Synthese (2025) 205:69
https://doi.org/10.1007/s11229-025-04911-0

ORIGINAL RESEARCH

Effective integration and models of information: lessons from integrative structure modeling

Agnes Bolinska[1] · **Andrej Sali**[2,3]

Received: 6 January 2024 / Accepted: 3 January 2025

Abstract
Integrative structure modeling is a method for using information from multiple sources to compute structural models of biomolecular systems. It proceeds via four steps: (i) defining the model *representation*, which determines the variables whose values will be computed; (ii) constructing a function for *scoring* alternative models according to how well they accommodate input information; (iii) *searching* a space of candidate models for acceptable models; and (iv) *analyzing* acceptable models to evaluate their fit with input information. These steps are iterated until a model adequate for addressing biological questions is found. In this paper, we draw lessons from integrative modeling about effective integration and about modeling. We describe what it means to integrate information from multiple sources: Integration amounts to distributing information among the four steps of integrative modeling. Theory and data alike can be sources of information; this process thus generates *models of information*, rather than models of theory or models of data. We then propose heuristics for distributing information and designing multiple iterations of modeling. Effective iteration requires prioritizing the most reliable information and minimizing the time required to obtain an adequate model. Rather than being constructed from theory and assessed using data, models are constructed from any available information and assessed in a coherentist manner.

Keywords Integration · Modeling · Information · Theory · Data · Structural biology · Evidence

1 Introduction

Modeling complex systems is challenging. Modelers typically have at their disposal a variety of empirical data, targeting different aspects of the target system or a different but related system. They must also take into account background theory. Empirical

Extended author information available on the last page of the article

Published online: 05 February 2025

data, background theory, and anything else that might be relevant to understanding the system are sources of *information*, which can be used as inputs for modeling. However, some pieces of information are less reliable than others. Data can be noisy; extrapolation from an experimental to a target system might be unwarranted; how a theory ought to be applied to the system can be uncertain; and there can be several ways to interpret a given piece of information. Moreover, it is typically not possible for a model to equally accommodate all information. Every model is compatible with different pieces of information to different degrees. And in many cases, the amount of information is vast, whereas time, funding, and other resources are limited. Modelers therefore face a pragmatic problem: How should they select from among a multitude of *prima facie* possible models, each of which accommodates different pieces of information to different degrees, making best use of time and other resources?

Philosophers have yet to adequately address this problem. The principle of total evidence says that one must take into account all available evidence in the formulation of one's hypotheses (Carnap, 1947), but is silent about how to do so when pieces of information come from multiple sources, target different parts or aspects of a system, are more or less reliable, and can be ambiguous about what they are evidence for. More recently, philosophers have discussed integration in several contexts, including the integration of fields, data, and explanations. But so far, these discussions represent a promising starting rather than end point for understanding how information from diverse sources and of variable reliability ought to be integrated in complex modeling problems.

Lindley Darden and Nancy Maull (1977), for instance, argue that two fields relying upon different empirical or conceptual tools to account for different aspects of the same phenomenon can be integrated when one field fills in gaps left open by the other. For example, early-twentieth-century cytology and genetics were both concerned with the nature of heredity. Cytology studied heredity by examining chromosomes using a light microscope, whereas genetics accounted for hereditary phenomena by postulating hypothetical entities, later known as genes. The integration of cytology and genetics began when Walter Sutton and Theodor Boveri noted analogues between chromosomes and genes. Viewed through a microscope, the chromosomes of diploid organisms appeared as distinct individuals found in pairs; similarly, genes were postulated to be distinct entities occurring in pairs. This and other similarities led Sutton and Boveri to propose what Darden and Maull call an *interfield theory,* the chromosome theory of Mendelian heredity: Genes were located on or in chromosomes. This theory in turn enabled further predictions. For example, given that the number of chromosomes was smaller than the number of genes, one could predict that some genes are linked on the same chromosome, rather than assorting independently (Darden & Maull, 1977).

Darden and Maull's work has been helpfully amended and elaborated. Baetu (2011), for instance, shows that molecular biology may fill in details that are "black boxed" in mechanisms proposed by classical genetics—and that, in doing so, can improve classical explanations. Similarly, Marco Nathan (2017) argues that merely bridging an explanatory gap does not constitute the integration of two fields. Rather, the concepts of one field must also be required for explanations in the other and *vice versa*, what Nathan calls *explanatory relevance*. Although this work is primarily con-

cerned with whether two fields may be integrated without the reduction of one to the other, it is nonetheless relevant for our purposes because such integration proceeds by way of integrating information. Cytology and genetics became integrated when information from one field was brought to bear on information from the other, via a sort of cross-referencing.

Nevertheless, this work does not address the fact that there can be several plausible interpretations of information, whether it comes from data or theory. When Darden and Maull write that genes were "hypothetical entities with known functions; chromosomes were entities visible with the light microscope with a postulated function" (1977, 51), they implicitly invoke a distinction between, on the one hand, what is *hypothetical* or *postulated* and, on the other, what is *known* or *observed*. Yet such a distinction is not always so clear-cut. Neither hypothetical nor observed entities are given, but instead come with a degree of uncertainty. Theories can be mistaken, and the hypothetical entities they posit may turn out not to exist (Kitcher, 1993; Laudan, 1977). We must interpret what we see through a microscope; interpretation is often far from straightforward, since images can be noisy and our interpretations of them theory-laden and prone to other biases (Hacking, 1981; Hanson, 1958; Kuhn, 1962).

As a consequence, there are often many ways in which information can be integrated. Interpreting a piece of information in one way may make it inconsistent with another piece of information (interpreted in a particular way). Further, sometimes perfect reconciliation between interpretations may be impossible. Different interpretations may license different conclusions about a phenomenon, with the interpretation of one piece of information contradicting that of another. Rather than information lining up neatly as it does in the case of interfield theories, it is sometimes unclear how, precisely, it can be integrated. This point is especially important in the age of big data, when there is orders of magnitude more information available, and therefore many more plausible candidate integrations; limitations of resources such as time and funding must be taken into consideration in this context. Darden and Maull (1977) and cognate approaches to interfield theories do not address the possibility of multiple ways of integrating information, and the pragmatic constraints that preclude considering all of them.

What about the literature more closely related to the aims of this paper, addressing the integration of data, explanations, and methods?[1] This literature discusses ample examples of such integration, acknowledging the variety of sources of data, methods of their generation, and forms in which they are presented. However, two lacunae remain. First, these examples are presented at a coarse grain; detailed descriptions of what such integration consists in or how it takes place are absent. Second, normative guidance for how to conduct integrative research effectively—and in particular, how to select among competing ways of integrating information—is also not offered. Anya Plutynski, for instance, points to several instances of integration in the study of carcinogenesis, writing that "familial data was integrated with subsequent work on the rates and character of retinal development" and that "work on the character

[1] See, for instance, Brigandt (2010, 2013a, 2013b), Leonelli (2013, 2016), MacLeod and Nersessian (2013), Mitchell (2002, 2003, 20092019), Mitchell and Gronenborn (2017), O'Malley and Soyer (2012), and Plutynski (2013, 2018).

of tissue renewal in the colon has informed understanding of how carcinogenesis develops in that tissue" (2013, 471). But she does not further describe what is meant by "integrating" or "informing" in these instances. Similarly, Sabina Leonelli suggests that "new data types, such as data about how *Miscanthus* behaves in the field, can be usefully integrated with data about *Arabodopsis* metabolism, resulting in new knowledge about how plants produce energy in both species" (2013, 509). Yet she too does not specify further what plant scientists actually do when they "integrate" this data. Neither Plutynski nor Leonelli provide normative accounts of what makes integrative strategies effective in their respective case studies.

Filling these lacunae is important. The value of integrative research strategies has been widely acknowledged, both by scientists and philosophers. Evolutionary systems biologists Michalodimitrakis and Islam, for instance, put it this way: "the interplay of modeling and experiments takes the research much further than either approach on its own" (2009, 28; quoted in O'Malley & Soyer, 2012, 64). A task for philosophers is to explain why. In virtue of what is integrative research superior to using a single method, dataset, or explanation alone? That is, what are the epistemically significant features of integrative research? Merely identifying some instances of integration in various scientific practices is not enough to answer these questions. Instead, we need to examine the mechanics of successful integration, that is, to gain an understanding of what, precisely, such integration consists in.

