

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Biochemical Kinds and Selective Naturalism

Permalink

<https://escholarship.org/uc/item/75m6x8hk>

Author

Havstad, Joyce Catherine

Publication Date

2014

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Biochemical Kinds and Selective Naturalism

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Philosophy (Science Studies)

by

Joyce Catherine Havstad

Committee in charge:

Professor William Bechtel, Chair
Professor Craig Callender
Professor Nancy Cartwright
Professor Cathy Gere
Professor James Griesemer

2014

©

Joyce Catherine Havstad, 2014

All rights reserved.

The Dissertation of Joyce Catherine Havstad is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014

DEDICATION
for my parents

TABLE OF CONTENTS

Signature Page.....	iii
Dedication.....	iv
Table of Contents.....	v
List of Figures.....	viii
Vita.....	x
Abstract of the Dissertation.....	xi
Introduction.....	1
References.....	7
Chapter 1: Messy Chemical Kinds.....	8
Section 1: Introduction.....	8
Section 2: Microstructuralism about Chemical Kinds.....	11
Sub-Section 2.1: Neither Strictly.....	16
Sub-Section 2.2: Nor Solely.....	18
Sub-Section 2.3: Hendry on Elements.....	21
Sub-Section 2.4: Hendry on Compounds.....	25
Section 3: A Different Kind of Chemical Kind.....	29
Sub-Section 3.1: Basic Protein Science.....	32
Sub-Section 3.2: Individuating Proteins.....	34
Sub-Section 3.3: Proteins, Microstructuralism, and Unification.....	40
Sub-Section 3.4: Slater, Tobin, and Goodwin.....	45
Section 4: The Argument from Iso-Ism.....	53
Section 5: Conclusion.....	55
References.....	58
Chapter 2: Protein Taxa.....	60
Section 1: Introduction.....	60
Section 2: A Primer on Proteins.....	65
Section 3: The Classification of Proteins.....	71
Sub-Section 3.1: Superstructural Considerations.....	73
Sub-Section 3.2: Functional Considerations.....	75
Sub-Section 3.3: Evolutionary Considerations.....	77
Section 4: Classificatory Complications.....	80
Sub-Section 4.1: Nuclear Receptors.....	83
Sub-Section 4.2: A Brief Note about Reference.....	86
Sub-Section 4.3: Phylogenies of the NHRSF.....	87
Sub-Section 4.4: Types of NRs.....	89

Sub-Section 4.5: The NHRSF by Expression Pattern.....	91
Sub-Section 4.6: Implications.....	93
Section 5: Conclusion.....	99
References.....	102
Chapter 3: Complex Objects.....	104
Section 1: Introduction.....	104
Section 2: Complexity.....	112
Section 3: Extending Mitchell's View.....	117
Section 4: Proteins as Complex Objects.....	122
Section 5: Conclusion.....	136
References.....	138
Chapter 4: Classification as a Tool for Discovery.....	140
Section 1: Introduction.....	140
Section 2: Complications.....	142
Section 3: Constraints.....	145
Sub-Section 3.1: Inaccessibility.....	145
Sub-Section 3.2: Information Overload.....	146
Sub-Section 3.3: Diffuse Effects.....	148
Sub-Section 3.4: Strain on Resources.....	149
Section 4: Classification as a Tool for Discovery.....	150
Sub-Section 4.1: Origins.....	151
Sub-Section 4.2: Application.....	153
Sub-Section 4.3: Evidence.....	155
Section 5: Demonstrating the Concept.....	158
Section 6: Conclusion.....	164
References.....	166
Conclusion.....	167
References.....	172
Appendix A.....	173
Section 1: Introduction.....	173
Section 2: Protein Science.....	174
Section 3: Molecular Biology.....	178
Section 4: Endocrinology.....	181
Section 5: The Nuclear Hormone Receptor Superfamily.....	185
Section 6: Conclusion.....	191
References.....	194
Appendix B.....	197
Section 1: Introduction.....	197
Section 2: GEL-X.....	203
Section 3: Applying the Ideas of <i>Laboratory Life</i> to GEL-X.....	213

Sub-Section 3.1: The Daily Activity of Science.....	214
Sub-Section 3.2: Literary Inscription and Origin Stories.....	220
Sub-Section 3.3: Statement Types and Cycles of Credit.....	225
Sub-Section 3.4: The (Social) Construction of Scientific Facts.....	233
Sub-Section 3.5: The Creation of Order out of Disorder.....	235
Section 4: Conclusion.....	236
References.....	239

