is closest (using Euclidean distance) to the center of, and finally updates the centers of each cluster to be the 3D centroid of the datapoints assigned to the cluster. This process is repeated until the cluster centers do not change significantly from one iteration to the next. The final assignments of each datapoint to a cluster provide a solution to the problem.

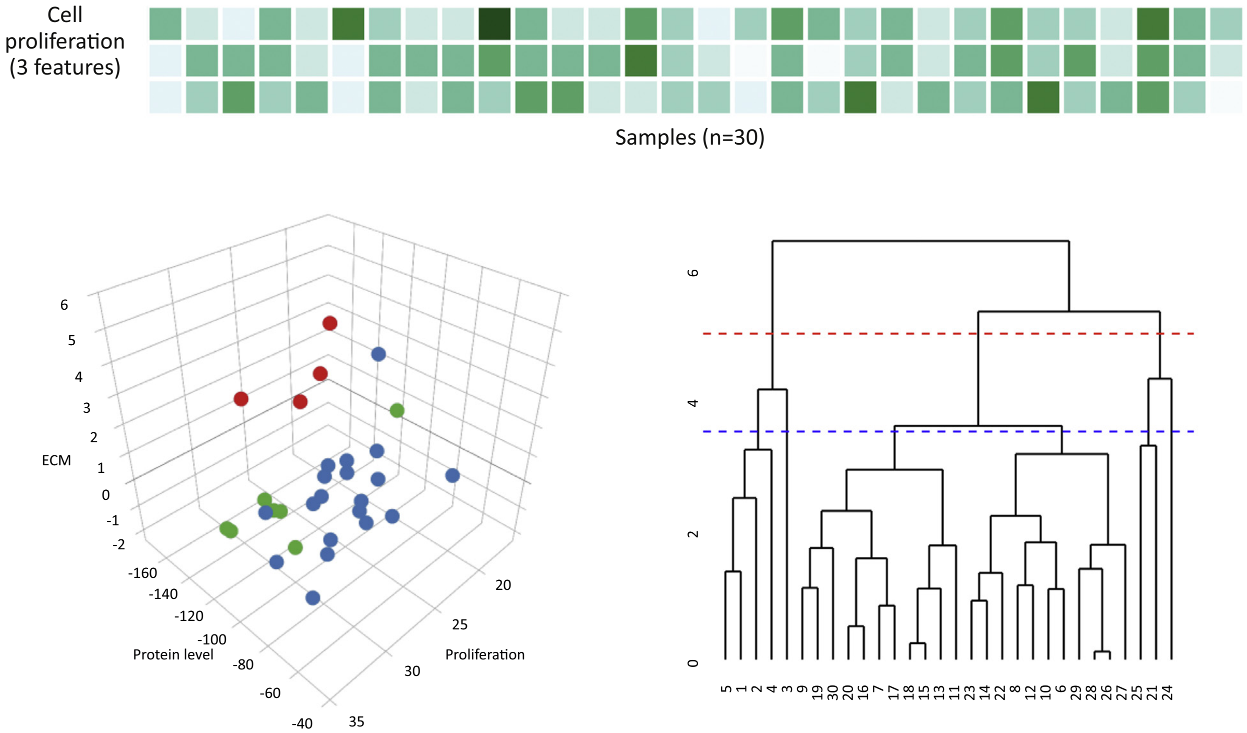
Another commonly used clustering technique is hierarchical clustering (Figure IC). One flavor of hierarchical clustering proceeds as follows: assign all the 30 datapoints to a single cluster, partition the cluster into two least-similar clusters (measured using Euclidean distance), then do the same for each cluster recursively until there is one cluster for each datapoint. This results in a hierarchical partitioning of the data, formally termed a tree. Clusters can be obtained by cutting this tree at a particular height. There is also a 'bottom-up' version of this algorithm – start with two most similar datapoints, merge, and work upwards – as well as different ways of finding similarity between clusters when splitting or merging. Several clustering techniques have been proposed and are used in practice ([75,76] for a comprehensive overview).



