Structural biology meets data science: does anything change?
Cameron Mura\(^1,3\), Eli J Draizen\(^1,3\) and Philip E Bourne\(^1,2\)

Data science has emerged from the proliferation of digital data, coupled with advances in algorithms, software and hardware (e.g., GPU computing). Innovations in structural biology have been driven by similar factors, spurring us to ask: can these two fields impact one another in deep and hitherto unforeseen ways? We posit that the answer is yes. New biological knowledge lies in the relationships between sequence, structure, function and disease, all of which play out on the stage of evolution, and data science enables us to elucidate these relationships at scale. Here, we consider the above question from the five key pillars of data science: acquisition, engineering, analytics, visualization and policy, with an emphasis on machine learning as the premier analytics approach.

Addresses
\(^1\) Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA
\(^2\) Data Science Institute, University of Virginia, Charlottesville, VA 22904, USA

Corresponding author: Bourne, Philip E (ppeb@virginia.edu)
\(^3\) CM and EJD contributed equally to this work.

Introduction
The term Structural Biology (SB) can be defined rather precisely as a scientific field, but Data Science (DS) is more enigmatic, at least currently. The intrinsic difference is two-fold. First, DS is a young field, so its precise meaning—based on what we practice and how we educate its practitioners—has had less time than SB [1,2] to coalesce into a consensus definition. Second, and more fundamental, DS is interdisciplinary to an extreme; indeed, DS is not so much a field in itself as it is a way of doing science, given large amounts of diverse and complex data, suitable algorithms and sufficient computing resources. Such is the breadth and depth of DS that it has been described as a fourth paradigm of science, alongside the theoretical, experimental and computational [3,4]. Because it is so vast and sprawling, a helpful organizational scheme is to consider four V’s and five P’s that characterize data and DS (Figure 1).

The four V’s describe the properties of data: volume, velocity, variety and veracity. The P’s are the five disciplinary pillars (P-i through P-v) of DS (Figure 1): (i) data acquisition, (ii) data reduction, integration and engineering, (iii) data analysis (often via machine learning), (iv) data visualization, provenance and dissemination, and (v) ethical, legal, social and policy-related matters. The P’s are interrelated, as are the V’s. For example, the fifth pillar leans into each of the other four: a host of privacy matters surround data acquisition, aggregation can have unforeseen security concerns, analytics algorithms can introduce unintended bias, and dissemination policies raise licensing and intellectual property issues. Similarly, many modes of data analysis (P-iii) rely on advanced visualization approaches (P-iv). The P’s also closely link to the four V’s. For example, P-i, the data acquisition pillar, clearly relates to volume and velocity. More subtle linkages also exist, e.g., between data analysis and variety in structural biology, hybrid approaches [5–7,8] involve joint integration/analysis of heterogeneous varieties of data (e.g., cryo-EM, mass spectrometry, cross-linking), for instance via a Bayesian statistical formulation of the structure determination process [9,10]. The philosophy and epistemology of DS is an entire field unto itself, and helpful starting points can be found in recent texts [11**].

The rest of this review focuses on the junction of data science and structural biology. We consider DS approaches that have been applied in SB recently, including examples from crystallography and protein interactions. We focus mostly on pillar P-iii (Figure 1), and specifically machine learning. In so doing, we largely ignore traditional disciplinary labels. For example, the junction of DS and SB could be viewed as simply expanding the field of structural bioinformatics [12]; but, such disciplinary labels and boundaries matter less than the actual scientific impact. Analogously, definitions of ‘the internet’ vary greatly, yet the impact of the internet on science is unmistakable. For convenience, we use the term ‘SB’ as including structural bioinformatics, simply to distinguish what has gone before versus what may lie on the horizon. We suspect much lies on the horizon: akin to the rapid growth [13] of databases such as the Protein Data Bank (PDB; [14]), our assessment of bibliometric data (Figure 2) suggests that data science will profoundly impact the biosciences, including structural biology. (The best-fit curve in Figure 2 is supra-exponential, with no inflection point in sight.) Conversely, can SB impact the broader field of DS? This has yet to occur in a definitive way, but, given the maturity of SB as a discipline, much can be learnt from it and its history; thus, we start with a short review of how SB might influence DS.
What structural biology has to offer data science