Understanding the mechanics of *successful* integration is key. Implicit in the widespread acknowledgement of the value of integrative research is the assumption that the research has been conducted in a principled way, with methodological rigor and care. We can imagine cases of integrative research in which methods, datasets, or explanations are integrated, but in an ad hoc, unprincipled, or sloppy manner. It would be hard to argue that such research would be superior to obtaining scientific knowledge from a single method, dataset, or explanation alone.

In this paper, we thus provide a detailed descriptive and normative account of integration by examining the practice of *integrative structure modeling* ("integrative modeling" for short) in structural biology. Structural biologists aim to construct structural models of biomolecular systems, i.e., to determine the relative positions and orientations of components such as atoms, residues, and secondary structure elements. For smaller systems like single proteins, they can use traditional methods like X-ray crystallography and NMR (nuclear magnetic resonance) spectroscopy to do so. However, for larger systems consisting of hundreds of macromolecules, such methods are insufficient. This is especially so given that biologists often want to understand these systems' dynamics—how they assemble and disassemble—and their functions—how they interact with other systems—to gain insight into their evolutionary history or develop medical therapies. To construct dynamic structural models of large systems, a variety of disparate information, coming from multiple experiments, physical theories, statistical analyses, and previously determined models, is required.[2] For example, the yeast Nuclear Pore Complex (NPC), a ~ 52-MDa channel mediating the exchange of small ions across the nuclear membrane, consists

[2] Structures of sufficiently large and complex systems cannot be currently solved using structure prediction relying on machine learning, such as AlphaFold (Callaway, 2020).

of ~ 550 protein subunits of ~ 30 different types. Constructing a structural model of the NPC required information from multiple sources, including X-ray crystallography, small angle scattering, and NMR spectroscopy, as well as comparisons to previously determined structures and molecular mechanics force fields (Alber et al., 2007; Rout & Sali, 2019; Sali, 2021).

Integrative modeling is a method for integrating all such information.[3] It proceeds via four steps: first, the model *representation*, which determines the variables whose values are to be computed by modeling, is defined; second, a function for *scoring* alternative models according to how well they accommodate input information is constructed; third, acceptable models—those that accommodate input information sufficiently well—are identified by *searching* a space of candidate models; finally, these models are *analyzed* to evaluate their precision and fit with input information. This process is iterated until an ensemble of acceptable models precise enough for addressing biological questions is found (Alber et al., 2007; Rout & Sali, 2019; Sali, 2021). We describe the mechanics of integrative modeling, showing what it means to integrate information using this method. We further offer normative guidance, proposing heuristics for effective integrative modeling.

Although we focus primarily on integrative modeling as a vehicle for integration, our analysis also offers a fresh perspective on modeling. In particular, we provide an alternative to the hierarchical view of models (Giere, 2010; Mayo, 1996; Suppes, 1962). The hierarchical view was developed to address a puzzle: Scientific models are abstract and idealized, whereas the systems they represent are concrete and, moreover, often not directly accessible. How, then, can we determine whether our models accurately represent these systems? The solution involves a hierarchy of different kinds of models that forges a connection between our abstract and idealized models and the concrete systems they represent. On this view, we construct our models solely on the basis of theory; they are thus referred to as *models of theory*. Although we cannot access the systems these models represent directly, we can gather data about them via experiment or observation. We may then identify patterns among these data, such as functions that accommodate them sufficiently well. In other words, we construct distinct models from the data called *models of data*. We assess our models of theory by comparing them to our models of data.[4]

According to the hierarchical view, theory and data have distinct functions: Theory is used for model construction; data is used for model assessment (via data models). In contrast, we will show that information from any source—including theory and data—can be used for model construction. Rather than models of theory or models of data, integrative modeling thus generates *models of information* (Bolinska, forthcoming). We further show that these models are assessed in a coherentist manner, according to how well they accommodate *all* such information.

The paper proceeds as follows. Section 2 describes the aim of integrative modeling as producing models of information, and the mechanics of integration, captured

[3] In discussing integrative modeling, we therefore extend Sandra Mitchell's (Mitchell, 2019, 2020; Mitchell & Gronenborn, 2017) work on integrating X-ray crystallography with NMR spectroscopy using joint refinement.

[4] The hierarchy of models also includes other elements (e.g., models of experiment), which we omit here.

by its four-step, iterative workflow. Section 3 proposes heuristics for effective integration in terms of how to distribute information among the steps of modeling and how to design each subsequent iteration of these steps. Section 4 draws lessons from our case study, showing how it moves philosophical discussions about integration and modeling forward. Section 5 concludes, highlighting parallels between integrative modeling and integrative research in other domains, showing how our account can help us to better understand and perhaps even improve such research.

2 The mechanics of integration in integrative modeling

What does it mean to "integrate" information? In this section, we describe the mechanics of integration in integrative structure modeling. We begin by discussing input information and the aim of integrative modeling as producing models of information (Sect. 2.1). Then, we show how information is converted into a structural model via its four-step, iterative workflow (Sect. 2.2).[5]

2.1 Input information and the aim of integrative modeling

Input information for integrative modeling can come from several types of sources. First, it can be determined by various experimental techniques, including X-ray crystallography, NMR spectroscopy, chemical cross-linking, and genetic interactions. Second, information can come from physical theory, such as a molecular mechanics force field that specifies preferred stereochemistry and non-bonded interactions. Third, it can come from statistical sequence-structure patterns, extracted from a large set of previously determined protein structures deposited in an online repository called the Protein Data Bank (Berman et al., 2000). Fourth, information can come from prior models, such as structures of subunits in a complex to be modeled. Finally, information can come from a scientist's intuition, hypothetical reasoning, or even guesswork. For the purposes of this paper, then, "information" should be understood broadly, as anything that might constrain which models are acceptable.[6]

Different pieces of information can target different parts of a system (e.g., different proteins in a complex of proteins) or different aspects of the same part (e.g., positions and distances between atoms). Further, our confidence in how we take information to constrain structural models can vary. For instance, we might be very confident that a particular segment of a protein sequence folds into an alpha-helix, given that most such sequence segments in other proteins are helical; in contrast, we might be entirely uncertain about a guess that two proteins in an affinity co-purified complex of mul-

[5] Due to space constraints, we highlight only those elements of integrative modeling most pertinent to this paper's aims. For a more comprehensive description of integrative modeling, see Rout and Sali (2019) and Sali (2021).

[6] Information in our sense is broader than Leonelli's notion of data as "any product of research activities, ranging from artifacts such as photographs to symbols such as letters or numbers, that is collected, stored, and disseminated *in order to be used as evidence for knowledge claims*" (2016, 77; see also Leonelli (2013)); it includes theory and educated guesses, neither of which are direct products of research activities or evidence per se.

tiple proteins contact each other (Rout & Sali, 2019; Sali, 2021). That is, some pieces of information are more *reliable* than others, where a piece of information is reliable to the extent that we can be confident that it constrains structural models correctly.

There are several sources of uncertainty in information, each of which can limit our confidence that a piece of information constrains structural models correctly. First, information can be sparse: There can be more degrees of freedom in the model than data points. If sparseness is not given due consideration, there is a risk of overfitting the data, proposing just one or a small handful of models when there may in fact be more that are equally consistent with input information. Second, information is subject to random and systematic error. For example, in X-ray crystallography, random error can come from noise in the X-ray flux and detector, systematic error from radiation damage or irregularities in the crystal. Third, information can be ambiguous. For example, it is generally not possible to determine which of three methyl protons is responsible for producing a given nuclear Overhauser effect (NOE) signal in solution NMR (Schneidman-Duhovny et al., 2014). Evaluating information with respect to these sources of uncertainty, then, enables assessing its reliability.