Trends in Biotechnology

Figure I. Classification. (A) A classifier *f* is trained to predict the level of the cell proliferation rate *Y*, which can be high (green), medium (grey), or low (orange), given the values of X_1, X_2, X_3, X_4 (the biomaterial design features). (B) k-NN classification. The class of the new datapoint (marked in red) is estimated as the majority class of its *k*-nearest neighbors. If *k* = 5, there are three 'medium' and two 'low' points in the neighborhood, and thus the new datapoint would be assigned the class 'medium'.

spectroscopy mapping – was analyzed using machine learning techniques, including feature selection and *k*-means clustering. By analyzing the clustering pattern of gene expression data, it was observed that the geometry of the scaffolds had a larger effect on stem cell function than did their chemical composition [43]. Reimer and colleagues [44] trained a classifier and



Trends in Biotechnology

Figure I. Clustering. (A) Cell response feature matrix. (B) *k*-Means clustering is performed to cluster the cell response features into *k* = 3 clusters, each of which can be considered as a phenotypic group. Each column of the feature matrix is plotted as a point. Colors indicate cluster assignments calculated using *k*-means. (C) Hierarchical clustering clusters the 30 datapoints into a 'dendrogram'. The sample number of each biomaterial is shown on the *x* axis. The dendrogram or 'tree' can be cut at any height to partition the data into clusters. Cutting it at a height of 5 (red broken line) results in two clusters and a height of 3.5 (blue broken line) results in four clusters. Abbreviation: ECM, extracellular matrix.

Box 4. The Curse of Dimensionality, Overfitting, Feature Selection, and Model Validation

High-content imaging and gene expression produce cellular response features of hundreds of dimensions. Similarly, accounting for a larger set of design parameters can produce tens of biomaterial design features. While this process results in a rich representation, it increases the complexity of analysis, commonly referred to as the 'curse of dimensionality'. This affects several techniques, including multiple regression: as the number of independent variables increase, it becomes more likely that the estimated model will overfit the data – that is, the model may perfectly fit the data that it was trained on, but may very poorly fit unseen data. Consider the example that used three design features to predict cell proliferation. With only four datapoints to learn from, in other words one more than the number of design features, the regression model would perfectly fit those four points. However, this model would perform very poorly if applied to predict the response for the remaining 26 datapoints because its degrees of freedom allowed it to overfit to the few examples.

The curse may be broken in a few ways. First, feature selection and dimensionality-reduction techniques can reduce the number of features or transform the feature into a different lower-dimensional feature space. Recognizing that highly correlated features capture similar information, a simple feature-selection approach is to reduce the number of features such that no two features have a Pearson correlation greater than a particular threshold. An alternative or addition to feature selection, dimensionality reduction can be performed, for instance, by finding linear combinations of the features that have the highest variance and are perpendicular to each other (termed principal component analysis, PCA [77]). Reviews on feature selection [78] and dimensionality reduction [79] discuss these techniques in detail.

A second way to break the curse is to use appropriate validation to test for overfitting and find the model that generalizes best across subsets of the data. In cross-validation, the dataset is partitioned into equal-sized subsets (typically, 5 or 10); one subset is held out, the model is trained on the rest, and the performance of the model is calculated on the held-out set. This process is repeated for each of the subsets, and the average performance across all iterations is used to evaluate the model. Other model validation techniques are discussed in [39].

analyzed the resulting model to identify topography parameters that maintain proliferation and Oct4 expression.

