Models of Information in Structural Biology

Agnes Bolinska

Abstract: On the hierarchical picture of models, theoretical models are constructed on the basis of theory and assessed by comparison to distinct models constructed from empirical data. Using the determination of the structure of the folded polypeptide chain as a case study, I instead argue that information from theory and data alike can be interpreted as constraints in the construction of *models of information*. On this view, more reliable information ought to be prioritized, sometimes forcing reinterpretations of less reliable information; information from theory and data is thus interdependent. I show how the reliability of information can be assessed, arguing that the evidence for a planar peptide bond was stronger and more secure than the evidence for a repeating subunit every 5.1 Å, and that disciplinary origin in physics or biology is immaterial to assessing reliability. I further show how models are assessed alongside interpretations of information in a coherentist manner: a better model accommodates more information, particularly reliable information; a model's inability to accommodate some information necessitates reinterpreting that information.

1. Introduction

Models often cannot be compared directly to their target systems, which are too small, too large, too far away, or otherwise inaccessible. How, then, is assessing them possible? A common answer comes from the hierarchical picture of models (Suppes 1962; Mayo 1996; Giere 2010). On this view, a *model of theory* is derived from general theoretical principles together with auxiliary assumptions enabling their application to a particular target system. For example, by specifying additional conditions and constraints, we can use Newton's laws of motion to derive a model of the earth-moon system (Giere 2010). We access the target system by gathering data about it via experiment or observation. We might, for instance, observe the moon's position at different points in time. Then, we identify a pattern among the data, positing a relationship between them called a *model of data*—for instance, by fitting a curve through the data points. That relationship can be compared to

the relationship between variables posited by the theoretical model. Although theoretical models cannot be compared to their target systems directly, they can be compared to them indirectly, via data models.[1] According to the hierarchical picture, theory and data play distinct epistemic roles and occupy separate spheres. Theory is used for the *construction* of theoretical models, which embody scientific *knowledge*. Data is used for *assessing* those models, with data models playing an intermediary role in such assessment; data thus serve as *evidence*.

This paper has two aims. First, to challenge the hierarchical picture of models by examining how scientists approached the problem of protein structure in the mid-twentieth century. I show that they did not construct one model from theory and a separate model from data, using the data model to test the theoretical one. Instead, scientists were constructing what I call *models of information*, where "information" is an umbrella term encompassing both theory and data. Theory and data played similar epistemic functions in this case: Both had to be interpreted as constraints in the construction of structural models. Constructing models of information, I argue, requires assessing the reliability of these interpretations by scrutinizing the evidence for them. I illustrate how this takes place in practice. I further argue that the disciplinary origin of a piece of information— whether it comes from physics or biology—is not pertinent to its reliability.

Thinking about modeling this way raises a challenge. If theory and data are both used to *construct* models, how, then, can those models be *evaluated*? This paper's second aim is to show that model evaluation is coherentist. In particular, scientists must simultaneously assess two things: first, how well available evidence supports particular interpretations of the information at hand, and second, how well a model can accommodate, account for, or make sense of *all* available information. Importantly, how we interpret one piece of information can affect how we interpret another. When interpretations conflict, I argue, those that are best supported by the evidence ought to be prioritized, forcing reinterpretations of less reliable information. Evaluating our interpretations of information and our models thus take place in tandem.

The paper proceeds as follows. Section 2 introduces the case study, showing that information from theory and data alike can be interpreted as constraints on structural

---

[1] The hierarchy of models includes other elements, such as models of experiment and background conditions (Suppes 1962; Mayo 1996; Giere 2010), omitted here.

models. Section 3 then shows how scientists evaluate the reliability of such interpretations by examining the evidence for them. Whether a piece of information originates in physics or biology, I argue, does not affect the reliability of its interpretation; rather, scientists must examine what Kent Staley (2004) calls the *strength* and *security* of evidence for each interpretation. Section 4 illustrates how models are evaluated. In addition to determining how well a model can account for all information, model evaluation must take assessments of interpretations of information into consideration; the picture of knowledge that emerges is coherentist through and through. Section 5 concludes, illustrating the implications of the view developed in the context of this case to other areas of science.

2. The Epistemic Function of Information

In this section, I argue that information from theory and information from data play similar epistemic functions: upon interpretation, they *constrain* structural models upon interpretation. However, a model compatible with one piece of information may be incompatible with another. Further, interpretations are interdependent: Interpreting a piece of information in a particular way shapes how other pieces of information are interpreted. I illustrate these claims using a case from the history of structural biology. In the 1930s and 40s, scientists knew that proteins were made up of polypeptide chains, which consisted of amino acid residues connected to one another in a particular way (Figure 1), and that they play pivotal roles in diverse bodily functions. The central problem was to determine how these chains are folded, that is, their three-dimensional configurations, such that they could play these roles.
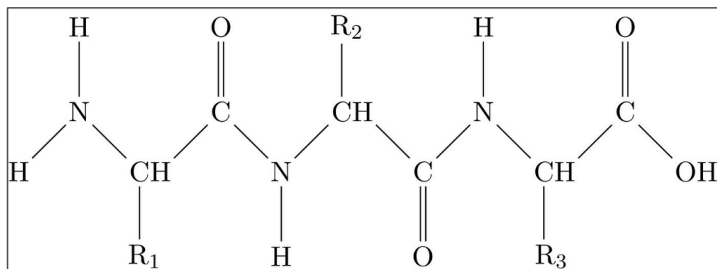

Figure 1. The general structure of the polypeptide chain. R groups represent side chains that differ for different amino acid residues.

Several pieces of information offered clues for solving this problem. Upon interpretation, each piece of information constrained structural models for the folded polypeptide chain in the sense that models incompatible with it were deemed *prima facie* unacceptable. For example, one influential piece of experimental data was an X-ray diffraction photograph of the protein keratin—a constituent of wool, fingernails, and hair—taken by William Astbury (Figure 2).[2] Astbury interpreted this photograph as indicating that the folded polypeptide chain contained a subunit that repeated every 5.1 Å (Astbury and Street 1931) and his interpretation became widely accepted (Olby 1994). Interpreted this way, the photograph constrained structural models in the sense that any model that didn't have such a subunit was considered to be incorrect.
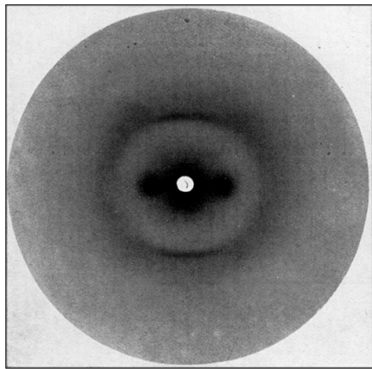

Figure 2. Astbury's X-ray diffraction photo of alpha keratin (Astbury and Street 1931).