Integrative modeling aims to generate models that accommodate all input information, where a model accommodates a piece of information to the extent that it is consistent with that information. For instance, suppose an experimental observation indicates a distance of less than 4 Å between two particular atoms. A model that locates those two atoms 3.5 Å apart accommodates this information better than one in which they are 4.5 Å apart. Because no model accommodates all information perfectly, the aim of integrative modeling is to generate models that accommodate all information *sufficiently well*. These models are thus akin to models of data positing a relationship among data in a given dataset, for instance, by fitting a curve to them. The difference is that information from *any* source can be used as an input for their construction; we may thus conceive of them as *models of information*.

Since there is typically more than one model that accommodates all information sufficiently well, integrative modeling aims to generate an *ensemble* of such models. The model ensemble—rather than an individual instance of a model within the ensemble—should be regarded as the desired outcome of the modeling process. It captures the uncertainty in the information arising from the sources discussed above.[7] Structural biologists often refer to the model ensemble as "the model," and we will follow suit where appropriate.

In addition to accommodating information sufficiently well, models must also be sufficiently complete, detailed, and precise for answering biological questions of interest. A model is sufficiently complete if it includes the part of the modeled system needed to answer the questions; for example, a question about enzymatic catalysis likely requires a model of an enzyme to include the active site. A model is sufficiently detailed if it depicts the modeled system with the granularity needed to answer the question; for example, a question about enzymatic catalysis likely requires a model that specifies atomic positions, rather than representing amino acid residues with

[7] This is different from cases in which multiple models are required because each makes different, often incompatible, idealizing assumptions (cf. Dickson, 2006; Fehr, 2006; Mitchell, 2002, 2003; Morrison, 2011, 2015; Weisberg, 2007, 2013).

coarse-grained beads. Finally, a model ensemble is sufficiently precise if the variability among models in the ensemble is sufficiently low; for instance, an ensemble of atomic models that vary on a scale larger than the size of an atom is unlikely to be sufficiently precise for questions about catalysis.

The outcome of successful integrative modeling is a sufficiently precise ensemble of acceptable models, each of which is complete and detailed enough for answering biological questions. Henceforth, we omit these qualifications, referring to such an ensemble simply as an "ensemble of acceptable models."

2.2 The integrative modeling workflow

Generating an ensemble of acceptable models is an iterative process that proceeds via four steps: defining the model representation, scoring a model against input information, searching for acceptable models, and analyzing the model (Fig. 1).

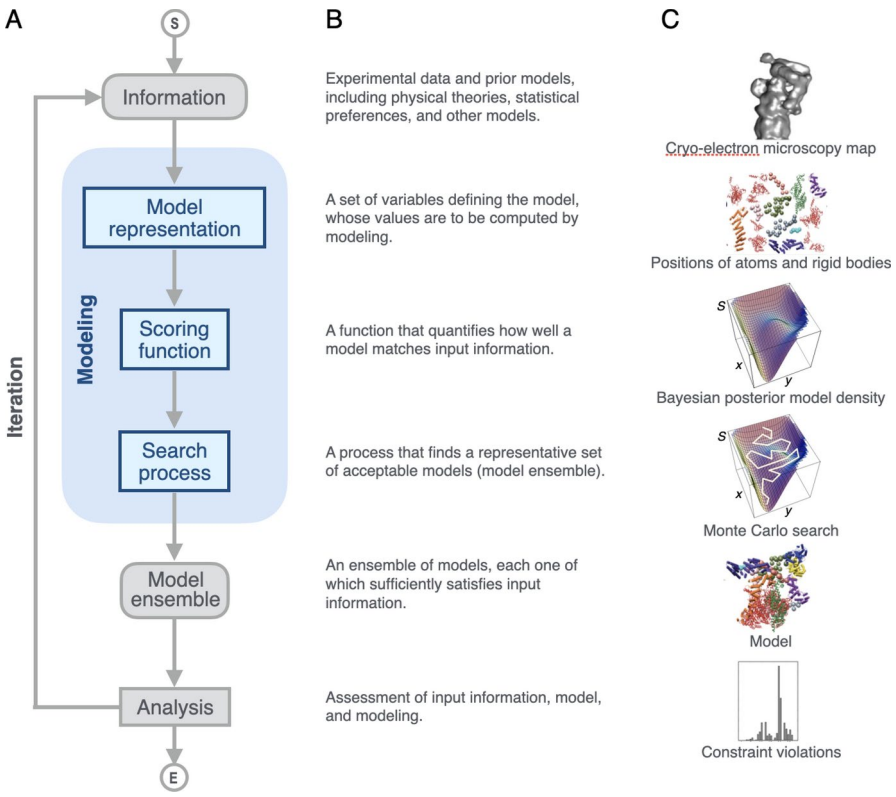

Fig. 1 Integrative structure modeling workflow. **A** Integrative modeling is an iterative process that converts input information about a biomolecular system into its structural model. The light blue rectangle indicates a single instance of model construction. **B** Each aspect of integrative modeling is defined. **C** Each aspect of integrative modeling is illustrated with an example

2.2.1 Defining the model representation

The first step of integrative modeling is to use some of the available information to define the model representation, which specifies the mathematical variables whose values will be determined by modeling and their allowed range. The most important model variables are typically positions of system components. For instance, a common aim is to determine positions of individual atoms. However, sometimes there is not enough information to do so, and a set of atoms can be represented by a larger sphere, such as a coarse-grained bead corresponding to an amino acid residue or protein subunit. System components can also be fixed with respect to each other into a rigid body corresponding to a previously determined structure. Further, some samples contain a mixture of structures; characterizing them may necessitate a model representation with separate sets of coordinates for each structure in the mixture, together with the structures' relative concentrations. Although positions of model components are most important, other model variables can also be included in the model representation.[8]

The number and nature of the variables that comprise the model representation determine how well a model can accommodate input information. In general, a model can better accommodate input information if its representation includes a greater number of co-existing structural states, fewer rigid bodies, and higher resolution particles—a point that will be important in the heuristics we propose in Sect. 3.3.

2.2.2 Scoring a model against input information

The model representation effectively defines a space of in-principle possible models, with each model in the space specifying values for each of the model variables. The next step is to determine how consistent each model is with the input information. The assumption is that models that are more consistent with more of the information are more likely to be correct. Thus, information can be used to construct a scoring function quantifying the match between a model and the input information and to compute its value.

Most commonly, a least-squares scoring function is used, corresponding to a weighted sum of spatial restraints:

$$S = \mathbf{\Sigma}_i \omega_i (X_i - X_i^o)^2,$$

where the sum runs over all spatial restraints i, X_i is the value of a restrained spatial feature in a model, X_i^{o} is its measured value, and ω_i is the weight of the restraint. Minimization of S will by design minimize the difference between the model and available information. Each restraint thereby quantifies the deviation of a computed property of a model from that specified by the input information. For example, a restraint (i) based on an NMR spectrum may compare the distance between two specific atoms in a model (X_i) with an experimental observation that this distance is

[8] For example, isotropic temperature factors indicating the fluctuations of atoms around their average positions in crystallography are often included.

less than 4.5 Å (X_i^o), weighted by our relative confidence in the measurement (ω_i). Summing over all such restraints, weighted according to our relative confidence in them, gives us the value for the scoring function S.[9] The scoring function enables us to determine which models are *acceptable*, where acceptable models are those that accommodate input information to a sufficient degree.

2.2.3 Searching for acceptable models

The space of in-principle possible models must then be searched to find all acceptable models—those that are sufficiently consistent with input information, as quantified by the scoring function.[10] In principle, the best search is a systematic enumeration that generates every possible model one by one with sufficient granularity. However, enumeration is rarely computationally feasible, given the size of biomolecular structures and the precision required to enumerate them. So stochastic sampling methods, such as various Monte Carlo schemes, can be used instead. These methods rely on heuristics that bias the search toward models that are more likely to be acceptable, without enumerating all models. They aim to map the shape of the scoring function landscape as a function of all model variables.