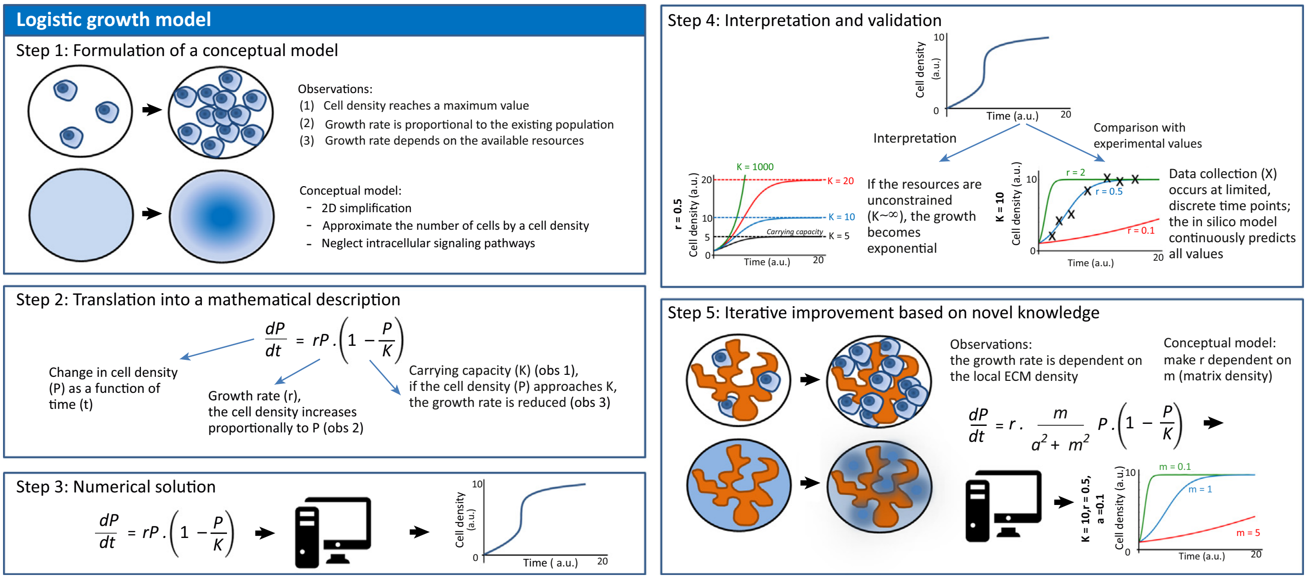
Computational Simulations: Modeling Cell–Material Interactions

Hypothesis-Driven Modeling

Computer simulations are becoming standard tools to gain insights into the complex interactions that occur in cell–biomaterial interactions, to assess new biological mechanisms, and generate novel hypotheses [45,46]. The construction of a hypothesis-driven **mathematical model** (Figure 1) first requires the formulation of a conceptual model, in other words to establish which components of the system of interest are involved and how they might interact with each other. For example, should the system be modeled in 3D, or is a 2D approximation sufficient, and which properties of the biomaterial should be captured, etc? Unlike data-driven approaches (such as machine learning), hypothesis-driven models require some *a priori* knowledge about the behavior of the system. Importantly, the model will always be a simplified version of the real system. Selecting which simplifications to make requires detailed knowledge of the system under study as well as mathematical skills, and thus a close collaboration between modelers, material scientists, and biologists can be very fruitful when they learn to speak each other's language and work in a tight feedback loop of *in silico* prediction and *in vitro* or *in vivo* validation.

In a second step, the ideas of the conceptual framework are translated into a mathematical form. For example, in Figure 1, the cell density P is, as a first simplification, modeled using two parameters: the growth rate r and the carrying capacity K.

In a final step, the mathematical equations are converted into computer code and solved numerically. Hypothesis-driven models are typically manipulated in a similar way as physical experiments: particular model parameters are perturbed and the resulting outcome is observed. However, hypothesis-driven models can avoid some experimental difficulties such as the following. (i) Models that include some stochasticity, or randomness, can be re-run


Logistic growth model
Step 1: Formulation of a conceptual model
Observations:
(1) Cell density reaches a maximum value
(2) Growth rate is proportional to the existing population
(3) Growth rate depends on the available resources
Conceptual model:
- 2D simplification
- Approximate the number of cells by a cell density
- Neglect intracellular signaling pathways
Step 2: Translation into a mathematical description
$\frac{dP}{dt} = rP.\left(1 - \frac{P}{K}\right)$
Change in cell density (P) as a function of time (t)
Growth rate (r), the cell density increases proportionally to P (obs 2)
Carrying capacity (K) (obs 1), if the cell density (P) approaches K, the growth rate is reduced (obs 3)
Step 3: Numerical solution
$\frac{dP}{dt} = rP.\left(1 - \frac{P}{K}\right)$
Cell density (a.u.)
10
0
Time (a.u.)
20
Step 4: Interpretation and validation
Cell density (a.u.)
10
0
Time (a.u.)
20
Interpretation
Comparison with experimental values
r = 0.5
Cell density (a.u.)
K = 1000
K = 20
K = 10
Carrying capacity
K = 5
20
10
5
0
Time (a.u.)
20
If the resources are unconstrained (K~∞), the growth becomes exponential
K = 10
Cell density (a.u.)
r = 2
r = 0.5
r = 0.1
10
0
Time (a.u.)
20
Data collection (X) occurs at limited, discrete time points; the in silico model continuously predicts all values
Step 5: Iterative improvement based on novel knowledge
Observations: the growth rate is dependent on the local ECM density
Conceptual model: make r dependent on m (matrix density)
$\frac{dP}{dt} = r.\frac{m}{a^2+m^2}P.\left(1 - \frac{P}{K}\right)$
K = 10,r = 0.5, a =0.1
Cell density (a.u.)
m = 0.1
m = 1
m = 5
10
0
Time (a.u.)
20