LIST OF FIGURES

Figure 2.1: Four levels of proteins structure.....	68
Figure 2.2: A helpful pair of diagrams from <i>Lehninger's Principles of Biochemistry</i> . On the left is a depiction of the tetravalent nature of carbon; on the right is a schematic of the basic structure of amino acids.....	68
Figure 2.3: Diagram of a (zinc finger) protein with both alpha helixes and beta sheets. The curly-cues or telephone-cord shapes represent the alpha helixes; the arrows represent the beta sheets.....	74
Figure 2.4: The DNA-binding helix-turn-helix motif.....	74
Figure 2.5: A snapshot that shows the superfamilies, families, subfamilies, and entries identified by the Atlas project by 1983, as well as the criteria of establishment for each division.....	79
Figure 2.6: On the left, the taxa of standard biological classification, progressing up the hierarchy from least to most inclusive. On the right, a prospective taxonomy for protein classification, similarly structured from least to most inclusive categories.....	82
Figure 2.7: A phylogeny of the nuclear hormone receptor superfamily.....	87
Figure 2.8: The five domains of a nuclear receptor.....	89
Figure 2.9: A physiologic nuclear hormone receptor superfamily.....	92
Figure 2.10: The PBD graphic of RAR α	94
Figure 2.11: From left to right, the taxonomic structures of the nuclear hormone receptor superfamily by phylogeny, typing, and expression pattern.....	95
Figure 3.1: Diagram of an alpha helix.....	126
Figure 3.2: Structure of the estrogen receptor (ER) ligand-binding domain.....	127
Figure 3.3: On the left, a protein with a triple zinc finger motif in complex with DNA. On the right, a repressor with a helix-turn-helix motif DNA-binding domain, also in complex.....	129
Figure 3.4: Three distinct classifications of the nuclear hormone receptor superfamily, one from each of the different classification systems.....	131
Figure 4.1: Step 1 – Indicate the different types of nuclear receptor with different shapes in a mechanistic nuclear hormone receptor superfamily.....	159

Figure 4.2: Step 2 – Associate a different color with each cluster of nuclear receptors in the physiologic nuclear hormone receptor superfamily, while also importing the type-indicating shapes from the mechanistic superfamily.....	160
Figure 4.3: Step 3 – Place an appropriately colored and shaped indicator next to each nuclear receptor in a phylogenetic nuclear hormone receptor superfamily.....	161
Figure B.1: On the left, the Salk Courtyard, facing the Pacific Ocean; on the right, a view of the North Building from the lawn.....	204
Figure B.2: Diagram of the interior of Guillemin’s laboratory.....	207
Figure B.3: Diagram of the interior of the GEL-X laboratory.....	208
Figure B.4: A comparison of laboratory hierarchies.....	211

VITA

2001–2007 Research Assistant, The Salk Institute for Biological Studies
2004 Bachelor of Arts, University of California, San Diego
2007 Master of Arts, San Diego State University
2008–2010 Editorial Assistant, *Philosophical Psychology*
2014 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

“Problems for Mechanism and Natural Selection” in *Philosophy of Science* 78(3): 512–523. July 2011.

“Human Reproductive Cloning: A Conflict of Liberties” in *Bioethics* 24(2): 71–77, February 2010.

“Biological Individuality and the Unit of Selection”— Unpublished Master’s Thesis, San Diego State University, May 2007.

“The Price of Dignity: Kantian Accounts of Moral Responsibility to Non-Human Entities” in *Dialogue* 49(2–3): 99–107, April 2007.

FIELDS OF STUDY

Major Fields: Philosophy and Science Studies

Studies in Philosophy of Science
Professors Craig Callender and Nancy Cartwright

Studies in Philosophy of Biology
Professors William Bechtel and James Griesemer

Studies in Science and Technology Studies
Professors Cathy Gere and Robert Westman

ABSTRACT OF THE DISSERTATION

Biochemical Kinds and Selective Naturalism

by

Joyce Catherine Havstad

Doctor of Philosophy in Philosophy (Science Studies)

University of California, San Diego, 2014

Professor William Bechtel, Chair

In this dissertation I show how the complexity of biological and chemical kinds often complicates attempts to classify these kinds. I construct an account—which I call ‘selective naturalism’—that explains the diverse yet grounded classification of complex biochemical objects like proteins. By distinguishing the properties and powers of these complex objects, as well as the inferences afforded by tracking sets of these properties and powers via classification, I can explain how various classificatory complications evident in the biological and chemical sciences are generated. I can also show how this kind of classification may be used as a tool for discovery in the sciences.