Some of the available information can be used to constrain the model space that is searched. For example, we can limit the search for positions of a membrane protein to the membrane or limit sampling to a single symmetry unit of a system (Kim et al., 2018). All else being equal, the smaller the search space, the more computationally feasible the modeling task.

2.2.4 Analysis of the model

Any candidate models generated by the search process are next analyzed to determine whether they satisfy all input information and are complete and detailed enough for answering biological questions. A model's failure to satisfy all input information—its inconsistency with some data, theory, statistical preference, or prior model—indicates that something has gone awry, though it is not always clear where the problem lies. Perhaps the model representation is incorrect or the input information is not as reliable as was assumed, provided that the scoring function form and its parameter values are accurate.

The analysis step terminates the modeling process once an ensemble of acceptable models is found. Otherwise, modeling proceeds via another iteration, the aim of which is to eliminate the source of inconsistencies. Is the problem with (some of) the information? Or is it instead with (some of) the ways in which it was used, that is, how it was interpreted with respect to the target system or apportioned among the steps of modeling? The answers to these questions have a normative dimension; we therefore reserve discussion for how to address them for Sect. 3.3.

[9] The best possible scoring function, however, is the Bayesian posterior model density, because it specifies the probability of a model M given information I, P(M/I) (Rieping et al., 2005).

[10] Strictly speaking, there might be infinitely many acceptable models because variables could be defined to arbitrarily many significant figures; thus, only a *representative* sample of models need be determined.

3 Heuristics for effective integrative modeling

In the previous section, we described the integrative modeling workflow, beginning by clarifying the notion of input information and the aim of integrative modeling, and then explicating the four steps of this process. In doing so, we have given an example of what integration consists in—of the mechanics of integration—in the context of structural biology. Rather than merely stating *that* information from multiple sources becomes integrated, we specified *how* such integration takes place: via the application of different pieces of information in each of the four steps of modeling. Information from one source might be used in defining the model representation, while information from another is used for constructing and evaluating the scoring function. Still other information can be used to guide the search through the space of in-principle possible models or to analyze the models generated by the search. We find, then, that modelers must make decisions about how to use different pieces of information in the integrative modeling workflow. We will show that some ways of doing so are better than others. In this section, we address the question: What does it mean to conduct integrative modeling effectively?

3.1 The efficiency of integrative modeling

We will understand effective integrative modeling in terms of efficiency. Normally, several iterations of modeling are required to compute an ensemble of acceptable models. Within an iteration, input information can typically be integrated in more than one reasonable way: There may be several conceivable model representations, interpretations of information, scoring functions, or searching algorithms. Yet decisions about how to conduct each step of the modeling process affect which model ensemble (if any) is constructed: Different model ensembles will result from different such decisions. For a typical modeling problem, it would take more time and computational resources than available to combine and recombine information into every possible permutation to see which models result and evaluate them. There is a limit to how many such permutations can even be considered.It is therefore imperative that integrative modeling be conducted *efficiently*. "Efficiency" is sometimes used pejoratively, suggesting corner-cutting or insufficient rigor. Here, we are instead concerned with maximizing the efficiency of rigorous investigations conducted according to explicit standards.

The efficiency of integrative modeling is a function of two factors: how many iterations are required to obtain an ensemble of acceptable models and how much time each iteration takes. We characterize an iteration of modeling as spanning the time from completing the previous iteration to the moment before starting the next one. Understood this way, each iteration includes not only the four steps of modeling, but also any time for deliberation before implementing these steps and any further information-gathering—for instance, by conducting more experiments—in preparation for the next iteration. This characterization of time per iteration enables us to define the efficiency of integrative modeling as the sum of iteration times.

We are now ready to put forward our normative proposal, our heuristics for maximizing the efficiency of integrative modeling. We begin by considering how informa-

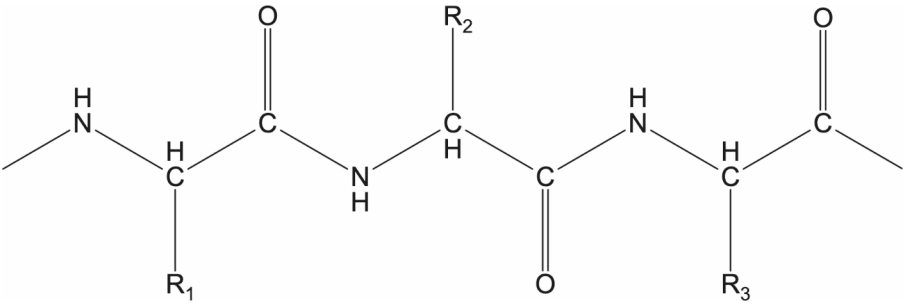

Fig. 2 The general structure of the polypeptide chain. R groups represent side chains that differ for different amino acid residues

tion should be distributed among the steps of modeling (Sect. 3.2). Then, we consider the role of iteration in effective integrative modeling (Sect. 3.3).

3.2 Using information effectively

In this section, we argue that it matters which piece of information is used for which step of modeling: Some ways of distributing information can, on average, generate an ensemble of acceptable models within fewer iterations. We do so by extending an analysis of Linus Pauling's determination of the structure of the folded polypeptide chain (Pauling et al., 1951), proposed by one of us (Bolinska, 2018).

3.2.1 Pauling's heuristic for determining the structure of the folded polypeptide chain

Bolinska (2018) understands the process of determining the structure of the folded polypeptide chain as a stepwise narrowing-down of a space of candidate models. The structure of the extended polypeptide chain had been established by the time Pauling was working toward determining its folded structure (Fig. 2). The space of candidate models therefore contained all of the ways in which the polypeptide chain might fold. Different pieces of information—either theoretical considerations or experimental data—could guide the elimination of portions of this possibility space. Stereochemical rules dictating energetically favorable molecular conformations and X-ray diffraction photographs of the protein keratin were especially important pieces of information. Which information should be considered first? Pauling chose stereochemical rules. He determined the structure of the folded polypeptide chain from known bond lengths and angles, representing the side chains of amino acid residues as R groups. Only after he determined a structure compatible with this information did he consult X-ray diffraction data.

Bolinska (2018) explains why his heuristic was successful. Stereochemical rules were highly confirmed theoretical considerations; it was unlikely that the structure of the folded polypeptide chain violated them. In contrast, X-ray crystallography was a relatively new experimental technique. X-ray diffraction photographs were blurry and seemed compatible with several interpretations. As a consequence, a model that

appeared to be incompatible with an X-ray diffraction photograph might nonetheless be correct. Pauling's alpha helix, for example, was incompatible with an influential X-ray diffraction photograph of the protein keratin taken by William Astbury, who interpreted it as indicating a repeating subunit every 5.1 Å (Astbury & Street, 1932).

Bolinska (2018) argues that considering stereochemical rules before X-ray diffraction photographs was a better heuristic than considering the latter before the former because it warranted greater confidence that, with the consideration of each piece of information, portions of the possibility space were eliminated *correctly*. Narrowing down the space of possible structures correctly is crucial: Mistakenly eliminating the right structure from the possibility space would necessitate starting the process over and determining where one went wrong. This heuristic minimizes how many iterations are, on average, required to get the right solution by reducing the likelihood of having to backtrack.

3.2.2 Applying Pauling's heuristic: prioritizing the most reliable information

Pauling's heuristic originates from a more general principle: Modelers should rely most heavily upon the most reliable information. Applying this principle to contemporary integrative modeling tells us to use the most reliable information to define the model representation and to guide searching. Because the model representation specifies which models are under consideration in the first place, it is impossible to find a model whose variables are not included in the representation. Defining the model representation incorrectly is thus a significant error: it precludes, from the outset, finding the correct model(s). Indeed, we might understand Pauling as effectively having defined the model representation when he took for granted the structure of the extended polypeptide chain, attempting to fold it (rather than some other structure) into an energetically favorable conformation. In the framework of integrative modeling, we might say that the variables he included in the model representation were positions of different polypeptide chain components, some represented at the atomic level (e.g., atoms in the polypeptide backbone), others at a coarser grain (e.g., side chains represented as R groups) (Fig. 2).