Trends in Biotechnology

Figure 1. Step-by-Step Illustration of the Construction of a Hypothesis-Driven Mathematical Model. (1) Formulation of a conceptual model where the key components and processes are defined. (2) Translation of the conceptual model into a mathematical form, exemplified in this illustration by a deterministic logistic growth function. (3) Computational solvers are used to calculate the numerical solution of the proposed equation. (4) Interpretation of the obtained numerical solution by testing various parameter values and by comparing the results to experimental values. (5) After acquiring additional biological knowledge or finding a significant difference between the model and the experimental data (Step 4), the model is iteratively improved.

multiple times from exactly the same starting condition to obtain a measure of uncertainty. (ii) The data can be non-destructively recorded at a higher frequency and spatial resolution than in an experimental setting, limited only by the temporal and spatial resolution of the numerical simulation. This allows the capture of all interesting dynamics versus the static and user-defined timing of data collection, as is schematically shown in Step 4 of Figure 1. (iii) All variables can be manipulated independently, including those that cannot be modified experimentally, at any magnitude. In Step 4 of Figure 1, K is varied with a constant r, and vice versa, but both variables can also be varied at the same time. Note that the carrying capacity K is typically an intrinsic property of the cell population, which is impossible to vary experimentally. The *in silico* model can, however, explore several values of K and the influence thereof on the cell density. Importantly, although valid model parameters (determined by physical laws) are required for meaningful predictions, running a model with 'unrealistic' values might yield interesting observations (e.g., which force distribution would a bone mechanoregulatory model predict for an 'infinite' stiff scaffold – stiffer than all scaffolds that have been produced to date?). Valid parameter ranges are also not necessarily known *a priori* but can be explored using a computational model. Several excellent reviews have more detailed information on hypothesis-driven models [47–49].

Software and Tools

Several software tools exist to convert mathematical equations into computer code, numerically solve them, and assist in the analysis of the results through dedicated visualization options. The tools can be divided into two broad categories: computer languages and comprehensive platforms. FreeFEM++ [50], SBML [51], and CellML [52] are languages specialized for modeling. MATLAB [53] is a commercial software and language geared towards numerical analysis and has several libraries for mathematical modeling. Open-source packages in Python also provide a similar functionality through libraries such as PySCeS, PyDSTool, and PySb [54]. C++, which requires significantly more programming experience, has numerical libraries that can enable building models as well as a simulation library tailored for biological models. For researchers with little background in numerical solvers, comprehensive platforms with intuitive user interfaces are available where the mathematics takes place 'behind the scenes'. Examples are VCell [55], CompuCell3D [56], Morpheus [57], and FEBio [58].

Examples of Hypothesis-Driven Modeling in Biomaterials Research

At the atomistic scale, modeling approaches such as **molecular dynamics** (MD) and **Monte Carlo (MC) simulations** use a 'classical mechanics treatment' with atoms and molecules as the basic modeling units. These approaches are for example directed at understanding the properties of (biological) membranes [59], the adsorption of biomolecules onto biomaterials [60], and the assembly and properties of supramolecular materials [61]. At the continuum scale, other modeling approaches are available, such as finite element analysis and computational fluid dynamical techniques, that are primarily applied to the optimization of scaffold design parameters related to mechanical and mass-transport properties [62]. For example, Byrne and colleagues [63] modeled the bone-scaffold system and optimized three design parameters – scaffold porosity, Young's modulus, and dissolution rate – to estimate the defect-specific loading requirements for the implant. Adachi and colleagues took a similar approach and optimized the scaffold microstructure to match the stiffness of healthy bone [64]. Others have used hypothesis-driven models to analyze the complex behavior of biological systems, for example Kang and coworkers [65] developed a computational model of foreign body response reaction on biomaterials; the model was able to predict a previously unobserved outcome. Similarly, Carlier and colleagues [66,67] developed and implemented a computational simulation model that was able to predict ectopic bone formation in response to calcium ion release from calcium phosphate scaffolds. The model was used to find the optimal scaffold design and cell culture conditions.