Open science

SB has pioneered open science through the provision of the PDB and many derivative data sources. The complete corpus of structural information in the PDB is free of copyright and is available for un fettered use, non-commercial or otherwise (P-v). Moreover, community practices—such as virtually no journal publishing an article without its data deposited in the PDB [15]—is a precedent that, if broadly adopted in other disciplines, would deepen the amount and diversity of data available for DS-like approaches in those other scientific and technical domains. The creation and free distribution of software (SW) tools has echoed this trend, as epitomized by the Collaborative Computational Project 4 (CCP4); developed and meticulously maintained since 1979 [16], the CCP4 suite has been a mainstay of the crystallographic structure-determination process. CCP4 and kindred projects, alongside myriad other SW tools and attendant data, have fostered an open discipline. DS draws upon data and ideas from a wide range of disciplinary areas, but some of these areas have been less open than SB, at least historically. To succeed, we believe that any DS must abide by the ‘FAIR’ principles, enabling researchers to Find, Access, Interoperate and Reuse data and analyses [17†]. SB has exercised this for decades, and is thus positioned to lead the way.

Reproducibility

In principle, reproducibility is the bedrock of the scientific enterprise. And, as a byproduct of open science, reproducibility has been central in SB, though often less so in other realms of DS. Cultural differences across various disciplines, often driven by (perceived) competitive pressures,
The recent surge in publishing activity for machine learning in the biosciences, shown here via bibliometric data obtained for the PubMed/MEDLINE (orange) and ISI Web of Science (blue) literature databases. The overlaid histograms show the number of publications in which the string ‘machine learning’ co-occurs in the title or abstract fields; the precise PubMed query was ‘machine learning[Title/Abstract]’, and the string ‘machine learning AND bio’ was used for an ISI Topic search. Both datasets were fit with the same four functions listed in Figure 1. For both PubMed and ISI, a subtle crossover occurs wherein a supra-exponential (form iv) gives a better fit than a pure power law; such highly nonlinear ‘J-curves’ or ‘hockey stick curves’ arise in systems subject to singularly disruptive forces (e.g., human population growth after the Industrial Revolution, climate temperatures in the past century). Intriguingly, the approximate year of crossover — 2010 for the ISI data, 2012 for PubMed — is generally regarded as the ‘breakthrough year’ for Deep Learning (e.g., Google Brain learned a ‘cat’ de novo, from YouTube data), enabled by advances such as GPU computing, algorithmic approaches such as ReLU and ‘dropout’, and vast stores of labelled training data (ImageNet). Judging by these charts, ML has begun driving a substantial transformation in the biosciences.

have dampened what could be the norm. In SB, the systematic, pipelined nature of many structure-determination approaches has facilitated reproducibility. A notable example is the effort, spurred by structural genomics, to annotate large-scale macromolecular crystallization experiments and to conduct careful target tracking [18]; in principle, such efforts afford a rich source of data, exploitable by DS via data mining and machine learning methods [19].

**Workflows, high-performance computing**

Reproducibility, in turn, is facilitated by workflows. Some workflow management systems (WMS) are domain-specific (e.g., Galaxy for genomics [20,21]), while others are more generic or monolithic (e.g., KNIME [22]); lightweight toolkits also exist, providing libraries to write custom parallel-processing pipelines (e.g., [23,24]). Again, structural genomics and other data-rich areas (e.g., large-scale biomolecular simulations) have prompted the development of WMS solutions. Closely related to workflows, recent technologies that have become best practices in DS—such as Jupyter notebooks (as a user interface) and Docker ‘containers’ (for virtualized runtime environments)—likely will be adopted more broadly in SB, as research questions become more quantitative and as data-intensive computational steps are pursued via distributed computing and other modes of HPC. Cloud computing and related approaches, such as the MapReduce paradigm (implemented in Hadoop), rapidly entered genomics and bioinformatics early on [25] and are becoming more widely adopted in other biosciences too, including SB [26]; other examples include large-scale biomolecular modeling for virtual screening and drug design [27,28] and, more recently, pipelines for cryo-EM structure determination [29].