Information from theory functioned in a similar way. For instance, chemical theory says that a molecule is allowed to rotate about single bonds, but not double bonds, which are planar. The peptide bond connecting amino acid residues into polypeptide chains is depicted in structural formulae as a single (C–N) bond. However, due to a phenomenon known as resonance, one of the electrons from the C=O bond spends some of its time at the peptide bond (Figure 3). As a result, the peptide bond has partial double-bond character, so it must be planar. The planarity of the peptide bond is a piece of information that, like Astbury's photograph, can constrain structural models of the folded polypeptide chain: Any model that allows rotation about the peptide bond does not qualify as acceptable.

---

[2] This photograph was part of an experimental study in which Astbury took X-ray diffraction photographs of samples of wool, horn, and hair both in unstretched ("alpha") and stretched ("beta") states. Photographs of these substances exhibited similar X-ray diffraction photographs (Astbury and Street 1931).
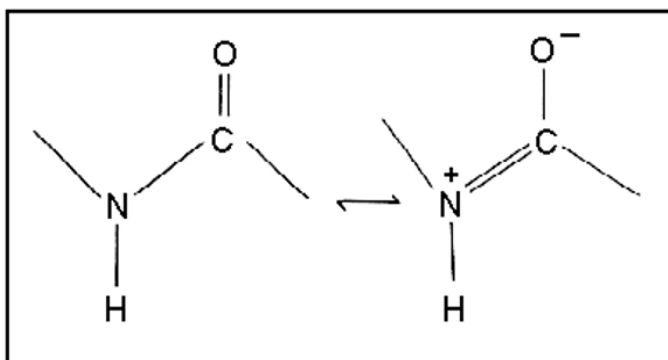
Figure 3. Resonance in the peptide bond.

A model that accommodates one piece of information (interpreted in a particular way) may be incompatible with another. For instance, in an early attempt to solve the problem of polypeptide chain folding, Sir Lawrence Bragg, John Kendrew, and Max Perutz's (1950) strategy was to list all structural models compatible with Astbury's interpretation of his photograph, selecting from among them the top candidates on the basis of other considerations. Although the structure they proposed had a repeating subunit every 5.1 Å, it rotated about the peptide bond. That is, the structure was compatible with information from Astbury's photograph but incompatible with information from chemical theory.

In contrast, Linus Pauling approached the problem by constructing physical molecular models on the basis of chemical principles, together with information about bond lengths and angles of individual amino acids. Consequently, his model—which he dubbed the alpha helix, and which turned out to be correct—had planar peptide bonds. However, rather than the 5.1 Å-repeat that Astbury's photograph indicated, it had a structural repeat every 5.4 Å (Pauling, Corey & Branson 1951). That is, Pauling's model accommodated information from theory but not from Astbury's photograph.

Notice that it made no functional difference whether information came from theory or data. Against the hierarchical picture of models, scientists did not construct one model solely from theoretical principles, comparing it to a distinct model constructed from data alone. Rather, they used any available piece of information—from theory or data—as a constraint telling them which models were acceptable. But in order to do so, they had to interpret that information with respect to the target system. That is, they had to understand it as telling them something about what the structure can and cannot look like—about

permissible and impermissible configurations for the folded polypeptide chain. We can conceive of the problem as beginning with a space of possible solutions, consisting of all possible folded configurations of the polypeptide chain, which can be successively reduced by taking into consideration of different pieces of information. The consideration of each piece of information, interpreted in a particular way, enables the elimination of portions of this possibility space.

However, several interpretations of a piece of information with respect to the target system may be possible. That is, a given piece of information may be interpreted as eliminating different portions of the possibility space. Further, the interpretation of any given piece of information can affect interpretations of other pieces of information. For instance, as a consequence of their commitment to accommodating information from Astbury's photograph, Bragg, Kendrew, and Perutz failed to notice that their model was incompatible with a planar peptide bond. Bragg lamented this oversight upon recognizing that Pauling's alpha helix was the correct structure for the folded polypeptide chain.[3] On the other hand, Pauling was aware that his model was unable to accommodate the information from Astbury's photograph. Yet he remained convinced that it was right, and that some other interpretation of the photograph would eventually be found.

Indeed, this is precisely what happened. It turned out that the reflection in Astbury's photograph seemingly indicating a repeating subunit every 5.1 Å was instead caused by a higher-order structure dubbed the *coiled-coil*: Two alpha helices, it turnde out, could be coiled around one another like strands of a rope.[4] Any piece of information may admit of multiple interpretations, some of which may be as yet unconceived at a given moment in time. And sometimes interpretations can conflict, showing that something is awry. How should such conflicts be resolved?

---

[3] Bragg would later describe this paper as "the most ill-planned and abortive in which [he had] ever been involved" (1965). According to Alexander Todd, professor of organic chemistry at Cambridge, he was especially "horrified" to have missed the planar peptide bond (Judson 1996, pp. 89-90).

[4] This higher-order structure was independently discovered by Pauling and Francis Crick (Judson 1996; Olby 1994).

3. Assessing Information

Imagine that you are a scientist working on the problem of polypeptide chain folding in the mid-twentieth century. Should you prioritize Astbury's photograph, the way that Bragg, Kendrew, and Perutz did, or should you instead, like Pauling, lean more heavily on the planar peptide bond? In this section, I describe how information ought to be assessed. I end up siding with Pauling, arguing that he rightly prioritized the planar peptide bond because it was more reliable than Astbury's photo.

3.1. Strength and Security

Some pieces of information constrain models more *reliably* than others. What does it mean for information to constrain models reliably? Recall that information is used to reduce the size of the space of possible models. By applying a piece of information as a constraint, we eliminate some portion of this possibility space. I propose that we understand the reliability of information in terms of our confidence that we have done so correctly. That is, the more confident we can be that the correct model remains in the possibility space (rather than having been eliminated) after applying information as a constraint, the more reliable the information. More precisely, we are concerned not with the reliability of the information itself, but of a particular interpretation of the information with respect to the target phenomenon. We are concerned, that is, about reasoning correctly from the information to the target phenomenon.[5] "Reliability of information" is shorthand for this more precise formulation.

As we have seen, our reasoning can be mistaken, since each piece of information admits of multiple interpretations. Astbury's photograph *seemed* to indicate that the folded polypeptide chain had a subunit repeating every 5.1 Å. However, it turned out that it could *also* be interpreted as indicating a coiled-coil higher-order structure of a folded polypeptide chain without such a subunit. Bragg, Kendrew, and Pertutz were mistaken in inferring from Astbury's photograph a model for the folded polypeptide chain with a 5.1-Å repeating

---

[5] In a similar vein, James Woodward (2000) argues that the reliability of data-to-phenomena reasoning can be assessed by examining processes of data production and data interpretation. The former can be improved, for instance, by constructing better instruments. Here, I focus on how we can assess the interpretation of information from data or theory.

subunit. In contrast, Pauling reasoned correctly from the planarity of the peptide bond to his model. These are descriptive claims.