Computational Requirements for Data Analysis – Do I Need To Buy a Supercomputer?

Computational requirements for machine learning depend on the amount and type of data, as well as on the algorithms being used. Despite the explosion in the size of the datasets produced in biological experiments, many data analysis problems tend to be 'small data' problems and can be solved easily on modern consumer-grade computers. However, large-scale high-throughput experiments, such as image-based experiments, can produce several terabytes of images from a single microscope per day [68], making it essential to have access to not only high-capacity storage but also to greater computing power. For instance, processing a 384-well plate of data from a multiplex assay such as Cell Painting [18] can take several days if performed on a single desktop computer. Further, techniques such as deep learning require the use of dedicated hardware to perform at scale [28] – the speedup using dedicated hardware is typically about a couple of orders of magnitude, reducing the processing time from several days to a few hours. Similarly, high-throughput gene expression assays require several stages of data processing for converting the raw data into differential expression levels, and necessitate the use of more-powerful computing systems [10].

To the rescue are commercial 'cloud computing' services, such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure, which obviate the need for researchers to set up their own physical computational infrastructure. Cloud computing services – services that can be accessed over the Internet – enable quickly setting up a personal cluster, which is a group of computers connected to each other. Importantly, the clusters can be easily scaled up or down (they are said to be 'elastic') depending on the need. The services also provide access to practically limitless data storage, including specialized databases. Specialized hardwares, such as those required for deep neural networks, are also available because powerful computers can currently have as many as 128 central processing units (CPU). Further, emerging software services such as Google Cloud Machine Learning will likely make it trivial to perform data analysis, requiring almost no software to be set up. Existing models can be selected, trained, tested, and validated using standardized workflows.

Finally, public database repositories allow researchers to deposit their data in a standardized manner. For instance, the Gene Expression Omnibus [69] is a database repository of high-throughput gene expression data and related techniques. Image Data Resource [70] is a recently launched public repository for image datasets. These resources not only decrease the burden on individual laboratories to manage their own data but also enable easy dissemination of their work.

Concluding Remarks and Perspective

We are entering an exciting era where massive amounts of experimental data are being produced, and computational methods can predict the behavior of biological systems using these data. By tapping into advances in the field of machine learning and computational simulations, biomaterial researchers can view their experiments through the lens of a ‘data scientist’ – using mathematical abstractions to transform their problems into ones that have well-established and generalized solutions. Machine learning techniques can be used even in the absence of any hypothesis about the biological system or material properties. By contrast, computational simulations are an essential part of the discovery process because they allow the researcher to make explicit hypotheses, test them, and make predictions using the models that encode their hypotheses.

The basics of the computational concepts involved are relatively straightforward, and the reference material cited here can bring the interested researcher up to speed on the details of techniques. However, establishing collaborations with computational researchers is invaluable – in our own experience, a close collaboration can dramatically accelerate the pace of research.

While the computational methods presented in this review are powerful, they bring several potential pitfalls that should be kept in mind. With machine learning techniques, there is a temptation to test out numerous predictive models – given the ease and availability of doing so – which can lead to overfitting of the data, even when all the common practices of model validation are followed [71,72]. Further, given the hundreds of variables involved, and the thousands of experiments that may be conducted, it is easy to fall prey to spurious correlations – correlations that happen by random chance, or by forgetting to adjust significance levels for multiple testing [72].

While several open questions remain (see Outstanding Questions), the biomaterials field is already well suited for taking on a computational approach to investigating cell–material interactions. Omics technology has given us the power to generate large datasets on cell–material interaction. By equipping ourselves with knowledge of computational tools, as well as effective collaborations, we will be prepared to channel the deluge of data towards making scientific breakthroughs.