Structural biology has relied upon HPC since the dawn of supercomputing in the 1960s. A recent example using HPC involves the phasing of diffraction data. Recognizing the wealth of structural information in the PDB, and that molecular replacement (MR) can be treated as embarrassingly parallel across all these structures, the BALBES [30] pipeline leverages all known 3D structures to create and then use MR search models in an automated manner. This approach was recently extended to fitting 3D models into cryo-EM maps [31]. Somewhat similar in spirit, *PDB_REDO* endeavors to automatically improve all PDB structures by re-refining 3D models against the original X-ray data, utilizing established refinement approaches (e.g., TLS) and grid computing [32,33]. As a final example, a recent and highly creative approach to crystallographic phasing has a strong DS feel: Encoding phase values as 9-bit strings (genes), and applying a genetic algorithm for sampling/optimization, Yeates et al. [34] developed a crowdsourced gaming platform for *ab initio* phasing, at least to low resolution. Such ‘citizen-science’ [35,36] approaches will likely play broader roles in SB (and DS) in the coming decade.

**Visualization**

Visualization has played a key role in SB since the 1950s, when the first macromolecular structures were determined. Concepts, principles and best practices for biomolecular visualization can be found in many reviews [37,38,39]; the supplement in [37] traces the historical development of this field. Recent advances have occurred in web browser-embedded, hardware-accelerated tools for interactive molecular visualization, such as the NGL Viewer [40]; in the future, a greater share of visualization work likely will occur within browsers. To transcend how molecular renderings are usually communicated (as static images), we suspect that much could be gained by comparing visualization techniques in DS and SB. Though iconic and highly informative, beware the ‘curse of the ribbon’: macromolecules are dynamic, multifaceted entities, and static renditions are but a starting point. For similar reasons, there is a need for molecular visualization platforms that transcend simple graphical viewers — that enable facile, flexible and extensible integration of other forms/modalities of data and novel visualization techniques [41], such as the data-intensive sort that often arise with big data. We believe that DS tools can address this need; note that texts are becoming available on this topic, such as the recent *Big Data Visualization* [42]. Ideas and methods from beyond...
SB—such as ‘chord diagram’ layouts in genomics [43], termed ‘hierarchical edge bundles’ [44] in computer graphics—can be applied in SB, for instance to visualize data associated with hierarchical clustering of protein structural differences (e.g., see the figures in [45]).

Finally, note that further areas of SB × DS overlap can be identified, but are not treated here because of space limitations. Three such examples are: (i) database (DB)-related issues, including structured versus unstructured data, relational versus non-relational DBs and query languages [46]; (ii) systems and network biology [47]; and (iii) ontologies and formal knowledge representation systems [48,49].

What data science analytics has to offer structural biology

DS analytics spans a vast territory, including applied mathematics, statistics and computer science. Here, we focus on two machine learning (ML) approaches—one which has received much recent attention (deep learning [DL]), and one for which we envision possible applications in SB (natural language processing [NLP]). A glossary is included (Box 1), and the Supplementary material offers: (i) a brief primer on ML, (ii) a concise historical note on early applications of neural networks in SB, and (iii) a short sketch of the general applicability of data sciences in structural biology and other biosciences.

Machine learning applied to biomolecular interactions

In a recent wave of activity, DL methods have been applied to model and predict protein–ligand and protein–protein interactions (PLI, PPI). Deep learning is a type of ML that employs deep (multi-layered) neural network (NN) architectures; training and deploying such architectures is now feasible because of the exceptional computational performance of modern GPU-equipped clusters.