INTRODUCTION

But the tradition from Aristotle down through the beginning of the twentieth century was that the hereditary elements, whatever they were and however they worked, determined both the transmission of traits and their development.

Wimsatt's *Re-Engineering Philosophy for Limited Beings* (2007)

This is a dissertation about biological and chemical kinds. It is especially about one particular group of kinds, proteins, which span the border between the scientific worlds of chemistry and biology. In chemistry, proteins are huge, and in more than one sense of the word: they are extremely large molecules; there are lots of them; and they are incredibly important to many current research programs. In biology, proteins are both everywhere and nowhere. As the products of gene expression, proteins are ubiquitous. But they are also rather neglected, relative to their genetic counterparts. As the lead quote of this introduction indicates, it is the hereditary material that has long been the focus of those studying the biological world, in both the scientific and philosophical traditions.¹

But that focus is changing. Proteins have now been the subject of chemical investigation for more than two centuries.² And over the last several decades, the role of proteins in biological studies has been increasing exponentially.³ Philosophers are even starting to turn their attention to proteins.⁴ And there is good reason for this. As this dissertation shows, proteins are philosophically fascinating. They are complex macromolecules, and an account of their complexity can be used to both challenge and

¹ Of course, in Aristotle's case and for many who followed after him, these were the same traditions.

² Proteins were first distinguished by Antoine Fourcroy in 1789; the term 'protein' was coined in 1838 by Jöns Jakob Berzelius (Tanford & Reynolds, 2001, chapter 1).

³ In 1983, Margaret O. Dayhoff's *Atlas of Protein Sequence and Structure* contained less than two thousand individual protein entries. Now estimations of the number of known proteins ranges up to somewhere around 65 million.

⁴ For instance, see Slater (2009); Tobin (2010); Goodwin (2011).

enrich many of our current philosophies of science.⁵ As I'll demonstrate throughout the dissertation, a study of proteins has much to contribute to ongoing debate about natural kinds, microstructuralism, scientific classification, complexity, and the context of discovery (just to name a few key topics).

In addition to this introduction, a summarizing conclusion, and two appendices, the dissertation contains four body chapters. The first and second chapters are about biochemical classification, but at distinct levels. Chapter 1 focuses on the individuation of various chemical kinds—including elements and proteins. Chapter 2 deals with classification at higher taxonomic levels—it is concerned with the organization of proteins, post-individuation. In the third chapter an account of proteins as complex objects is constructed, and then used to explain the classificatory complications of chapter 2. In chapter 4 a novel aspect of scientific practice—the use of classification as a tool for discovery—is reported, and then also explained via the account of complex objects. After the summarizing conclusion there is a historical and then a sociological appendix.

In what remains of this introduction I'll describe, in a bit more detail, the philosophical puzzles that this dissertation addresses—using a careful study of proteins, and at times other biochemical kinds as well. I will provide a schematic outline of the selected puzzles and my response to each of them, proceeding chapter-by-chapter. This should equip the reader for evaluation of the puzzles, responses, and arguments as they

⁵ Throughout this dissertation, I will do my best to reserve the term 'complex' for ontological ascriptions and 'complicated' for epistemological ones. In fact, although I want to be quite liberal with respect to what constitutes complexity—using the term to mean something as simple as “has lots of interacting parts” or some such—I want to be quite strict in applying the term to *things*, rather than to *understanding*. When something is difficult to understand I will call it complicated rather than complex.

proceed throughout the document.

I begin with a discussion of chemical kinds. Though certain biological and chemical kinds (particularly species and elements) once shared a somewhat celebrated status as paradigmatically natural kinds, biological kinds are now considered rather messy and difficult. Chemical kinds, however, have retained their paradigmatically natural status. In the first chapter of the dissertation (“Messy Chemical Kinds”) I argue that this divisive characterization—of chemical kinds as paradigmatically natural, and biological kinds as messy—is misleading. I focus on the chemical side of things, showing that neither elements nor compounds behave in a particularly uniform or simple way. I argue that chemical kinds need to be evaluated on a case-by-case basis, and that uniform judgments about the naturalness of chemical kinds are unsupported.