Acknowledgments

We are grateful to the editor and the anonymous reviewers whose inputs helped to improve the article significantly. We also thank Prof. Dr. Morgan Alexander, Dr Andrew Hook, and Dr Matthew B. Baker for fruitful discussions. A.V., A.C., and J.d. B. acknowledge the financial contribution of the Province of Limburg. A.C. gratefully acknowledges the Dutch Science Foundation (NWO) for a VENI grant (15075). The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement 289720.

Outstanding Questions

How do we create a standard for parameterizing biomaterials? Although different parameterizations exist, there is currently no consensus in the field on how to standardize them.

How should data from different sources be integrated? As more approaches for parameterizing cellular states become available with high throughput, the field faces the challenge of combining these measurements for making predictions.

Can the application of machine learning be extended beyond prediction to a simulation of cell–material interactions? Recent advances have been made in simulating physical models using machine learning; doing so for biological systems is likely to be much more challenging.

How should the community gear itself up to adopting computational methods for solving problems? What incentives should funding organizations and institutions put in place to nudge researchers to adopt these methods? What role should scientific journals play in encouraging research in this direction? How should the next generation of scientists be trained to develop competence in both computing and biomaterial science? How should the field make itself attractive to the machine learning and mathematical modeling community such that more researchers from those domains are drawn to conducting research in biomaterials?

Given the need for data repositories in the field, what funding mechanisms can be designed to encourage the community to create such resources?

What is the holy grail in biomaterials, and what is the role of computational techniques in finding it?

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tibtech.2017.05.007.