But we can make a normative claim, too. In particular, I'd like to suggest that the peptide bond's planarity was more reliable than the 5.1-Å repeat inferred from Astbury's photograph. Moreover, its assessment as such was available—at least in principle—to historical actors.

Let us examine how scientists assess the reliability of interpretations of information. They can do so by asking how well supported these interpretations are by available evidence. So, first of all, they should assess the *strength* of the evidence: how compelling or persuasive it is. But they should also assess its *security*: how susceptible the evidence is to defeat from the failure of an auxiliary assumption on which it relies (Staley 2004).[6] The idea here is that evidence depends upon several auxiliary assumptions, each of which is itself supported by other evidence. Evidence is like a bridge supported by auxiliary assumptions serving as pilings. An engineer can make a bridge less susceptible to collapse either by reinforcing these pilings or by adding more pilings. Similarly, evidence can be made more secure either by strengthening the evidence for one (or more) of the auxiliary assumptions upon which it depends, or by adding further auxiliary assumptions (Staley 2004).

3.2 Assessing the Planar Peptide Bond

How strong and secure was the evidence for a planar peptide bond? In their paper announcing structural details of the alpha helix, Pauling and his collaborators Robert Corey and Herman Branson describe the planar peptide bond as follows:

 "This structural feature has been verified for each of the amides that we have studied. Moreover, the resonance theory is now *so well grounded* and its experimental substantiation *so extensive* that there can be *no doubt whatever* about its application to the amide group" (1951, p. 206; my emphasis).[7]

---

[6] Staley (2004) frames his discussion in terms of the strength of evidence and the security of evidence *claims*; however, he agrees that the notion of security can be applied to evidence itself (personal communication).
[7] Amides are compounds that contain an acyl group (R–C=O) connected to a nitrogen atom; the amide group is called a peptide bond when it is part of a polypeptide chain.

Pauling, Corey, and Branson took the planar peptide bond to be reliable: They were confident that they had used this piece of information to constrain the model correctly. This confidence was grounded in the fact that the evidence for a planar peptide bond relies on several auxiliary assumptions—that resonance theory in general is true and that resonance exists in several different amides—each of which is well supported by the evidence. That is, Pauling, Corey, and Branson were arguing that the evidence for a planar peptide bond is strong and secure.

We may scrutinize their argument by examining these auxiliary assumptions and why they were taken to support a planar peptide bond. The notion of resonance was first introduced in the helium atom, by Werner Heisenberg in 1926, and extended to the covalent bond between hydrogen and oxygen in water by Walter Heitler and Fritz London in 1927. Pauling pioneered the application of physical principles to chemistry, publishing a paper in which he demonstrated resonance in the carbon tetrahedron (Pauling 1928). Applying quantum mechanical considerations to molecular structures eliminated a sharp distinction between single and double bonds; Pauling showed that there was instead a continuum between purely single and purely double bonds, with interatomic distance giving a measure of percentage single/double bond composition (Pauling, Brockway, and Beach 1935). Other considerations from physical theory enabled predictions of molecular structure. According to the second law of thermodynamics, a molecule ought to adopt a configuration with the lowest free energy. Thus, if the free energy of the hybrid between several forms of a molecule was lower, the molecule would adopt that conformation—a phenomenon known as *resonance stabilization* (Pauling and Sherman 1933).[8]

These theoretical considerations aligned with mounting empirical evidence revealing intermediate bond distances between single and double bonds. Pauling showed that there was resonance in the carbonyl and amide groups in urea, oxamide, and oxamic acid—molecules with similar functional groups to those in the polypeptide chain (Pauling and Sherman 1933). In the following decade, more evidence for a planar peptide bond emerged: The structures of several peptides and amino acids, including diketopiperazine (Corey 1938), glycine (Albrecht and Corey 1939), DL-alanine (Levy and Corey 1941), glycylglycine (Hughes and Moore 1942; 1949), and L_S-threonine (Shoemaker et al. 1950), were

---

[8] For a detailed description of the history of the resonance concept and its application to the peptide bond, see Olby (1994).

published. Moreover, much of this work was replicated and refined in subsequent studies.[9] We can thus see why Pauling, Corey, and Branson were so confident in the planar peptide bond.

Nevertheless, these theoretical considerations and empirical evidence do not by themselves establish a planar peptide bond in the polypeptide chain; rather, a series of inferences is required to do so. Let us therefore look more closely at how a planar peptide bond was established in one of these molecules, diketopiperazine. For, in fact, the reasoning that established a planar peptide bond in diketopiperazine mirrored the reasoning for its planarity in the polypeptide chain: It required assessing the evidence for resonance, which itself comes from studies of other molecules, together with theoretical considerations.

Corey determined the structure of diketopiperazine using X-ray crystallography. From the crystallographic data, he found that it has a planar hexagonal ring structure—that is, that the ring is flat rather than puckered—and determined its bond distances and angles. He argued that diketopiperazine resonates among four structures (Figure 4), citing two pieces of information as evidence for this claim. First, that the C–O and C–N bond distances have values that are "characteristic of resonance in this type" (1938, p. 1603), as observed in earlier studies of molecules such as urea (Wyckoff and Corey 1934). Corey reasoned that we could extrapolate from these other molecules to diketopiperazine. Second, that the planar configuration of the ring is to be "expected for resonance among structures I to IV, in consequence of the stereochemical properties of the C–N double bonds in the ring" (ibid.). That is, from the stereochemistry typical of double bonds in general—that rotation about them is prohibited—we can infer a planar ring in diketopiperazine in particular. In contrast, Corey observed, all known hexagonal ring structures containing only single bonds are puckered. Therefore, "the observed coplanarity provides strong evidence for resonance" (ibid.).