A similar argument applies to the searching step. Using information in searching delimits the range of values the model variables can adopt. Just as we cannot find a model whose variables are not included in the model representation, we also cannot find a model where we do not search. A historical case can illustrate this point. In an attempt to determine the folded structure of the polypeptide chain that preceded Pauling's, Sir Lawrence Bragg, John Kendrew, and Max Perutz (1950) listed twenty possible structural models, selecting from among them those that were compatible with the 5.1-Å repeat indicated by Astbury's photograph. We can understand Bragg, Kendrew, and Perutz as having used information from that photograph to guide their search through the space of possible structures. This turned out to be a mistake. Their proposed structure violated a stereochemical rule: It allowed rotation about the peptide bond, which has partial double-bond character and is therefore planar. Meanwhile, Astbury's interpretation of his photograph as indicating a 5.1-Å repeat was eventually found to be mistaken (Judson, 1996; Olby, 1974).

In contemporary integrative modeling, using unreliable information to guide searching can also mislead. For instance, using the information that the NPC has eightfold symmetry for searching means that we only search a single symmetry unit and assume that the structure is replicated accordingly (Kim et al., 2018). Since doing so precludes the possibility of finding a structure without such a symmetry, only the most reliable information should be used for this step.

Information in which we have intermediate confidence is best suited for scoring. The scoring function enables us to quantify our confidence by weighting the importance of accommodating information relative to its reliability. For example, an NMR NOE data point restrains the maximal distance between a pair of atoms, via an upper distance bound term in the scoring function, and an electron microscopy density map restrains the shape of a model, via a correlation coefficient between the map and model. The weights we select for these terms in the scoring function reflect our relative confidence in the corresponding data.

Information in which our confidence is low may best be left out of model construction altogether, and instead reserved for analysis. For instance, suppose we know that mutating a residue in a protein prevents it from forming a complex with another protein. This information can be interpreted in two ways: either the mutation prevents the formation of the complex because it is located in the interface between the proteins, or it modifies the interface through allostery, and is located elsewhere. Given uncertainty about which of these interpretations is correct, it is unclear what conclusions this information warrants (Kaake et al., 2021). Reserving this information for analysis makes the construction of models inconsistent with it possible. If those models are indeed found to be inconsistent with the information, we have a ready explanation for the inconsistency: There is a problem with (our interpretation of) the information, rather than the models. However, we may find that the models we construct without this information are consistent with it. In this case, we may use the information differently in a subsequent iteration of modeling—a point we discuss in the next section.

3.3 Iterating effectively

In the previous section, we showed that the distribution of information among the steps of integrative modeling can reduce the number of iterations required to obtain an ensemble of acceptable models. We then introduced heuristics for how to distribute information *within* an iteration of modeling. We now discuss various functions of iteration and provide heuristics for iterating effectively.

One such function is to refine models, making them more precise. At the end of the previous section, we argued that unreliable information should be reserved for analysis so that it cannot influence model construction. We suggested that a model's inconsistency with that information could be readily explained, given the information's unreliability. However, we also noted that a model may instead turn out to be consistent with the unreliable information. In such a case, a process of refinement can be initiated. For example, one piece of information available for modeling the Nup84 complex was data from X-ray crystallography, suggesting a particular interface between two proteins. However, differences between the crystallographic

experimental context and a protein's native environment sometimes preclude extrapolating from experimental results to that environment. Researchers thus wondered whether the crystallographic interface reflected how these proteins come together in vivo (Fernandez-Martinez et al., 2012). As in the case that we considered at the end of the previous section (Kaake et al., 2021), it was unclear what conclusions were warranted: the proteins could come together in the same way, but they might not. Researchers thus constructed models without this information, reserving it for analysis. When analysis revealed the models to be consistent with the crystallographic data, researchers' skepticism about whether the crystallographic interface reproduced cellular conditions was eliminated. Their newfound confidence justified using the crystallographic data in model construction in subsequent iterations, resulting in a more precise model (Fernandez-Martinez et al., 2012). When iteration is used for refinement, the outcome of earlier iterations of modeling warrants the redistribution of information in subsequent iterations.

However, in addition to determining how to apportion information among the steps of modeling, modelers must also make other decisions. For instance, they must decide how *accommodating* the model representation should be, where a more accommodating model representation is one that has a greater number of variables (i.e., is more fine-grained, flexible, or has more states; see Sect. 2.2.1). How should they do so? We suggest that modelers should err on the side of a modeling protocol that finds fewer rather than more acceptable models, preferring a less accommodating model representation in earlier iterations of modeling. Recall that the goal of modeling is to find a sufficiently precise ensemble of acceptable models, each of which is sufficiently complete and detailed for answering biological questions (Sect. 2.1). In practice, completeness and detail can be established by defining a model representation that contains the required components at a sufficient level of granularity for answering biological questions. Therefore, at the end of an iteration of modeling, the remaining task is to determine whether the ensemble of acceptable models is sufficiently precise to be useful; this outcome constitutes the aim of modeling, terminating the iterative process.

However, there are two ways in which modeling can fall short of this aim: by finding no acceptable models, or by finding an ensemble of acceptable models that is insufficiently precise for answering biological questions. Further, before conducting an iteration of modeling, a modeler does not know whether the iteration will end in an ensemble of acceptable models that is sufficiently precise to be useful; no acceptable models; or an insufficiently precise ensemble of acceptable models. Given this uncertainty about the outcome of an iteration of modeling, if one way of falling short of the aim of modeling is better than the other, modelers ought to err on the side of that outcome. This reasoning is familiar from the argument from inductive risk. Scientific reasoning is inductive and therefore prone to error. There are two kinds of error a scientist can make: concluding that there is an effect when in fact there isn't one (i.e., a *false positive*) or concluding that there is no effect when in fact there is one (i.e., a *false negative*). When they test hypotheses, scientists must set significance thresholds. According to the argument from inductive risk, where they set these thresholds biases the direction in which they err, either toward false positives or false

negatives.[11] A more general principle can be extracted: When we make decisions under uncertainty, we should consider the consequences of different ways of erring, rather than simply aiming for our desired outcome.

We can apply this principle to integrative modeling as follows. One way of erring, finding an insufficiently precise ensemble of acceptable models, would generally necessitate performing more experiments, which are typically more time-consuming and costly than modeling. The only way to increase the precision of the output model is to use more information in the input; achieving this goal often requires more experimental resources. In contrast, finding no acceptable models merely requires conducting further iterations of modeling, which is generally quicker and cheaper. Our guiding principle for maximizing efficiency, then, is that, all else being equal, modelers should err on the side of finding fewer acceptable models. Erring in this direction helps to minimize the average time required for an iteration of modeling.

Erring on the side of fewer acceptable models tells us to adopt the least accommodating conceivable representation given the desired use of the model. If a less accommodating representation enables us to find an ensemble of acceptable models, we find out immediately. Further, less accommodating representations increase the efficiency of searching for acceptable models because they have fewer variables, generally resulting in fewer models in the search space. All else being equal, an iteration that includes a less accommodating model representation takes less time than one that includes a more accommodating representation.

Adopting the least accommodating conceivable representation can initiate a process of self-correction (Chang, 2004), whereby the representation is made more accommodating until a sufficiently precise ensemble of acceptable models is found. For example, in modeling the structure of the yeast Spindle Pole Body, researchers were uncertain whether one component formed an extended rigid rod or there instead existed a pivot point at which the sequence folded upon itself. Based on results from X-ray crystallography, they began by assuming it to be an extended rigid rod, a less accommodating representation. They were unable to find a model that satisfied all information, so they next chose a model representation with a pivot point. Adopting this more accommodating representation enabled them to find a model satisfying all information (Viswanath et al., 2017).