References

1. Barthes, J. *et al.* (2014) Cell microenvironment engineering and monitoring for tissue engineering and regenerative medicine: the recent advances. *BioMed. Res. Int.* 2014, 921905
2. Murphy, W.L. *et al.* (2014) Materials as stem cell regulators. *Nat. Mater.* 13, 547–557
3. Discher, D.E. *et al.* (2009) Growth factors, matrices, and forces combine and control stem cells. *Science* 324, 1673–1677
4. Dorland, W.A.N. (2007) *Dorland's Medical Dictionary for Health Consumers,* Saunders
5. Crowder Spencer, W. *et al.* (2016) Material cues as potent regulators of epigenetics and stem cell function. *Cell Stem Cell* 18, 39–52
6. Feller, L. *et al.* (2015) Cellular responses evoked by different surface characteristics of intraosseous titanium implants. *BioMed. Res. Int.* 2015, 171945
7. Wen, J.H. *et al.* (2014) Interplay of matrix stiffness and protein tethering in stem cell differentiation. *Nat. Mater.* 13, 979–987
8. Chaudhuri, O. *et al.* (2015) Substrate stress relaxation regulates cell spreading. *Nat. Commun.* 6, 6364
9. Watt, F.M. and Huck, W.T. (2013) Role of the extracellular matrix in regulating stem cell fate. *Nat. Rev. Mol. Cell Biol.* 14, 467–473
10. Subramanian, A. *et al.* (2017) A next generation connectivity map: L1000 Platform and the first 1,000,000 profiles. *bioRxiv* 136168
11. Qu, X.A. and Rajpal, D.K. (2012) Applications of Connectivity Map in drug discovery and development. *Drug Discov. Today* 17, 1289–1298
12. Abraham, Y. *et al.* (2014) Multiparametric analysis of screening data: growing beyond the single dimension to infinity and beyond. *J. Biomol. Screen.* 19, 628–639
13. Johannessen, C.M. *et al.* (2015) Integrating phenotypic small-molecule profiling and human genetics: the next phase in drug discovery. *Trends Genet.* 31, 16–23
14. Groen, N. *et al.* (2015) Exploring the material-induced transcriptional landscape of osteoblasts on bone graft materials. *Adv. Healthc. Mater.* 4, 1691–1700
15. Eyckmans, J. *et al.* (2013) Mapping calcium phosphate activated gene networks as a strategy for targeted osteoinduction of human progenitors in vitro and in vivo. *Biomaterials* 34, 4612–4621
16. Groen, N. *et al.* (2017) Linking the transcriptional landscape of bone induction to biomaterial design parameters. *Adv. Mater.* 29, 1603259
17. Caicedo, J.C. *et al.* (2016) Applications in image-based profiling of perturbations. *Curr. Opin. Biotechnol.* 39, 134–142
18. Bray, M.-A. *et al.* (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* 11, 1757–1774
19. Angelo, M. *et al.* (2014) Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* 20, 436–442
20. Eliceiri, K.W. *et al.* (2012) Biological imaging software tools. *Nat Methods* 9, 697–710
21. Fennema, E. *et al.* (2013) Spheroid culture as a tool for creating 3D complex tissues. *Trends Biotechnol.* 31, 108–115
22. Yuan, H. *et al.* (2010) Osteoinductive ceramics as a synthetic alternative to autologous bone grafting. *Proc. Natl. Acad. Sci.* 107, 13614–13619
23. Celiz, A.D. *et al.* (2014) High throughput assessment and chemometric analysis of the interaction of epithelial and fibroblast cells with a polymer library. *Appl. Surf. Sci.* 313, 926–935
24. Sliwoski, G. *et al.* (2014) Computational methods in drug discovery. *Pharmacol. Rev.* 66, 334–395
25. Chow, L.W. and Fischer, J.F. (2016) Creating biomaterials with spatially organized functionality. *Exp. Biol. Med.* 241, 1025–1032 http://dx.doi.org/10.1177/1535370216648023
26. Urquhart, A.J. *et al.* (2007) High throughput surface characterisation of a combinatorial material library. *Adv. Mater.* 19, 2486–2491
27. Jordan, M.I. and Mitchell, T.M. (2015) Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260
28. LeCun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444
29. Mitchell, T.M. (1997) *Machine Learning,* McGraw-Hill
30. Russell, S.J. and Norvig, P. (2002) *Artificial Intelligence: A Modern Approach,* Prentice Hall
31. Lutz, M. (2013) *Learning Python,* O'Reilly Media
32. McKinney, W. (2012) *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython,* O'Reilly Media
33. Wickham, H. and Grolemund, G. (2016) *R for Data Science,* O'Reilly
34. Garreta, R. and Moncecchi, G. (2013) *Learning Scikit-Learn: Machine Learning in Python,* Packt Publishing
35. Kuhn, M. (2008) Caret package. *J. Stat. Softw.* 28, 1–26
36. Hall, M. *et al.* (2009) The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* 11, 10–18
37. Berthold, M.R. *et al.* (2008) KNIME: the Konstanz information miner. In *Data Analysis, Machine Learning and Applications* (Preisach, C., ed.), pp. 319–326, Springer
38. Hook, A.L. *et al.* (2012) Combinatorial discovery of polymers resistant to bacterial attachment. *Nat. Biotechnol.* 30, 868–875
39. James, G. *et al.* (2013) *An Introduction to Statistical Learning,* Springer
40. Fernandez, M. *et al.* (2014) Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO_2 capture. *J. Phys Chem. Lett.* 5, 3056–3060
41. Epa, V.C. *et al.* (2012) Modelling human embryoid body cell adhesion to a combinatorial library of polymer surfaces. *J. Mater. Chem.* 22, 20902–20906
42. Autefage, H. *et al.* (2015) Sparse feature selection methods identify unexpected global cellular response to strontium-containing materials. *Proc. Natl. Acad. Sci.* 112, 4280–4285
43. Kumar, G. *et al.* (2011) The determination of stem cell fate by 3D scaffold structures through the control of cell shape. *Biomaterials* 32, 9188–9196
44. Reimer, A. *et al.* (2016) Scalable topographies to support proliferation and Oct4 expression by human induced pluripotent stem cells. *Sci. Rep.* 6, 18948
45. Semple, J.L. *et al.* (2005) Review: in vitro, in vivo, in silico: computational systems in tissue engineering and regenerative medicine. *Tissue Eng.* 11, 341–356
46. Geris, L. (2014) Regenerative orthopaedics: in vitro, in vivo . . . in silico. *Int. Orthop.* 38, 1771–1778
47. Scholma, J. *et al.* (2014) Biological networks 101: computational modeling for molecular biologists. *Gene* 533, 379–384
48. Xiong, F. and Megason, S.G. (2015) Abstracting the principles of development using imaging and modeling. *Integr. Biol.* 7, 633–642
49. Brodland, G.W. (2015) How computational models can help unlock biological systems. *Semin. Cell Dev. Biol.* 00, 62–73
50. Hecht, F. (2012) New development in FreeFem++. *J. Numer. Math.* 20, 251–266
51. Myers, C. (2011) SBML and synthetic biology. *Nat. Preceed.* Published online September 6, 2011. http://dx.doi.org/10.1038/npre.2011.6343.1
52. Cooling, M.T. and Hunter, P. (2015) The CellML metadata framework 2.0 specification. *J. Integr. Bioinform.* 12, 86–103
53. Hunt, B.R. *et al.* (2014) *A Guide to MATLAB: For Beginners and Experienced Users,* Cambridge University Press
54. Lopez, C.F. *et al.* (2013) Programming biological models in Python using PySB. *Mol. Syst. Biol.* 9, 646