---

[9] For instance, a three-year study on L_S-threonine (Shoemaker et al. 1950) built on earlier work (McCoy, Meyer, and Rose 1935; Meyer and Rose 1936).
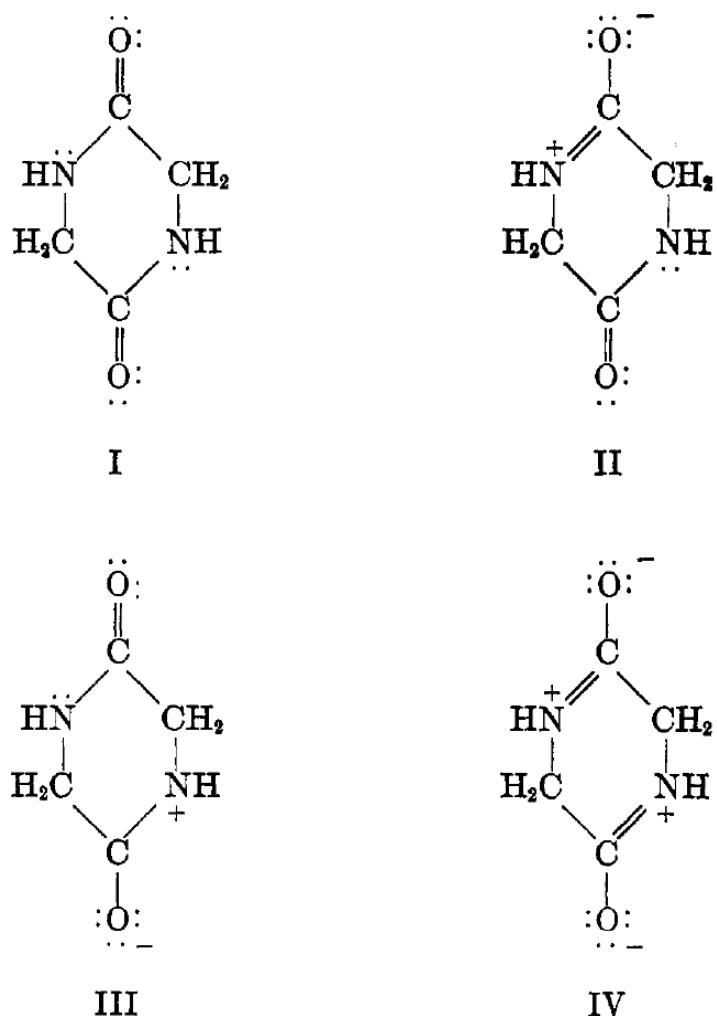
Figure 4. Resonance in diketopiperazine (Corey 1940, p. 231).

To conclude that the peptide bond in diketopiperazine was planar, then, Corey used empirical evidence from other small molecules, together with theoretical considerations. Extrapolating from the empirical evidence required Corey to assume that similar bond lengths indicated similar structures. He also assumed that stereochemical principles precluded rotation about double bonds, just as they did in other molecules. This assumption was further supported by the observation that hexagonal rings containing only single bonds are puckered, in contrast to the coplanarity of the atoms observed in the diketopiperazine ring. We can see, then, that the evidence for a planar peptide bond in

diketopiperazine was supported by several auxiliary assumptions, each of which were themselves well supported by other evidence. It was strong and secure.

In 1940, Corey first described the structure of the extended polypeptide chain, detailing its bond lengths and angles. Just as he did for diketopiperazine, Corey inferred the structure of the polypeptide chain from the structures of smaller molecules—in this case, diketopiperazine and glycine. He identified bond distances and angles in diketopiperazine and glycine corresponding to those in the polypeptide chain, noting their similarities to one another (Figure 5). He then inferred bond distances and angles for the polypeptide chain from averages of the diketopiperazine and glycine bond distances and angles. He further argued that "[t]here seems to be every reason for believing that the type of resonance evidenced by diketopiperazine, urea, etc. will be present in the polypeptide chain" (1940, p. 234).

TABLE 1

Corresponding interatomic distances found in molecules of diketopiperazine and glycine

	DIKETOPIPERAZINE	GLYCINE
	Å.	Å.
C—O	1.25	1.26
α-C—N	1.41	1.39
C—C	1.47	1.52
OC—N	1.33	

Figure 5. Corresponding interatomic distances in diketopiperazine and glycine (Corey 1940, p. 234).

As with diketopiperazine, evidence for a planar peptide bond in the polypeptide chain depended upon several auxiliary assumptions. In order to assess the reliability of the planar peptide bond in the polypeptide chain, we must therefore assess the evidence for each of these assumptions. A central assumption was that bond distances and angles in small molecules were similar to those in larger molecules. Indeed, small molecules generally do provide strong evidence for larger molecules with similar structures; this is a foundational assumption of chemistry, a cornerstone enabling chemical reasoning. Further, X-ray diffraction photographs of crystallized small molecules were generally clearer than

those of larger molecules, so the data based on smaller molecules was itself reliable.[10] Thus, the evidence for a planar peptide bond was strong and secure, being supported by many auxiliary assumptions that were themselves supported by strong evidence.

3.3 Assessing the 5.1-Å Structural Repeat

Let us now turn to Astbury's interpretation of his photograph as indicating a 5.1-Å structural repeat. I will show that the evidence supporting this interpretation was relatively insecure and weak, as compared to evidence for a planar peptide bond. Astbury had a reputation for drawing premature conclusions from his data; according to his colleague Ian MacArthur, he "brought his findings to market in the green ear, but would not clear the weeds nor suffer the system and technique necessary for the harvest" (1961, p. 332). Whereas the planarity of the peptide bond was established on the basis of studies of several small molecules over about a decade, the 5.1-Å repeat was based exclusively on Astbury's studies of alpha keratin. Historian Robert Olby writes,

> "[i]f he had been studying some other substance—polyoxymethylene, or a synthetic polypeptide, no doubt he would have arrived at a different solution. But he was, as [his assistant] H. J. Woods recalled, 'completely interested in keratin…' All the ideas were designed to try to explain keratin, not polyoxymethylene, rubber, or the then unknown synthetic polypeptides" (1994, p. 53).

The claim that the folded polypeptide chain would have a 5.1-Å repeat was based exclusively on the keratin data, without taking into consideration other information that could be relevant. Compared to the evidence for a planar peptide bond, then, it was insecure, resting heavily on few auxiliary assumptions.

What's more, the evidence for the 5.1-Å repeat was not as strong as the evidence for a planar peptide bond. By its very nature, information from X-ray crystallography admits of several interpretations. X-ray diffraction photographs are produced when a beam of X-rays is shone at a crystallized molecule. The electron clouds of the molecule diffract the beams; however, unlike in light photography, the beams are not subsequently refracted. Thus, the X-ray diffraction pattern gives only partial information about the structure that was causally

---

[10] As we will see shortly, X-ray diffraction photographs of larger molecules such as keratin tended to be blurry.

responsible for its production. Indeed, this is precisely why Astbury's photograph could be compatible both with a 5.1-Å structural repeat and a coiled-coil higher-order structure lacking this repeat.

This limitation of X-ray crystallography—that it provides partial information about structure, and that multiple structures are therefore compatible with a given image—was known at the time of the historical actors. Corey describes X-ray diffraction photographs of fibrous proteins like keratin as being "of great assistance in constructing a picture of the *probable arrangement* of the molecules in these substances" (1940, p. 230; my emphasis). However, he notes that "they *fail entirely to give any direct information* regarding the distances between discrete atoms in the polypeptide chain or, indeed, in any part of the protein molecule" (1940, p. 231). These photographs, in other words, give us clues about molecular structure, but cannot supply all the information needed for its determination.