My position is at odds with the received view. The pre-eminent account of chemical kinds is Robin F. Hendry’s (e.g., 2006). Hendry’s account of chemical kinds is *microstructuralist*—meaning the account posits that membership in chemical kinds is conferred by microstructural properties. But I claim that Hendry’s account does not accurately characterize chemical kinds, because not even in Hendry’s flagship case of elements is membership in chemical kinds simply conferred by microstructure.

As I show in chapter 1, membership in element kinds is neither strictly nor solely conferred by microstructural properties. Rather, membership in element kinds is conferred by a *combination of factors*—including the microstructural property of nuclear charge, correlated macroscopic features of chemical behavior, and the vested interests of chemists in those chemical behaviors. Additionally, elements are not the only kind of chemical kind; Hendry, for instance, also singles out compounds and mixtures for special

attention within his account of chemical kinds. Almost half of all registered compounds are proteins or protein-coding sequences, so I also look at the conditions for membership in protein kinds. I show that here, too, membership is conferred by a combination of factors—but in this case, these factors span microstructural, mechanical, functional, and evolutionary properties.

Proteins are newly of interest to philosophers, and thus far the discussion has been structured as a choice between pluralism or monism with respect to these macromolecules. But as I show in chapter 2 (“Protein Taxa”), this is a false choice—there are both monistic and pluralist aspects to the classification of proteins, located at different points in the classificatory process. Proteins are incredibly complex macromolecules, yet despite this complexity, there is robust scientific consensus with respect to the individuation of particular proteins. This consensus, however, does not hold at higher levels of protein classification—there is no sole, comprehensive way of organizing individuated proteins with respect to one another. In other words, although token proteins are consistently and routinely typed, these protein types can be, and often are, related to one another in different ways. Both of these activities are classificatory ones, but by making a distinction between the former (*individuation*) and the latter (*organization*), I can precisely characterize the state of protein classification. At least as it currently stands, protein individuation is overwhelmingly monist, while protein organization is thoroughly pluralist.

This leads to a discussion in chapter 3 (“Complex Objects”) of complex objects in general, and how their complexity complicates attempts to classify them. Complexity has long been a topic of interest in the philosophy of biology (e.g., Wimsatt 1972), but it has

recently been the focus of renewed attention (e.g., Mitchell 2003, 2009). I adjust Sandra D. Mitchell's account of complex systems in order to apply it to complex objects; then I use the adapted account to explain complex object classification. I argue that organizational pluralism arises despite monistic individuation of complex objects because complex objects have many *properties*, which interact to generate different *powers*, many of which grant different *affordances*. This complex framework of properties, powers, and affordances allows for a selection process to occur in the organization of these individuated objects. I call this process by which the classification of complex objects can occur 'selective naturalism'. Incentivized by the granting of different affordances, different sets of properties and powers are picked out and tracked, giving rise to a diverse set of grounded classification schemes—to a kind of constrained classificatory pluralism.

Finally in chapter 4 ("Classification as a Tool for Discovery"), I explain how this constrained classificatory pluralism can be used to help rather than hinder scientific practice. Other philosophers have described cases of *interdisciplinary classificatory pluralism* (e.g., Dupré 1993). But in my study of proteins, I have observed *intradisciplinary classificatory pluralism*—the use of different classification schemes in not just one particular field, but also within a single lab, and even by individual scientists. Drawing on extensive historical, sociological, and scientific immersion in a sub-field of protein science, I show how intradisciplinary classification schemes can be used to overcome certain constraints on research, and to direct research towards those places where progress is most likely to be made. Though philosophers have worried about the negative effects of classificatory pluralism on scientific investigation (e.g., Ereshefsky 1998), I document something positive—a case whereby selective naturalism fashions

pluralist classification into a tool for discovery.

So, the final chapter of the dissertation both documents a novel aspect of scientific practice—the use of classification as a tool for discovery—and explains how this practice works and why—drawing again on the account of complex objects and their classification. In the conclusion I chart the progress of the dissertation through the relevant territories in protein science, complex classification, and the study biochemical kinds. Implications of the arguments and possible future avenues of application are explored.

Finally, there are two appendixes to the dissertation. Much of the information about protein science presented throughout the dissertation was gathered through extensive historical and sociological study. Appendix A is the historical appendix: it provides a history of protein science in general and nuclear receptor research in particular. These historical facts inform much of the dissertation, but particularly chapters 2 and 4. Appendix B is the sociological appendix: it documents a laboratory study that I conducted in order to give myself broad familiarity with the experience of conducting research in molecular biology. It informs all those aspects of the dissertation that pertain to scientific practice, but especially chapter 4.