Accurately predicting and modeling PLIs (structural poses, energetics) would advance many areas, both basic (e.g., evolutionary analyses of ligand-binding properties) and applied (e.g., drug design and discovery). Historically, this field has largely relied on two distinct methodological approaches: quantitative structure-activity relationships (QSAR) and in silico docking. Virtual screening, wherein one docks against large libraries of small compounds, is an established example of DS in SB; as a newer example, note that workflow-based approaches to high-throughput crystallographic fragment screening have a significant DS component [50]. Extensions of the basic QSAR and ligand-docking approaches also call upon DS. For example, recognizing that a protein exists as an ensemble of thermally-accessible conformational states in solution, simulations have been combined with docking in the ‘relaxed complex’ scheme to capture receptor flexibility [51]. Similar in spirit, data-intensive ‘ensemble-based’ methods [52] can enable dynamic pharmacophore models (e.g., [53]) to be devised.

In a recent approach, a workflow to discover ‘cryptic’ (and druggable?) binding sites was developed by integrating comparative structural analyses, pocket-detection algorithms, fragment docking, molecular simulations, and an ML classifier [54]. In another data-driven, structure-based

Glossary

The following terms, organized thematically, appear in this review or are pervasive in the literature. As part of the DS jargon, the terminologies may be unfamiliar and is therefore included here for convenience.

Statistical and machine learning

Classifier: An algorithm or function that maps input data into one of at least two categories (or classes). For instance, if only two classes are possible (e.g., True or False, Exon or Intron), and input data are integers, then the modulo operation (mod 2) could serve as a binary classifier.

Model: A formal relationship between input data and some set of outputs; another way to view this is as a mapping, association rule or mathematical function. As a concrete example, we have an ideal, one-dimensional spring hanging frictionless surface. Say we collect dense (finely-sampled) data on the precise position $y$ of the spring’s terminus at many time-points (i.e., we have a time-series, $x(t)$). To elucidate the system’s behaviour in terms of our data, we may propose an explanation, say, $y=x+e$; $A. ñt=A.y$, where $A$ is amplitude, $f$ is phase and $v_0$ is angular frequency. This functional form is what we mean by a model: the precise parameters will vary from system to system (differential springs, stiffnesses, etc.), and what matters instead is the functional form of the mapping (in this case, the equation of motion models simple harmonic oscillation). For more complex systems—e.g., recognizing patterns in images, delineating protein structures—such simple, closed-form expressions generally do not exist (nevertimdes us being able to propose them a priori); statistical approaches come to the rescue by offering ways to learn a model for the set of input → output associations.

Regression: A statistical approach to estimate relationships amongst variables; e.g., linear regression will estimate a linear relationship (or slope) between two or more variables, which can be used for purposes of prediction and classification.

Random forest: An ensemble of decision trees.

Decision tree: A classifier that follows if-then-else decision rules to traverse a directed graph, thus predicting an output. The rules, or nodes of the tree, are the features of the model sorted by information gain split on certain values of a feature.

Supervised, unsupervised learning: See the Supplementary material for a description of these terms.

Support vector machine (SVM): A classifier that finds a linear discriminatory boundary between classes, generally via regression in a higher-dimensional space or application of kernel methods (most simply, a ‘kernel’ can be viewed as a measure of similarity between two feature sets, e.g., the dot product).

Neural networks and deep learning

Neural network (NN): Also known as an ‘artificial NN’ or ‘multilayer perceptron’ in the older literature, these are mathematical networks of nodes (hence, neurons) and connections (terminals), the system’s terminus or ‘output’ class (e.g., Receiver operating characteristic, ROC curve).

Feedforward NN: An NN architecture wherein information flows through the network unidirectionally, from the input layer to the output layer. This is possible because the edges (links) are directed from one node to another; this network topology is a type of directed acyclic graph (DAG), and other DAG-based NN architectures are conceivable.

Convolutional NN (CNN): ANN that applies convolutional operations, which take local, connected, sub-regions of an input (or matrix) as units. Inputs are typically 2D images, which is a 2D matrix of pixels, where the sub-regions are smaller pieces of the image, or 3D volumes where smaller cubes traverse the volume.