Moreover, in the period of our case study, X-ray crystallography was still in its infancy. It was introduced in 1912 by Max von Laue, Walter Friedrich, and Paul Knipping, and had only just begun to be applied to large biomolecules in 1923 (Olby 1994). The complexity of these biomolecules, together with technical limitations, severely restricted the quality of photographs that could be produced, even by the most skilled crystallographers like Astbury. We can see this especially by contrast to contemporary X-ray diffraction photographs (Figure 6). Just as many different images can be inferred from a blurry light photograph, so too are many structural models compatible with a blurry X-ray diffraction photograph. All else being equal, the noisier our data, the less confident we can be in any given interpretation of that data: more interpretations lie within the data's error bars. Indeed, Astbury himself writes that "[i]t would not be justifiable at this stage to insist too strongly on the validity of such chemical interpretation as the present X-ray data suggest" (Astbury and Street 1931, p. 89).[11] Thus, even before the discovery of the coiled coil, there was reason to be wary of according any particular interpretation of an X-ray diffraction photograph too much evidentiary weight. Compared to the evidence for a planar peptide bond, the evidence for Astbury's interpretation was weak.

---

[11] He further notes that "it may well be that the indications of X-ray analysis do actually point the way to a solution, if only we may interpret them correctly" (Astbury and Street 1931, pp. 89-90). Astbury thus recognized both how challenging it could be to interpret X-ray diffraction photographs correctly and their promise to help with difficult problems.
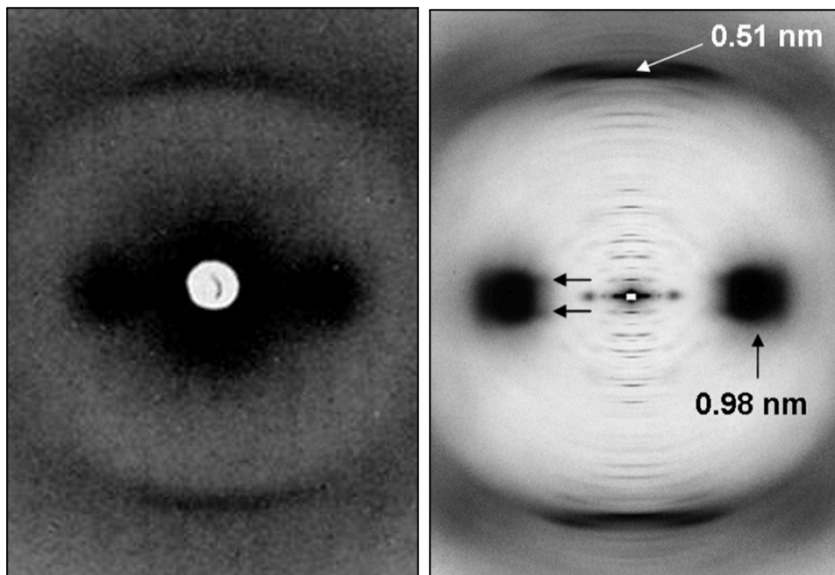
Figure 6. Astbury's photo of alpha keratin (left) (Astbury and Street 1931) compared to a contemporary X-ray diffraction photograph (right) (Parry 2021).

3.4 On Disciplinary Boundaries and the Role of Physics

I have argued that interpretations of information ought to be evaluated with respect to how reliable they are—how confident we can be that we have applied them correctly as constraints on structure, eliminating only incorrect models from the space of structural candidates. I then showed how assessments of these interpretations can be made: by considering the strength and security of the evidence for them. From these considerations, I argued that the planar peptide bond was a more reliable constraint on the structure of the folded polypeptide chain than the 5.1-Å repeat interpretation of Astbury's photograph.

However, there is an important influence on Pauling's work—and especially on his argument for a planar peptide bond—that we have yet to consider: his deep familiarity with physics. Pauling completed his graduate training at Caltech, an institution in which physical chemistry was closely informed by contemporary developments in quantum mechanics. Richard C. Tolman, an early member of Caltech's administration, believed quantum mechanics could be used to solve chemical problems, organizing seminars on this topic. Pauling attended these seminars, in addition to lectures by Arnold Sommerfeld in the early 1920s. Further, he spent 1926 in Europe as a John Simon Guggenheim Fellow, attending lectures by Sommerfeld, Erwin Schrödinger, and Peter Debye (Olby 1994). His aim was to

apply the new quantum mechanics that was emerging at the time to understanding molecular structure and the nature of the chemical bond (Pauling 1970).

What role did crossing disciplinary boundaries and Pauling's familiarity with physics play in his success? Does the fact that the argument for a planar peptide bond ultimately originated in physics make a difference to the assessment of its reliability as a constraint? According to Olby, what set Pauling's approach apart from Astubry's and Bragg, Kendrew, and Perutz's was his "*confidence* that quantum mechanics could be used as a check and a guide in the selection of probable molecular structures, and in particular to accept or reject the bond distances and angles derived from X-ray diffraction pictures" (1994, p. 268; original emphasis). In contrast, in Britain there were "flourishing schools of *protein* crystallography which had *not* been nurtured in an environment of the new physical chemistry" (ibid.). One might thus argue that Pauling's success resulted from his strong foundation in physics, rather than his prioritizing more reliable information.

But such an argument would be misleading. Although Pauling's familiarity with quantum mechanics indeed gave him confidence in his reasoning about the folded polypeptide chain structure, it wasn't this familiarity *per se* that gave him this confidence. Nor was the planar peptide bond's provenance in physics what made it a reliable piece of information. Rather, it was the strength and security of the evidence for the planar peptide bond that did so. As we saw above, some of that evidence happened to come from physics. But the reliability of information is not a product of disciplinary origin. Notably, Bragg, Kendrew, and Perutz each had extensive backgrounds in physics. Further, X-ray crystallography is itself a technique founded on physical principles. Bragg and his father Sir William Bragg were pioneers of X-ray crystallography, and Kendrew and Perutz relied heavily on it in their work. In the 1930s, Astbury worked with William Bragg at University College London. A guiding aim of his research program was to apply insights from physics (and chemistry) to problems in biology and medicine. He popularized the term "molecular biology," describing it as a field that applied experimental apparatuses and mathematical techniques from physics and chemistry to biological molecules (Hall 2014). Yet the scientists in Britain did not draw the same conclusions from the available information that Pauling did. It wasn't a background in physics that was responsible for Pauling's success.

What was it that set Pauling apart, then? Seymour Jonathan Singer, who had worked in Pauling's lab, describes his achievement as follows:

"What Linus did [...] was to insist that from his data on crystal structures of simple molecules, he could extrapolate. For example, that the peptide bond had to be planar. So he went ahead and imposed these restrictions. The length of the bond between the carbon and nitrogen atoms was much shorter, according to the X-ray data, than if the bond were single. So it had to have substantial double-bond character. So it had to be flat" (quoted in Judson 1996, p. 86).

What Pauling did and others failed to do was to correctly assess the significance of certain information, prioritizing it over other information. Pauling recognized that the folded polypeptide chain *had to have* certain features, such as a planar peptide bond, based on careful assessments of the strength and security of the evidence upon which these features were based. Once these features were imposed as constraints on structural models, other information was up in the air—including widely accepted information, such as the 5.1-Å structural repeat.