Adopting the least accommodating representation can also initiate a process of trial-and-error. The aim of integrative modeling is to use all available information in the determination of the structure of a biomolecular system. But as it is currently implemented in software, there is a limit to what information can be included in a single iteration. For example, researchers possessed data indicating that the SEA complex contained either one or three copies of protein subunits Sea4 and Seh1, but it wasn't clear which possibility was more likely. They started by including just one subunit in the model representation (erring on the side of finding fewer acceptable models). When no acceptable models were found, they considered a representation with three subunits in the next iteration, producing a model that satisfied all information (Algret et al., 2014).

[11] The argument from inductive risk further asserts that such thresholds cannot be set without considering non-epistemic values (Douglas, 2000).

More generally, when no acceptable models are found, either there are problems with one (or more) of the first three steps of modeling or the information is less precise than expected. Given this indeterminacy, a modeler should begin by assessing whether a problem might have arisen in one of the first three steps of modeling. The alternative, performing more experiments to help interpret information and its precision more accurately, would require more time per iteration.

4 Lessons from integrative modeling

Let us take stock. We began with the observation that modeling complex systems poses significant challenges. Different information targets different parts or aspects of the modeled system and some pieces of information are more reliable than others. There are often many possible ways to integrate information, and pragmatic limitations preclude considering all of them. How, then, should a modeler proceed? We examined integrative modeling in structural biology as a case study for answering this question. We described its mechanics, showing how integration takes place via its four-step, iterative workflow, and offered heuristics for integrating information effectively.

Integrative modeling provides us a rich source of insight, both about integration (which need not involve modeling) and about modeling (which need not involve integrating information from varied sources). We therefore draw some lessons about integration and modeling from our analysis.

4.1 Understanding integration

Our primary goal is to use integrative modeling as a foothold for better understanding integration. What is the value of integrative research? As many have noted, we need integrative research for sufficiently complex problems because they resist solution by traditional means. For example, James Griesemer shows that David Wake's integrative approach to evolutionary biology enabled him to solve problems that "would not have been solved so well by non-integrative approaches" (2013, 525). Similarly, Ingo Brigandt argues that it is "necessary to integrate different theoretical models and modes of explanations" to account for the evolutionary origin of novelties (2010, 304). Putting multiple pieces of information together gives us a picture of the whole that would be difficult or impossible to obtain otherwise. Integrative research can enable, for instance, the localization of errors via triangulation (MacLeod & Nersessian, 2013) or the coordination of different causal models, sometimes operating at different levels, resulting in a fuller explanation of target phenomena than each could provide on its own (Bechtel, 2013; Mitchell, 2009; Plutynski, 2013). Integrative research is thus "vital to some instances of success in science" (Brigandt, 2013b, 461).

But our case study reveals that the mere necessity of integration for solving complex problems doesn't wholly account for its value. Unlike pieces of a jigsaw puzzle that only fit together in a single way, there are often many ways in which information from various experimental and theoretical sources can be integrated, since there can

be several plausible interpretations of information and ways of weighing the importance of a model's accommodating each one. A philosophical account of integration ought to shed light on how decisions about the interpretation and accommodation of information that are the very substance of integration should be made.

We have shown that integrative research is effective when it includes a way of prioritizing the most reliable information, enabling it to place the most stringent constraints on acceptable solutions to a research problem. If this heuristic seems obvious, recall what can happen if it is not followed. It is plausible that, had Bragg, Kendrew, and Perutz explicitly considered the reliability of available information, they might not have made the blunder they did. Further, and more importantly, it is not always clear what it means in practice to prioritize the most reliable information. The heuristics for effective integrative modeling we've proposed offer a concrete way of doing so: Use the most reliable information for representation and searching; less reliable information for scoring; and the least reliable information for analysis.

Extending more general philosophical considerations about the epistemic value of iteration (Chang, 2004), our case study highlights its functions in effective integrative research. Iteration can enable the redistribution of information among the steps of modeling by increasing researchers' confidence in some pieces of information. Even information that initially appears to be unreliable can become useful when it is found to be consistent with models produced without it. That is, integrating *some* information in a particular way enables reassessment of the reliability of *other* information, warranting its use as a more stringent constraint in subsequent iterations. Assessing consistency with other information, rather than agreement with the target system, enables this reassessment of information's reliability. Effective integrative research, then, enables researchers to extract as much *new* information as possible from the information they already have, thereby maximizing its value.

Iteration also plays a role in defining the model representation. When they define the model representation, modelers ought to consider not only which model representation is best suited to their goals, but also which way of getting it wrong—making it too accommodating or not accommodating enough—would be more productive. The more general takeaway is that, in any iterative process, researchers should consider not only the outcome of a particular iteration, but also different ways of erring. Doing so enables them to err in the direction that best promotes an efficient iterative process.

Finally, the iterativity of integrative modeling enables the reassessment of how information was integrated within a given iteration. Finding no acceptable models at the end of an iteration indicates that an error has been made. Subsequent iterations constitute a systematic attempt to identify its source. This function of iteration is crucial. We began with the observation that, in complex modeling problems, there is often a lot of information and many possible ways in which it could be integrated, but time and other limited resources preclude trying all of them. Several permutations of available information can nonetheless be assessed. Although trial and error is involved in this assessment, it need not take place in an ad hoc, haphazard fashion. Our heuristics for designing iterations of integrative modeling introduce some systematicity to the process.

4.2 Understanding modeling

With these lessons about integration in hand, let us now turn to what we can learn about modeling. We will show that our analysis also sheds light on the nature of model construction and evaluation, as well as on the significance of abstraction in these processes. We do so by contrast to the hierarchical view of models, which was introduced by Patrick Suppes (1962) and has since been developed (Giere, 2010; Mayo, 1996). According to this view, *models of theory* are constructed from theoretical principles together with various auxiliary assumptions that enable their application to particular systems. For example, a model of the earth-moon system can be constructed from Newton's laws of motion together with additional constraints (Giere, 2010). Since we often do not have direct access to the systems our models represent, we cannot compare them directly to these systems. Instead, we can collect data via experiment or observation—for instance, noting the position of the moon at various points in time. We can then infer a relationship among the data called a *model of data*. We use models of data to assess models of theory. The key point for our purposes is that theory is used for the *construction* of (theoretical) models, data for their *assessment*, with models of data acting as intermediaries in this process.[12]

In contrast, we framed the challenge addressed by integrative modeling in terms of finding a model that best accommodates all available information—from data, theory, or any other source. Accordingly, we saw that information from theory and data alike may be used in model construction, that is, for defining the model representation, constructing a scoring function, and searching for acceptable models. We end up with neither models of theory nor models of data, but rather with *models of information* (Bolinska, forthcoming). Similarly, information from any source may be used for model evaluation, that is, reserved for the analysis step. Rather than using theory to construct models and data to assess them, the construction and assessment of models alike may rely upon information from either theory or data. Theory and data can serve the same epistemic ends; there is no sharp delineation between them.

What emerges is a coherentist view of knowledge. No single piece or kind of information acts as a foundation upon which models are constructed. Instead, the reliability of all information must be assessed and reassessed using the iterative integrative modeling workflow. The reliability of information must be assessed, first, to enable the appropriate distribution of information among the steps of modeling that we argued for in Sect. 3.2. If it proves impossible to construct a model on the basis of a given distribution of information—which reflects a given assessment of reliability—then reliability must be reassessed and information redistributed accordingly in a subsequent iteration of modeling. In the analysis step, models are assessed according to how well they accommodate *all* available information; when no model can accommodate all information sufficiently well, those models that accommodate more reliable information are favored. Rather than taking some pieces of information to be known or given, the reliability of every piece of information—and the models constructed on the basis of particular reliability assessments—are evaluated in an

[12] Although the hierarchical view has drawn criticism, this basic role differentiation for theory and data is widely accepted (Bokulich & Parker, 2021; Karaca, 2018; Leonelli, 2019).

iterative fashion. Revising assessments of reliability, models, or both is encouraged if a coherent picture does not emerge.[13]

This view takes seriously the fact that we cannot access biomolecular complexes independently of the experimental means at our disposal (Chang, 2022; Matthiessen, 2022). The assessment of models therefore takes place entirely within the realm of modeling, rather than by comparison to the world outside this realm. Again, we may invoke the analogy to data models. Consider the curve-fitting problem: How should we select between infinitely many curves that capture a given dataset to varying degrees? This problem arises because data are a product of both the target phenomenon and measurement error. We want the curve we select to accurately reflect the target phenomenon, not noise introduced by the measurement process (Forster and Sober 1994). That is, we want to accommodate the data sufficiently well without overfitting them. Because we cannot compare the curve directly to the phenomenon it represents, we must use a principled means to select a curve *from the data itself*. For instance, Forster and Sober (1994) invoke the Akaike Information Criterion as a way to balance the competing desiderata of goodness-of-fit and simplicity.