REFERENCES

- Dayhoff, M. O., Barker, W. C., & Hunt, L. T. 1983. Establishing homologies in protein sequences. *Methods in Enzymology* 91: 524–545.
- Dupré, J. 1993. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.
- Ereshefsky, M. 1998. Species pluralism and anti-realism. *Philosophy of Science* 65: 103–120.
- Goodwin, W. 2011. Structure, function, and protein taxonomy. *Biology & Philosophy* 26: 533–545.
- Hendry, R. F. 2006. Elements, compounds, and other chemical kinds. *Philosophy of Science* 73: 864–875.
- Mitchell, S. D. 2003. *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
- Mitchell, S. D. 2009. *Unsimple truths: Science, complexity, and policy*. Chicago: University of Chicago Press.
- Slater, M. H. 2009. Macromolecular pluralism. *Philosophy of Science* 76: 851–863.
- Tanford, C., & Reynolds, J. 2001. *Nature's robots: A history of proteins*. Oxford: Oxford University Press.
- Tobin, E. 2010. Microstructuralism and macromolecules: The case of moonlighting proteins. *Foundations of Chemistry* 12: 41–54.
- Wimsatt, W. C. 1974. “Complexity and organization,” in K. F. Schaffner & R. S. Cohen (eds.), *PSA 1972*, 67–86. Dordrecht: D. Reidel.
- Wimsatt, W. C. 2007. *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

CHAPTER 1: MESSY CHEMICAL KINDS

Now these arguments—against microstructuralism and for pluralism—are local to biology. Chemistry, I argue, is different: the interests that govern its classifications are more unified, and in most cases membership of its kinds is conferred by microstructural properties.

Hendry's *Elements, Compounds, and Other Chemical Kinds* (2006)

Abstract: The received view of chemical kinds is microstructuralist. To be microstructuralist about a kind is to think that membership in said kind is conferred by microstructure, or by microstructural properties. Although I agree that microstructure is a crucial component of chemical kindhood, I do not think that microstructure is the only component of chemical kindhood. To put the point another way: there is ambiguity concealed by the term ‘conferred’. Although membership in chemical kinds is partially conferred by microstructural properties, it is not conferred by all microstructural properties (strictly) or only by microstructural properties (solely). Here I argue that neither in the flagship case of elements, nor in the wide-ranging case of proteins, is membership in chemical kinds solely or strictly conferred by microstructure.

Section 1: Introduction

It is common knowledge in the philosophy of science that biological kinds are messy. Biological species, once thought to be paradigmatic examples of natural kinds, are generally considered natural kinds no longer—at least, they’re no longer thought of as straightforwardly natural kinds. If these kinds are still “natural” (whatever that means), then their naturalness must be complicated in some way—for example, perhaps biological kinds are cluster kinds (e.g., Boyd 1991, 1999), or some such thing (e.g., Millikan 1999; Ellis 2001; Chakravartty 2007).

But chemical kinds remain stalwart natural kinds—for evidence of this claim, just consider the frequency and stature of statements like “gold is the element with atomic number 79” and “water is H₂O” in philosophical discourse.¹ Now one might ask, “why is gold atomic number 79 and water H₂O?” And the standard answer is “microstructure.” Relatedly, “why are biological kinds messy and chemical kinds, well, not messy?” Again, the common reply is “microstructure.”

Take the currently predominant account of chemical kinds: Robin Findlay Hendry’s. Hendry is a microstructuralist with respect to chemical kinds (and author of the lead quote). As Hendry defines it, “microstructuralism about a natural kind is the thesis that membership of that kind is conferred by microstructural properties” (2006, p. 865). Correspondingly, microstructuralism about a chemical kind (gold, water, etc.) is the thesis that membership in that chemical kind is conferred by microstructural properties (atomic number, molecular components, etc.).

In his “Elements, Compounds, and Other Chemical Kinds” (2006), Hendry argues that chemical kinds in general are microstructuralist—on the basis of elements (like gold) and one compound (water). But I think that this move is tenuous at best, for at least three reasons. For one, there are literally millions of compounds, of which water is just one (and not a particularly representative one at that). So it is a mistake to infer even from the case of water to other compounds.