Deep CNN (DNN): Most simply, a NN architecture that includes multiple hidden layers. Backpropagation: A method to update learnable weights of the NN interconnects between nodes by transmitting errors backwards (in the direction from the output layer towards the input layer); this backwards propagation of errors, in turn, corresponds to the network improving on a particular prediction, i.e., the network can be said to ‘learn’. More concretely, backpropagation proceeds by applying the chain rule to compute the gradient of the error (the loss function) at each filter (node) for a given layer, and iteratively using the gradient values to update the weights; therefore, this is fundamentally a gradient descent algorithm, as found in many classes of optimization problems.

Loss function: A function to compute the error between the true and predicted values. For example, this could be as simple as the Euclidean distance between estimated and true (target) values.

Dropout: A technique to address overfitting by removing randomly selected subset of nodes, in a single layer, during training (and forward and backward pass); this allows the NNs to learn more robust features by forcing different possible subsets of nodes, typically, on the order of 50% of nodes are silenced.

Regularization: A technique to optimally balance the perils of underfitting/overfitting to training datasets.

Epoch: One forward and one backward pass of all training data. Many epochs (typically ranging from 30 to 100) are usually required before a NN model converges.

Natural language processing

Corpus: Most simply, a collection of information. This term, prevalent in the NLP field, is frequently used to generically refer to written data (books, journals, etc.), where it often means a comprehensive collection on a particular topic (all writings by particular authors, or about a particular protein, etc.).

Topic: Most simply, a statistical distribution of words, each word being drawn from a well-defined set of words (as in vocabulary); a topic can also be viewed as a theme. In many ways, a given document is defined by its collection of most prevalent topics.

Topic modeling (TM): A set of unsupervised algorithms to discover the topics in a corpus of unstructured information, generally by applying statistical algorithms to analyze and model word distributions.

Latent structure: A highly general concept, referring to there being some general correlation (or non-randomness, or ‘structure’) among the hidden (latent) random variables that define the probabilistic distributions underlying modes such as DaT in TM, we seek to learn these relationships between hidden variables (i.e., the structure), which manifests in the form of (non-random) topics.

Latent Dirichlet allocation (LDA): A type of TM wherein a document (a ‘bag of words’) is viewed as a probabilistic distribution over a set of topics; in its topical view, it is a generative model in which the observed data (the document, its words, their distributions) are taken as having been generated via sampling a hidden distribution (a random process, or, if there is some latent structure, a non-random process). Briefly, each document’s set of topics is taken to be Dirichlet-distributed, and the words in a document are allocated to its various topics based on this distribution (in the Bayesian sense of joint distributions, prior, etc.), the Dirichlet distribution is the conjugate prior to the multinomial distribution that is taken as explaining the distribution of topics.)
Recent work on predicting PLIs has directly employed ML, including for the interrelated goals of virtual screening, affinity prediction and pose prediction. The application of statistical and ML approaches, in particular deep neural nets, to the PLI problem was reviewed recently [57**]. Here, we mention only that a surge of new work has applied convolutional neural nets (CNNs) to the PLI problem — references [58**,59–63] comprise a partial list from just the past year. Notably, these purely ML-based approaches rely on human expertise only in the early stage of choosing structural descriptors (hydrophobicity, ionizability, etc.), which are input features for NN training. The protein structure (either as a complex, or just receptor) can be treated as a 3D image, wherein atoms that compose the structure are assigned to discrete volumetric elements (voxels). CNNs excel at learning from 2D image data [64**], suggesting that their 3D counterparts, 3D deep CNNs, can be used for volumetric analysis. Leveraging these ideas, the 3D DCNN of DeepSite achieved state of the art performance, having been trained on known protein–ligand structures [58**].