But this recognition was no mean feat. According to Singer, "what was astonishing about Pauling, what makes him great, was that he was willing to move to a concept on the basis of certain data whose relevance was not clear to others. Only Linus showed the willingness to take the inductive leap" (quoted in Judson 1996, p. 86). It wasn't his training in physics, but instead his ability to correctly assess the relative reliability of different pieces of information, applying them correctly as constraints on the structure, that set Pauling apart from others.

4. Assessing Models (and Information): A Coherentist View

I have argued that the planar peptide bond was a more reliable constraint than the 5.1-Å repeat. It was thus imperative to prioritize it in the construction of models: If it proved impossible to construct a model with both a planar peptide bond and a 5.1-Å repeat, models with a planar peptide bond ought to be preferred. In this section, we turn to the evaluation of these models—which, I argue, is interwoven with the assessment of information. I show that model evaluation involves both the assessment of different pieces of information and how well a model accommodates each piece of information.

In his seminal (1962) paper, Patrick Suppes introduced the hierarchical picture of models to address a puzzle about theory assessment. To assess the empirical adequacy of a theory, one needs to determine whether the theory stands in an appropriate relation to relevant experimental data. But assessing empirical adequacy directly is typically not possible. Some theoretical notions have no observable correlates in an experiment. Further, theories can contain continuous functions and infinite series, whereas data are necessarily discrete and finite. Assessing the empirical adequacy of a theory, in other words, requires one to commensurate otherwise incommensurable entities. Suppes argues that doing so necessitates a hierarchy of models connecting theory to data. He conceives of what he calls a *model of theory* as a set-theoretic structure, a possible realization of a theory. Analogously, he defines a *model of data* as a possible realization of the data—a model that fits the data well enough, as assessed using various statistical tests. By containing only those aspects of the experiment with corresponding parameters in the theory, data models can, in turn, be used in statistical tests of a theory's adequacy.

On the hierarchical picture of models, data models can be used to assess theoretical models because data and theory are independent of one another: A theoretical model is constructed exclusively from theory, a data model exclusively from data.[12] Yet, in our case study, model-building required scientists to *interpret* each piece of information as a constraint on the target system, and the interpretation of each piece of information—from theory or data—depended on how one interpreted *other* information. Rather than being kept apart in the process of model construction, interpretations of data were influenced by interpretations of theory and *vice versa*.

In fact, the situation is even more complex. I've illustrated the way in which two pieces of information can be interpretationally linked. But determining structural models typically requires the consideration of much more than two pieces of information. In Section 3.2, we examined the role that studies of smaller molecules played in reasoning about the planar peptide bond. These studies also provided information about bond lengths and angles in the polypeptide chain (Corey 1940). And there were other considerations, too—for instance, that the folded polypeptide chain should be held together stably by

---

[12] The underdetermination of data models by data has been noted by several authors. Multiple data models can be inferred from a given dataset (MacAllister 1997); interpretation is required to select from among them (Harris 2003). Data can be used as evidence for a variety of claims depending on how they are interpreted in different contexts (Leonelli 2016).

hydrogen bonds and that differences between amino acid residues would not affect polypeptide chain folding. Also in play were assumptions from X-ray crystallography, such as the folded polypeptide chain having an integral repeat (Judson 1996; Olby 1994).[13] The task scientists faced was how to apply *all* of these considerations—and what to do when this proved to be impossible.

What emerges, on this view, is a kind of coherentism about scientific knowledge, which applies both to the interpretation of individual pieces of information *and* to the assessment of models. Rather than building a theoretical model and assessing it against a data model, scientists were instead constructing a model that could make sense of all information at their disposal, interpreted in plausible ways. Pauling describes his earliest attempt to determine the structure of the folded polypeptide chain as follows:

> "Our X-ray work with crystals and electron-diffraction work with gas molecules had provided a large amount of information about bond lengths in simple molecules, and I decided to see to what extent this information could be made compatible with the X-ray diffraction diagrams of the fibrous proteins. [...] I thought that the general structure theory should permit predictions to be made not only of the values of bond lengths and bond angles, but also about the formation of hydrogen bonds and the planarity of the amide group." (Pauling 1970, p. 1003).

The question for Pauling wasn't whether he could find a model that fits the data or matches up to the world. Rather, he was interested in reconciling information from many sources by constructing a model in which each piece of information is compatible with every other. This information included different kinds of experiment, conducted on different molecules, and theory.

Further, the attempt to find a model that makes different pieces of information mutually compatible also facilitates assessing interpretations of this information. Just as a good model can accommodate all available information, good interpretations are those that can be accommodated by a reasonable model. Consider Pauling's reasoning about Astbury's photograph as indicating a 5.1-Å repeat:

---

[13] The integral repeat, like Astbury's photo, also proved to be misleading: It was a guiding assumption in Bragg, Kendrew, and Perutz's (1950) reasoning.

"I knew that what Astbury said wasn't right, because our studies of simple molecules had given us enough knowledge about bond lengths and bond angles and hydrogen-bond formation to show that what he said wasn't right." (quoted in Judson 1996, p. 82).

Pauling's conviction that there was something wrong with Astbury's interpretation stemmed from the fact that it didn't square with other information constraining polypeptide chain folding.

Nevertheless, Pauling "didn't know what *was* right" (quoted in Judson 1996, p. 82). The incompatibility of different pieces of information with one another doesn't itself reveal where the problem lies. How, then, should conflicts between incompatible interpretations of information be resolved? In Section 3, I argued that interpretations backed by strong and secure evidence are more reliable and therefore ought to be prioritized over interpretations supported by relatively weak and insecure evidence. However, assessing strength and security isn't always easy; if it were, Pauling's determination of the alpha helix wouldn't have the historical significance it rightly does.

Model construction can thus further aid the resolution of conflicting interpretations. Consider Pauling's 1937 attempt to find a model for the folded polypeptide chain. No matter how hard he tried, Pauling could not construct a model that simultaneously accommodated theoretical considerations, information about bond lengths and angles, *and* the 5.1-Å repeat. So, contrary to his initial assessment that Astbury's interpretation must be wrong because it didn't square with other information, Pauling took seriously the possibility that he was "making some unjustified assumption about the structural properties of the molecules" (quoted in Judson 1996, p. 82). Thus, he decided that he and Corey "should determine the structures of amino acids and simple peptides, to see if there wasn't something important that [they] were overlooking—some structural feature that didn't show up in the more distantly related simple molecules that [they] had been studying" (ibid.). The impossibility of finding a model that accounted for all available information led Pauling to a reassessment of the information he had considered most reliable. Something had to give. Either Astbury's interpretation of his photograph was wrong, or (at least) one of the assumptions that Pauling had made was. By 1948, after over a decade of research, Pauling concluded that

"there were no surprises about these molecules" (ibid.). He and Corey "had made [their] information more precise but hadn't changed [their] understanding in any qualitative sense" (ibid.).