Similarly, in integrative modeling, models are assessed according to how well they accommodate different pieces of information, rather than by comparison to their target systems. But assessing these models is not just curve-fitting. As we pointed out, the challenge that integrative modeling addresses is that vast amounts of information of variable reliability must be accounted for, each targeting different parts or aspects of the target system. Addressing this challenge requires further tools, which the iterative process of integrative modeling gives us.

Our models-of-information framework enables us to set aside philosophical questions about abstraction (e.g., Cartwright, 1983; Godfrey-Smith, 2009; Jones, 2005), focusing instead on their practical ramifications. Scientists are often concerned precisely with these, that is, with how much detail to include in a model. For instance, Chris Eliasmith and Oliver Trujillo (2014) describe large-scale brain modeling as follows:

> "Is there a right 'level of detail'? We believe that this is simply an ill-posed question. [...] [T]he appropriate scale is determined by balancing two things: first, the questions that need to be answered and second, the computational resources" (2014, 3).

Our case study offers concrete normative guidance for how to strike this balance. Integrative modeling requires modelers to *decide*, when they determine the model representation, which variables to include in the model and at what level of detail. That is, they must adopt particular abstractions, assuming them for the sake of model construction. Our heuristics show how the level of detail should be determined initially and how it can be reassessed in later iterations of modeling: Begin with a less accommodating model representation, i.e., include less detail. If no acceptable model can be found, adopt a more accommodating representation in the next iteration.

[13] For a detailed account of models of information and the coherentist picture of scientific knowledge supported by this view, see Bolinska (forthcoming).

Philosophical questions about abstraction have to do with how well a model can represent its target. But as mentioned above, the relationship between model and target cannot be assessed using some independent means. Instead, all we can do is build certain abstractions into our models, assessing those models over the course of multiple iterations. In other words, the issues our paper addresses are in some sense prior to questions about abstraction and representation. We take a key contribution of our paper to be that, even if we set these issues aside, significant challenges must still be addressed, arising from the volume, variable reliability, and ambiguity of available information.

5 Conclusion: Beyond integrative modeling

We conclude by suggesting how our lessons from integrative modeling can be applied to domains outside structural biology. The challenges that integrative modeling was designed to address are also present in other domains. Especially in an age of big data, scientists often have a lot of information at their disposal. Each piece of information, whether it comes from theory or from data, may admit of many interpretations. Therefore, there can be many ways to integrate this information. Moreover, resource limitations incentivize efficiency. Researchers must be strategic about *how* they integrate information, prioritizing some ways of interpreting and integrating information over others.

Further, the steps of integrative modeling often have analogs—even in research that does not use modeling per se. In any domain, researchers must either begin by selecting the variables whose values will be determined (as in integrative modeling) or, more generally, by specifying a research problem and what qualifies as an acceptable solution. They must determine, for instance, how precise or detailed a solution they seek and what aspect of the system in question to target. Second, they must evaluate solutions to the research problem—be they models, as in our case study, or other epistemic products, like theories or explanations. Even if that evaluation is not explicit or quantitative, researchers must nonetheless determine how well putative models, theories, or explanations can accommodate or account for different pieces of information. Third, any research problem has a space of possible solutions—again, whether this space is explicitly acknowledged as such or not. Fourth, some form of analysis must take place once a plausible solution has been identified. Moreover, research in many domains involves iteration. Therefore, the heuristics for effective integrative modeling we propose can be extended to other domains. They may serve as a blueprint for how researchers can take *all* available information, put it into a model, and evaluate that model in a coherentist way.

There are limitations, though. Eric Hochstein (2023) has argued that, for large-scale models of the brain, integration need not consist in a single unified model or theory. Rather, shared metaphysical commitments between multiple models make extrapolation between them possible; integration consists not in a unified model, but

in the possibility of extrapolation that such shared commitments afford.[14] Whether a form of modeling akin to integrative modeling is possible in other domains will depend, then, on what their goals are. Given that integrative modeling aims to determine the structures of biomolecular complexes—the positions and orientations of their component parts—it is easy enough to conceive of information as constraints on structural models. In fields in which explanatory goals are less well defined, more multiplicitous, or more open-ended, this kind of integration may not be feasible or desirable. Nevertheless, given the analogs to integrative modeling in other domains adumbrated above, we contend that many problems are amenable to our account. Further specifying the prospects and limitations of its extension to disparate problems is a task for future work.

Acknowledgements We are grateful to Helen Berman, Carl Kesselman, Kate White, Brinda Vallat, and Jitin Singla for discussions about integrative modeling in the context of the Pancreatic Beta Cell Consortium. We presented this paper at the poster session of the 2022 Philosophy of Science Association Biennial meeting and in the Center for Philosophy of Science's Featured Former Fellow series at the University of Pittsburgh; we thank our audiences for helpful questions and comments. For insightful feedback on drafts of this paper, we thank Riana Betzler, Adrian Erasmus, and members of Hasok Chang's research group at Cambridge University. We are also grateful to two anonymous reviewers for comments that helped us improve the paper. Revision of this manuscript took place at the 2024 AJI Writer's Retreat, funded by the University of South Carolina's Ann Johnson Institute for Science, Technology, and Society and Department of Philosophy. We further acknowledge funding from NIH/NIGMS grants R01GM083960 and P41GM109824 (AS).

Funding Open access funding provided by the Carolinas Consortium.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., et al. (2007). Determining the architectures of macromolecular assemblies. *Nature,450*(7170), 683–694.

Algret, R., Fernandez-Martinez, J., Shi, Yi., Kim, S. J., Pellarin, R., Cimermancic, P., Cochet, E., et al. (2014). Molecular architecture and function of the SEA complex, a modulator of the TORC1 pathway. *Molecular & Cellular Proteomics: MCP,13*(11), 2855–2870.

[14] See also Bolinska (2024) for a discussion of what a unified model of protein structure would look like and extrapolation between models constructed in different contexts.