As with PLIs, protein–protein interactions (PPIs) are critical to much of cell biology, and are another focal point of recent ML efforts. Were all binary PPIs known, they could be used to build whole species interactomes [65] and inter-species (e.g., host-pathogen) interactomes [66,67], which, in turn, would aid elucidation of signaling pathways [68], metabolic networks (Recon3D [69]), and evolutionary pathways [70]. ML can be used to predict which two proteins interact and what specific residues (‘hot-spots’) mediate the interaction (i.e., binding sites). If both binding sites (or interfaces) are known, they can be used to model structures of their complexes.

Thus far, the optimal information for predicting interacting residues has been at the sequence level, using residue co-evolution. Intuitively, residues that co-evolve between two proteins are likely to contact one another. Such sites can be predicted using ML and DL methods like maximum entropy models or 2D-CNNs [71–75]; a drawback to such approaches is the need for sufficiently large protein families. For purposes of structure prediction, the same approach can be used to predict residue–residue contacts from one protein family alignment. One can also predict PPIs from structure if a query protein is homologous (based either on sequence or structure) to one protein in a known PPI. If the identity of only one interaction partner is known, and the binding sites in the other partner unknown, binding sites and partners can be predicted by structurally aligning a query to crystal structures of complexes, using either local (e.g., PRISM [76]) or global (e.g., IBIS [77]) 3D superimpositions. Residues from the query protein that align to one side of an interface are predicted to be a part of the binding site.

ML methods can also predict binding-site residues given the 3D structure of only one partner. Here, atomic and residue-level features (e.g., hydrophobicity, phylogenetic conservation) are calculated for all structures in the PDB. True binding site residues are taken from crystallized complexes, split into monomers, and used to train a classifier (SVMs, Decision Trees, etc.). Unfortunately, such predictors have suffered from low precision and recall [78], perhaps because the 3D spatial details of the data are not retained but rather enter the model only as ‘flattened’ features (or, assumptions of independence are applied incorrectly). Given current limitations and difficulties, it is unsurprising that DL is now starting to be applied to PPI modeling and prediction. While only biomolecular interactions are discussed here, we envision that contemporary DL approaches, such as variational autoencoders, will play major roles in areas such as structure prediction [79], protein design and evolutionary analyses [80].

Natural language processing applied to biomolecular assemblies

NLP is a form of machine learning concerned with processing and analyzing language, written or spoken. Here, ‘processing’ can mean many things: analyzing frequencies and co-occurrences of words and higher-order units (n-grams), parsing texts in syntactic/grammatical analyses, information retrieval, machine translation, language comprehension (and synthesis), and beyond. The scope of this topic is vast, and helpful biosciences primers are available [81–83]. NLP methods play key roles in routine tasks such as search/query (databases, knowledge-bases), information extraction and text summarization; indeed, you may have reached this article via a web-query using the PubMed search engine. PubMed is but one example of NLP in biomedical informatics, and it remains an active area of research; recent examples include a ‘neural word embedding’ approach for document matching in PubMed [84] and development of a flexible term ↔ concept matching system for biomedical pipelines [85]. Apart from search and mining in biomedical literature, might NLP impact structural biology in other ways?

NLP-like approaches have been applied to detect the subcellular localization of proteins [86,87] and, recently, to predict structures of protein complexes [88]. Notably, using ML–enhanced NLP, versus a purely text-mining-based NLP approach, was found to significantly improve the structural predictions of complexes [89]. Note that both sorts of problems — subcellular localization and structural modeling — are distinctly spatial, or image-based, as opposed to textual. For this reason, we expect that a relatively new and highly-generalized approach to NLP, termed topic modeling (TM), holds great promise in the biosciences. In TM, ‘topics’ are extracted over a corpus of unstructured data (e.g., a set of books) using a probabilistic machine learning framework;
fundamentally, this is achieved by examining the distributions of words (a ‘bag of words’ ansatz) under a generative statistical model, such as the latent Dirichlet allocation (LDA). An introductory review of TM and a recent overview of TM-like approaches in bioinformatics can be found in refs [90*] and [91*], respectively. To extend TM to other areas — including even the learning of topics (themes) from non-textual data like protein structures — the basic issue is one of defining a suitable mapping of one’s problem to TM’s core framework of document \(\rightarrow\) topic \(\rightarrow\) word. As a potential horizon, we suggest that TM may be applicable to the analysis of protein folds and other biomolecular structures. Such an application of NLP to what is a fundamentally geometric problem would find precedent in the pioneering development of a generative Bayesian hierarchical model for scene classification from raw image data [92].