Pauling's inability to construct a model accommodating all available information prompted more research, the aim of which was to reassess some of that information—in this case, information about bond lengths, bond angles, and the planar peptide bond. This research strengthened the evidence for that information. In doing so, it also reinforced Pauling's conviction in his model and in Astbury's interpretation being mistaken. The aim of modeling is thus both to find an adequate model—one that makes all information mutually compatible—and to further facilitate the assessment of information. The failure to find a model that accommodates all information can lead to the reassessment of evidence for an interpretation. Pauling's initial confidence in his assessment that Astbury's interpretation of his photograph was mistaken was strengthened after his attempt to construct a model using this interpretation, and the further empirical work he conducted with Corey. However, note that this work could have had a different outcome. Had Pauling and Corey found a mistake in their earlier work, this may be warranted prioritizing the 5.1-Å structural repeat over the planar peptide bond.

Once a given interpretation is confirmed to be reliable, the task then shifts, from constructing a model that accommodates *given* interpretations, to finding an *alternative* interpretation of the most weakly supported information. Consider again Astbury's interpretation of his photograph as indicating a 5.1-Å structural repeat. In Section 2, we saw how the coiled-coil higher-order structure of protein offered an alternative interpretation of this photograph. Other evidence for an alternative interpretation also emerged. In 1949, scientists from the commercial company Cortauld's synthesized a fibrous polypeptide with a somewhat different composition from keratin (Ambrose and Hanby, 1949; Bamford, Hanby and Happey, 1949a, 1949b). The X-ray diffraction pattern for this fiber lacked the reflection indicating a 5.1-Å repeating subunit. This additional information further undermined Astbury's interpretation of his photograph—and strengthened Pauling's (and the scientific community's) assessment that the alpha helix was correct (Olby 1994).

Thus, although we can assess the reliability of interpretations of information in (relative) isolation, by assessing the strength and security of evidence adduced in their favor, we can also do so holistically, by examining how well those interpretations can be

reconciled with interpretations of other information, and how well a model accommodates all information. Perhaps a given piece of information doesn't seem to have good support if we focus only on the strength and security of evidence adduced in its favor. This could be the case for a variety of reasons, including simply mistakes in reasoning, on an individual or community-wide level.[14] However, if it coheres with other information that *does* have good support, and if it is part of a model that makes all information mutually compatible, that gives us good reason to rely heavily on that information. The converse is also true. We might have a piece of information that appears to be well supported by the evidence, but if it doesn't cohere well with other information, then we should question its interpretation. For instance, although Astbury's interpretation of his photograph had gained widespread acceptance, Pauling showed that it was not in fact sufficiently well supported by the evidence.

In Section 2, I argued that there was no sharp distinction between information from theory and information from data. We can now see that the distinction between information and models as distinct epistemic entities, with one serving as evidence for the other, is also blurred. A model that best accommodates the most reliable information itself becomes a new, reliable piece of information, which can warrant the reassessment of other information. With each failed attempt at model construction, one must ask, what is the most weakly supported information, and how might it be reinterpreted? The evaluation of information and models is interdependent.

5. Conclusion

Although the hierarchical picture of models has drawn criticism, the basic distinction between models of data and models of theory has largely been retained.[15] In

---

[14] For instance, an assumption from crystallography was that the folded polypeptide chain would have an integral repeat. This was an assumption that Pauling was willing to abandon, but most adopted (Judson 1996; Olby 1994).

[15] Leonelli (2019), for instance, argues that the hierarchical view has limited scope: it does not apply to data that cannot be subjected to statistical manipulation, such as images, or to cases of exploratory research, where there isn't a theoretical model that is tested by the data. She also suggests that the hierarchical picture cannot account for the fact that the activities of data acquisition and data manipulation are intertwined. Karaca (2018) similarly argues that the hierarchical picture lacks a model of data acquisition. Bokulich and Parker (2021) argue that the hierarchical view forces us to identify data models with the theoretical models, leaving "the relation between data models and the world at best unanalyzed, and at worst erased" (p. 31).

contrast, I have argued that we should abandon this distinction for the case of modeling the folded polypeptide chain structure. In this case, scientists constructed *models of information*, using theory and data alike as constraints on possible structures. At the same time, they assessed and reassessed different interpretations of available information. These interpretations were interdependent. Extending Neurath's metaphor, we might imagine them building a raft not from uniform planks, but from pieces with different shapes. Exchanging one of these pieces constrains which other pieces we can use in its place: Replacing, for instance, a square piece with a triangular one half its size requires some other piece to fill in the gap that opens up.

On this view, interpretations of information are evaluated alongside models constructed on their basis. Such interpretations must be compatible with one another *and* with the models they are used to construct. The interpretation of information—from theory or data, from physics or biology—and the evaluation of models are not separate steps, but instead require a holistic, iterative strategy. Scientists must decide how to interpret different pieces of information, applying abstract theoretical claims and messy experimental data alike to the problem at hand. Interpreting each piece of information requires assessing a chain of evidential dependencies, that is, considering how each piece of information is supported by other information. Doing so in turn requires determining which inferences are and are not permitted on the basis of each piece of information. I offered some examples of how this is done, tracing the reasoning for a planar peptide bond in the polypeptide chain to a planar peptide bond in diketopiperazine and other small molecules. We could continue evaluating the evidence for each piece of information, tracing the reasoning even further back—in the same way we might answer a child's persistent series of "why?"s.

One part of this process is interpreting each piece of information with respect to the target system, transforming it into a constraint. Taken individually, what does it tell us about what a model of the target can and cannot look like? However, another is determining how best to apply *all* such constraints. Constraints typically cannot be applied simultaneously, but instead just one (or a few) at a time. This process may reveal a conflict between the interpretation of one piece of information and the interpretation of another, forcing reassessment of both. The challenge is to determine which interpretation to retain, eliminating the portion of the possibility space indicated by applying it as a constraint, and which to abandon, finding an alternative that makes sense of the retained interpretation.

I would like to conclude by suggesting that the case I have used to argue for the models-of-information view is not unique. In contemporary structural biology, it is typically not possible to construct models on the basis of theory alone (Mitchell & Gronenborn 2017). Especially in the determination of the structures of large biomolecular complexes, data from multiple experimental techniques is often required. Contemporary structural biologists have even more information at their disposal than the scientists in our case study did. In addition to X-ray crystallography, many other experimental techniques such as NMR spectroscopy, cryo-EM, and cross-linking with mass spectroscopy are now available. Theoretical techniques such as *ab initio* protein structure prediction methods and bioinformatic approaches are much more powerful (Sali et al. 2003). Further, structural biologists rely upon models deposited in the Protein Data Bank (PDB), a repository holding experimentally and computationally determined protein and nucleic acid structures (Berman et al. 2000). Such models can help them to solve related structures, or be used in the determination of the structure of a larger complex.