55. Loew, L.M. and Schaff, J.C. (2001) The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol.* 19, 401–406

56. Swat, M.H. *et al.* (2012) Multi-scale modeling of tissues using CompuCell3D. *Methods Cell Biol.* 110, 325

57. Starruß, J. *et al.* (2014) Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology. *Bioinformatics* 30, 1331–1332

58. Maas, S.A. *et al.* (2012) FEBio: finite elements for biomechanics. *J. Biomech. Eng.* 134, 011005

59. Saxton, M.J. and Jacobson, K. (1997) Single-particle tracking: applications to membrane dynamics. *Annu. Rev. Biophys. Biomol. Struct.* 26, 373–399

60. Tang, Y.H. and Zhang, H.P. (2016) Theoretical understanding of bio-interfaces/bio-surfaces by simulation: a mini review. *Biosurf. Biotribol.* 2, 151–161

61. Baker, M.B. *et al.* (2015) Consequences of chirality on the dynamics of a water-soluble supramolecular polymer. *Nat. Commun.* 6, 6234

62. Giannitelli, S.M. *et al.* (2014) Current trends in the design of scaffolds for computer-aided tissue engineering. *Acta Biomater.* 10, 580–594

63. Byrne, D.P. *et al.* (2007) Simulation of tissue differentiation in a scaffold as a function of porosity, Young's modulus and dissolution rate: application of mechanobiological models in tissue engineering. *Biomaterials* 28, 5544–5554

64. Adachi, T. *et al.* (2006) Framework for optimal design of porous scaffold microstructure by computational simulation of bone regeneration. *Biomaterials* 27, 3964–3972

65. Kang, M. *et al.* (2016) Computational modeling of phagocyte transmigration for foreign body responses to subcutaneous biomaterial implants in mice. *BMC Bioinform.* 17, 111

66. Carlier, A. *et al.* (2011) Designing optimal calcium phosphate scaffold–cell combinations using an integrative model-based approach. *Acta Biomater.* 7, 3573–3585

67. Manhas, V. *et al.* (2017) Computational modelling of local calcium ions release from calcium phosphate-based scaffolds. *Biomech. Model. Mechanobiol.* 16, 425–438

68. Meijering, E. *et al.* (2016) Imagining the future of bioimage analysis. *Nat. Biotechnol.* 34, 1250–1255

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

70. Williams, E. *et al.* (2016) The Image Data Resource: a scalable platform for biological image data access, integration, and dissemination. *bioRxiv* 089359

71. Skocik, M. *et al.* (2016) I tried a bunch of things: the dangers of unexpected overfitting in classification. *bioRxiv* 078816

72. Saeb, S. *et al.* (2016) Voodoo machine learning for clinical predictions. *bioRxiv* 059774

73. Boyd, S. and Vandenberghe, L. (2004) *Convex Optimization,* Cambridge University Press

74. Draper, N.R. and Smith, H. (2014) *Applied Regression Analysis,* John Wiley & Sons

75. Jain, A.K. *et al.* (1999) Data clustering: a review. *ACM Comput. Surv.* 31, 264–323

76. Xu, R. and Wunsch, D.C., 2nd (2010) Clustering algorithms in biomedical research: a review. *IEEE Rev. Biomed. Eng.* 3, 120–154

77. Jolliffe, I.T. (2013) *Principal Component Analysis,* Springer New York

78. Guyon, I. *et al.* (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182

79. Van Der Maaten, L. *et al.* (2009) *Dimensionality Reduction: a Comparative Review,* Tilburg University Report No.: 2009–005. Published online October 29, 2009. https://www.tilburguniversity.edu/upload/59afb3b8-21a5-4c78-8eb3-6510597382db_TR2009005.pdf