Astbury, W. T., & Street, A. (1932). The X-ray studies of the structure of hair, wool, and related fibres. I. General. *Philosophical Transactions of the Royal Society of London, Series A,230*, 75–101.
Baetu, T. M. (2011). Mechanism schemas and the relationship between biological theories. In P. McKay, J. Williamson, & F. Russo (Eds.), *Causality in the sciences* (pp. 407–424). Oxford University Press.
Bechtel, W. (2013). From Molecules to Behavior and the Clinic: Integration in Chronobiology. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(4), 493–502.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research,28*(1), 235–242.
Bokulich, A., & Parker, W. (2021). Data models, representation and adequacy-for-purpose. *European Journal for Philosophy of Science,11*(1), 31.
Bolinska, A. (2018). Synthetic versus analytic approaches to protein and DNA structure determination. *Biology & Philosophy,33*(3), 26.
Bolinska, A. (2024). A monist proposal: Against integrative pluralism about protein structure. *Erkenntnis,89*, 1711–1733.
Bolinska, A. (Forthcoming). Models of information in structural biology. In French, S., & Hermida, M. (Eds.) *The Philosophy of Biophysics*. MIT Press. Preprint version, PhilSciArchive. https://philsci-archive.pitt.edu/24213/
Bragg, W. L., Kendrew, J. C., & Perutz, M. F. (1950). Polypeptide chain configuration in crystalline proteins. *Proceedings of the Royal Society,203A*, 321–357.
Brigandt, I. (2010). Beyond reduction and pluralism: Toward an epistemology of explanatory integration in biology. *Erkenntnis,73*, 295–311.
Brigandt, I. (2013a). Systems biology and the integration of mechanistic explanation and mathematical explanation. *Studies in History and Philosophy of Biological and Biomedical Sciences,44*(4 Pt A), 477–492.
Brigandt, I. (2013b). Integration in Biology: Philosophical Perspectives on the Dynamics of Interdisciplinarity. *Studies in History and Philosophy of Biological and Biomedical Sciences, 44*(4), 461–465.
Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature,588*(7837), 203–204.
Carnap, R. (1947). On the application of inductive logic. *Philosophy and Phenomenological Research,8*(1), 133–148.
Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press.
Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.
Chang, H. (2022). *Realism for realistic people*. Cambridge University Press.
Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science,44*, 43–64.
Dickson, M. (2006). Plurality and complementarity in quantum mechanics. In H. K. Kellert, H. E. Longino, & C. K. Waters (Eds.), *Scientific pluralism.* University of Minnesota Press.
Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science,67*(4), 559–579.
Eliasmith, C., & Trujillo, O. (2014). The use and abuse of large-scale brain models. *Current Opinion in Neurobiology,25*, 1–6.
Fehr, C. (2006). Explanations of the evolution of sex: A plurality of local mechanisms. In H. K. Kellert, H. E. Longino, & C. K. Waters (Eds.), *Scientific pluralism.* University of Minnesota Press, Minnesota.
Fernandez-Martinez, J., Phillips, J., Sekedat, M. D., Diaz-Avalos, R., Velazquez-Muriel, J., Franke, J. D., Williams, R., et al. (2012). Structure–function mapping of a heptameric module in the nuclear pore complex. *The Journal of Cell Biology,196*(4), 419–434.
Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science,45*, 1–36.
Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese,172*, 269–281.
Godfrey-Smith, P. (2009). Abstractions, idealizations, and evolutionary biology. In Barberousse, A., Morange, M. & Pradeu, T. (eds.), *Mapping the future of biology: evolving concepts and theories* (Boston Studies in the Philosophy of Science 266) (pp. 47–56). Dordrecht: Springer Netherlands.
Griesemer, J. (2013). Integration of approaches in David wake's model-taxon research platform for evolutionary morphology. *Studies in History and Philosophy of Biological and Biomedical Sciences,44*, 525–536.
Hacking, I. (1981). Do we see through a microscope? *Pacific Philosophical Quarterly,62*(4), 305–322.
Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge University Press.
Hochstein, E. (2023). Integration without integrated models or theories. *Synthese,202*, 76.

Jones, M.R. (2005). Idealization and abstraction: A framework." In Jones, M R., & Cartwright, N. (eds.), *Idealization XII: Correcting the model* (Poznań Studies in the Philosophy of the Sciences and the Humanities 86) (pp. 173–217). Amsterdam and New York: Rodopi.

Judson, Horace Freeland. (1996). *The Eighth Day of Creation: Makers of the Revolution in Biology*. Expanded edition. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Kaake, R. M., Echeverria, I., Kim, S. J., Von Dollen, J., Chesarino, N. M., Feng, Y., Yu, C., et al. (2021). Characterization of an A3G-VifHIV-1-CRL5-CBFβ structure using a cross-linking mass spectrometry pipeline for integrative modeling of host-pathogen complexes. *Molecular & Cellular Proteomics.,20*, 100132.

Karaca, K. (2018). Lessons from the large hadron collider for model-based experimentation: The concept of a model of data acquisition and the scope of the hierarchy of models. *Synthese,195*(12), 5431–5452.

Kim, S. J., Fernandez-Martinez, J., Nudelman, I., Shi, Yi., Zhang, W., Raveh, B., Herricks, T., et al. (2018). Integrative structure and functional anatomy of a nuclear pore complex. *Nature,555*(7697), 475–482.

Kitcher, P. (1993). *The advancement of science: Science without legend, objectivity without illusions*. Oxford University Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. University of California Press.

Leonelli, S. (2013). Integrating data to acquire new knowledge: Three modes of integration in plant science. *Studies in History and Philosophy of Biological and Biomedical Sciences,44*(4 Pt A), 503–514.

Leonelli, S. (2016). *Data-centric biology: A philosophical study*. University of Chicago Press.

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science,9*(2), 22.

MacLeod, M., & Nersessian, N. J. (2013). Coupling simulation and experiment: The bimodal strategy in integrative systems biology. *Studies in History and Philosophy of Biological and Biomedical Sciences,44*(4 Pt A), 572–584.

Matthiessen, D. (2022). Empirical techniques and the accuracy of scientific representations. *Studies in History and Philosophy of Science,94*(August), 143–157.

Mayo, D. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mitchell, S. D. (2002). Integrative pluralism. *Biology and Philosophy,17*(1), 55–70.

Mitchell, S. D. (2003). *Biological complexity and integrative pluralism*. Cambridge University Press.

Mitchell, S. D. (2009). *Unsimple truths: Science, complexity, and policy*. University of Chicago Press.

Mitchell, S. D. (2019). Perspectives, representation, and integration. In M. Massimi & C. D. McCoy (Eds.), *Understanding perspectivism: Scientific challenges and methodological prospects* (pp. 178–193). Routledge.

Mitchell, S. D. (2020). Through the fractured looking glass. *Philosophy of Science,87*, 771–792.

Mitchell, S. D., & Gronenborn, A. M. (2017). after fifty years, why are protein X-ray crystallographers still in business? *The British Journal for the Philosophy of Science,68*(3), 703–723.

Morrison, M. (2011). One phenomenon, many models: Inconsistency and complementarity. *Studies in History and Philosophy of Science Part A,42*(2), 342–351.

Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. Oxford University Press.

Nathan, M. J. (2017). Explanatory unification. *The British Journal for the Philosophy of Science,68*, 163–186.

Olby, Robert C. (1974). *The Path to the Double Helix: The Discovery of DNA. Seattle.* University of Washington Press.

O'Malley, M. A., & Soyer, O. S. (2012). The roles of integration in molecular systems biology. *Studies in History and Philosophy of Science Part c: Studies in History and Philosophy of Biological and Biomedical Sciences,43*(1), 58–68.

Pauling, L., Corey, R. B., & Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America,37*(4), 205–211.

Plutynski, A. (2013). Cancer and the goals of integration. *Studies in History and Philosophy of Biological and Biomedical Sciences,44*(4 Pt A), 466–476.

Plutynski, A. (2018). *Explaining cancer: Finding order in disorder*. Oxford University Press.

Rieping, W., Habeck, M., & Nilges, M. (2005). Inferential structure determination. *Science,309*(5732), 303–306.

Rout, M. P., & Sali, A. (2019). Principles for integrative structural biology studies. *Cell,177*(6), 1384–1403.
Sali, A. (2021). From integrative structural biology to cell biology. *Journal of Biological Chemistry*. https://doi.org/10.1016/j.jbc.2021.100743
Schneidman-Duhovny, D., Pellarin, R., & Sali, A. (2014). Uncertainty in integrative structural modeling. *Current Opinion in Structural Biology,28*(October), 96–104.
Suppes, P. (1962). Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science.* Stanford University Press.
Viswanath, S., Bonomi, M., Kim, S. J., Klenchin, V. A., Taylor, K. C., Yabut, K. C., Umbreit, N. T., et al. (2017). The molecular architecture of the yeast spindle pole body core determined by Bayesian integrative modeling. *Molecular Biology of the Cell,28*(23), 3298–3314.
Weisberg, M. (2007). Three Kinds of Idealization. *Journal of Philosophy,104*(12), 639–659.
Weisberg, M. (2013). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Authors and Affiliations

Agnes Bolinska[1] · Andrej Sali[2,3]

✉ Agnes Bolinska
bolinska@mailbox.sc.edu

Andrej Sali
sali@salilab.org

[1] Department of Philosophy, University of South Carolina, Columbia, SC 29208, USA

[2] Department of Bioengineering and Therapeutic Sciences, Quantitative Biosciences Institute (QBI), San Francisco, CA, USA

[3] Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA 94157, USA