What's more, many features of contemporary structural biology are common to other areas of science. Theoretical principles must often be used in conjunction with experimental data to construct models. Data may not come from a single experiment, but instead from multiple sources. They can have varying levels of resolution and can target different parts or aspects of the target system. Further, data from experiments on related systems can also be relevant to our understanding of the target. We want, in other words, to construct *models of information*, telling us what all available empirical data and theory, possibly from different disciplines—taken together—tell us about the target. Understanding the construction and assessment of models of information in the case of polypeptide chain folding can thus also shed light on these other areas of science.

References

Albrecht, G., & Corey, R. B. 1939. "The crystal structure of glycine." *J. Am. Chem. Soc* 61(5): 1087–1103.

Ambrose, E. J. & Hanby, W. E. 1949. "Evidence of Chain Folding in a Synthetic Polypeptide and in Keratin." *Nature* 163: 483–84.

Astbury W. T., and Street, A. 1932. "The X-ray studies of the structure of hair, wool, and related fibres". *I. General. Philos Trans R Soc Lond Ser A* 230:75–101.

Bamford, C. H., Hanby, W. E. & Happey, F. 1949a. "Evidence for α-Protein Structure in Polypeptides." *Nature* 164: 138–39.

Bamford, C. H., Hanby, W. E. & Happey, F. 1949b. "The α-β Transformation in a Polypeptide." *Nature* 751–52.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42.

Bokulich, Alisa, and Wendy Parker. 2021. "Data Models, Representation and Adequacy-for-Purpose." *European Journal for Philosophy of Science* 11 (1): 31.

Bogen, James, and Woodward, James. 1988. "Saving the Phenomena." *The Philosophical Review* 97:303-352.

Bolinska, Agnes. 2018. "Synthetic Versus Analytic Approaches to Protein and DNA Structure Determination." *Biology and Philosophy* 33 (3-4): 26.

Bragg, Lawrence, J. C. Kendrew, and M. F. Perutz. 1950. "Polypeptide Chain Configurations in Crystalline Proteins." *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 203 (1074): 321–57.

Bragg, Lawrence. 1965. "First Stages in the X-ray Analysis of Proteins." *Reports on Progress in Physics* 28 (1): 1–14.

Carpenter, B. & Donohue, J. (1950). "The Structure of the Amino Acids, I. The Configuration of Crystalline DL-Serine." *J. Am. Chem. Soc.* 72(6): 2705-2713.

Corey, Robert B. 1938. "The Crystal Structure of Diketopiperazine." *J. Am. Chem. Soc.* 69: 1598–1604.

Corey, Robert B. 1940. "Interatomic Distances in Proteins and Related Substances." *Chem. Rev.* 26: 227–36.

Hacking, I. 1992. The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Science as practice and culture* (pp. 29–64). Chicago: The University of Chicago Press.

Hager, Thomas. 1995. Force of Nature: The Life of Linus Pauling. New York: Simon and Schuster.

Hall, Kersten. 2014. *The Man in the Monkeynut Coat: William Astbury and How Wool Wove a Forgotten Road to the Double-Helix*. Oxford: Oxford University Press.

Harris, Todd. 2003. "Data Models and the Acquisition and Manipulation of Data." *Philosophy of Science* 70 (5): 1508–17.

Hon, Giora. 2003. "The Idols of Experiment: Transcending the 'Etc. List'" in Hans Radder (ed.) *The Philosophy of Scientific Experimentation*. Pittsburgh: University of Pittsburgh Press.

Hughes, Edward W. & Moore, Walter J. 1942. "The Structure of β Glycylglycine." *J. Am. Chem. Soc* 64: 2236–37.

Hughes, E. W. & Moore, W. J. 1949. "The Crystal Structure of β-Glycylglycine." *J. Am. Chem. Soc* 71: 2618.

Judson, Horace F. 1996. *The Eighth Day of Creation*. New York: Cold Spring Harbor Laboratory Press.

Karaca, Koray. 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): 5431–52.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. University of Chicago Press.

———. 2019. "What Distinguishes Data from Models?" *European Journal for Philosophy of Science* 9 (2): 22.

Levy, H. A. & Corey, R. B. 1941. "The Crystal structure of dl Alanine." *J. Am. Chem. Soc.* 63: 2095–2108.

MacAllister, James. 1997. "Phenomena and Patterns in Datasets." *Erkenntnis* 47 (2):217-228.

MacArthur, Ian. 1961. "Prof. W. T. Astbury, F. R. S." *Nature* 191, 331-32.

McCoy, R. H., Meyer, K. H., & Rose, W. C. 1935. "The Connection of Glutathione with the Action of Insulin." *Journal of Biological Chemistry*, 112, 283-302.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Meyer, K. H., & Rose, W. C. 1936. "The Nucleic Acids of Plant Tissues. I. The Separation of Purines, Pyrimidines, Nucleosides, and Nucleotides by Paper Chromatography." *Journal of Biological Chemistry*, 115, 271-290.

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): 703–23.

Olby, Robert. 1994. *The Path to the Double Helix: The Discovery of DNA*. New York: Dover Publications.

Parry, D. A. 2021. 'Structures of the ß-Keratin Filaments and Keratin Intermediate Filaments in the Epidermal Appendages of Birds and Reptiles (Sauropsids)'. *Genes*, 12(4), 591.

Pauling, L. 1928. "The Shared-Electron Bond." *Proceedings of the National Academy of Sciences*

of the United States of America, 14(4), 359-362.

Pauling, L., Brockway, L. O. & Beach, J. Y. 1935. "The Dependence of Interatomic Distance on Single Bond–Double Bond Resonance." *J. Am. Chem. Soc.* 74: 2705–09.

Pauling, L., Corey, R. B. & Branson, H. R. 1951. "The Structure of Proteins: Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain." *Proc. Natn. Acad. Sci. U.S.A.* 37: 205–11.

Pauling, L. & Sherman, J. 1933. "The Nature of the Chemical Bond. VI. The Calculation from Thermochemical Data of the Energy of Resonance of Molecules Among Several Electronic Structures." *J. Chem. Phys.* 1: 606–17.

Sali, Andrej, Robert Glaeser, Thomas Earnest, and Wolfgang Baumeister. 2003. "From Words to Literature in Structural Proteomics." *Nature* 422 (6928): 216–25.

Shoemaker, D. P., Donohue, J., Schomaker, V., & Corey, R. B. 1950. The Crystal Structure of L_S-Threonine. *Journal of the American Chemical Society, 72(5)*, 2328-2335.

Staley, Kent. 2004. "Robust Evidence and Secure Evidence Claims." *Philosophy of Science 71*: 467–488.

Suppes, P. 1962. Models of data. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science*. Stanford: Stanford University Press.

Woodward, Jim. 2000. "Data, Phenomena, and Reliability." *Philosophy of Science*, 67, pp. S163-S179.

Wyckoff, R. W., & Corey, R. B. 1934. "Spectrometric measurements on hexamethylene tetramine and urea." *Zeitschrift für Kristallographie-Crystalline Materials* 89(1-6): 462-68.