Conclusion
In addressing the topic posed here — SB meets DS — we have considered the influences of these fields on one another, given their respective stages of maturity. SB’s rich history could positively influence the five pillars of DS (Figure 1). For example, data collection and processing often entails well-established workflows, standards and practices (e.g., structure validation), such that the results (3D structures) can be taken as ‘gold standard’ data in downstream analyses. Moreover, SB uses some approaches, like ontologies (for standardization, automated relationship discovery), that are not as prevalent in DS, but which could enjoy broader application. Notably, the data-access and software-sharing policies that have evolved in SB communities for decades can serve as positive models for DS.

Conversely, DS is being driven by economic, political and social factors that reach far beyond science itself (technology, commerce, etc.), and which inspire scientists to further innovate across the five pillars of DS. The links to SB are many and varied, and here we have touched on but two of them (ML more broadly, and NLP in particular). Details of ML and NLP approaches are beyond the scope of this work, which has only sought to briefly sketch potential synergies between these DS-based approaches and SB.

We answer our question, then, with a resounding yes: DS is already impacting SB, and we argue that the converse could be true, too. The rate of change is less certain, but is clearly steep: Figure 2, based on the recent biosciences literature, suggests that we are at the cusp of a major impact. Realizing the full benefits of this gold-rush moment will require more multidisciplinary training of students, support from scientific bodies, deep funding and, most importantly, a general willingness by the respective scientific communities. These are interesting times indeed.

Conflict of interest statement
The authors declare no conflict of interest.

Acknowledgements
We thank D Mietechen, S Veretnik and Z Zhao (UVA) for helpful discussions and critical reading of the manuscript. Portions of this work were supported by the University of Virginia (PEB) and National Science Foundation Career award MCB-1350957 (CM).

Appendix A. Supplementary data
Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.sbi.2018.09.003.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
-° of outstanding interest

9. A recent, succinct overview of the motivation, opportunities, and data-related challenges that arise in integrative structural modeling.
11. Kelleher JD: This recent text, from MIT’s Essential Knowledge series, offers a clear and concise, yet wide-ranging and highly informative, introduction to DS. Tierney B: Data Science. The MIT Press; 2018.

Articulates many best practices for data, including examples familiar to many SB readers (e.g., wwPDB, UniProt) as well as less familiar concepts (e.g., ‘machine-actionable’ data formats).


A recent, pedagogical review of cloud computing that covers many salient issues, including the benefits and drawbacks to this HPC approach in biosciences research.


This review is both thoroughly detailed and vast in scope, ranging from the visualization of quantum mechanical calculations (few dozen atoms) to microtubule bundles (billions of atoms).


This comparative analysis of the performance (storage, indexing, retrieval, etc.) of various database approaches illustrates some of the challenges and opportunities posed by data in the biosciences.


A recent, authoritative review of network analysis that is wide-ranging (beyond biomedical DS), highly detailed, and lucidly presented.


This ambitious effort illustrates how DS approaches (machine learning, feature selection, etc.) can integrate sequence- and structure-based information to map cryptic binding sites in human proteins.


A review of deep learning for PLIs that complements the present work, particularly in its lucid coverage of methodological issues (model representations, benchmarks, algorithms).


One of the first reports of 3D-CNNs applied to PLIs; methods for voxelization and DL algorithms are described in a clear and straightforward manner.


64. Goodfellow I, Bengio Y, Courville A: This recent text has become a modern classic on deep learning (and CNNs more generally); it includes chapters on mathematical background, as well as treatments of CNNs and other advanced/specialized topics. Deep Learning. Cambridge: London: MIT Press; 2017.