For another, there are myriad chemical kinds. Hendry divides them initially into elements, compounds, and mixtures (2006, p. 864)—but later he admits of more

¹ Primarily thanks to Kripke (1972/1980) and Putnam (1973), examples of these cases are simply too numerous to survey. For a general introduction to the issue, please see the SEP’s entry on “Natural Kinds” (Bird & Tobin, 2008).

categories, additionally including atoms, ions, molecules, metals, and acids (p. 873). He also mentions isotopes and isomers at other points in the piece. Again, there is a problem with sample size: the case of elements plus water is an insufficient basis for extrapolation to all chemical kinds.

Finally, and most philosophically interesting, not even elements are purely microstructuralist kinds—in other words, membership in element kinds is not entirely conferred by microstructure alone, by all or only microstructural properties. Or so I shall argue here.

In this chapter I use facts about elements, as well as proteins and other messy chemical kinds, to argue against the received microstructuralist view of chemical kinds. I make two main claims:

- i. Membership in chemical kinds is not, even in the case of elements, strictly or solely conferred by microstructure.
- ii. Chemistry does not have a special status (relative to, say, biology) as an exemplary science with unified (i.e., purely microstructuralist) interests that govern its classification.

My first claim constitutes a dispute with Hendry and other microstructuralists about the characterization of certain chemical kinds as microstructural kinds. My second claim constitutes a dispute with the inference from these particular cases to chemical kinds in general, and in contrast with biological kinds. Claim (ii) follows from claim (i).

I begin the argument for claim (i) in section 2, focusing on the case of elements (with some discussion of water), while also laying some groundwork for the argument for claim (ii). In section 3 I first present the case of proteins as another challenge to the microstructuralist, and then make the argument for claim (ii) in earnest. Several other

philosophers have recently discussed proteins in relevant ways (e.g., Slater 2009; Tobin 2010; Goodwin 2011), and I respond to their work at the end of the section. In the penultimate section of the chapter, section 4, I consider a potential objection to section 3's protein-based case against microstructuralism, and then offer a reply—which I call the argument from iso-ism—that draws on various different kinds of chemical kinds. In section 5, the conclusion, I summarize the arguments for and consider some implications of both my first and second claims.

My conclusion is that the situation with respect to what confers membership in chemical kinds, and what status these kinds therefore have, is complicated—just as it is for biological kinds.² Although I agree with Hendry and most other philosophers of chemistry that microstructure is a crucial component of chemical kindhood, I do not think that microstructure is the only component of chemical kindhood. Neither in the flagship case of elements, nor the immense set of macromolecular compounds generally known as proteins, does any simple or unqualified version of microstructuralism look tenable. The microstructuralist is left with very little to support either their view of chemical kinds as purely and simply microstructuralist, or their image of chemistry as a field with unified classificatory interests.

Section 2: Microstructuralism about Chemical Kinds

Let me reiterate the received microstructuralist view of chemical kinds. To be microstructuralist about a kind is to think that membership in said kind is conferred by

² Though the ways in which membership in biological and chemical kinds are complicated are not necessarily the same between the fields, or even among the different cases within a field.

microstructural properties. So a microstructuralist about chemical kinds thinks that membership in chemical kinds is conferred by microstructural properties.

To some degree, I concur with the microstructuralist that membership in chemical kinds is conferred by microstructural properties. But I'd like to separate two senses of the term 'conferred' at work here. I think that in what might be called a logical, or first-order sense of 'conferred', membership in chemical kinds *is often conferred by* microstructural properties. Membership in element kinds, for example, is strictly and logically conferred by one particular microstructural property, the number of protons an atom has.³

But in a historical, or second-order sense of 'conferred', I think that membership in chemical kinds *is not conferred by* microstructural properties. In the case of elements, for instance, I claim that membership is historically conferred by the vested interests of chemists in assorted macroscopic properties like chemical stability and reactivity, rather than microstructure.

To put the point another way, consider two questions. First is the straightforwardly logical query: what determines membership in element kinds? To which both the microstructuralist and myself will answer: number of protons. But now consider a second, more contextual question: what determines what determines membership in element kinds? To which I answer: the interest of chemists in certain, somewhat correlated macroscopic features, and their decision to settle on those features and properties as the relevant ones.

³ By 'strictly' here I simply mean that members of the kind completely converge with respect to this property. As in, all members of the kind share the property. More on 'strictly' in a bit—and I should also note that in many chemical case, microstructure is only one component of what confers membership in the kind, even in the logical sense of 'confers'. Again, more on this later.

I'll expand on and defend these claims in what follows. But let me situate the example I've been using (elements) within the broader context (chemical kinds in general).

Chemistry is a broad field, dedicated to the study of matter. The basic chemical units of matter are atoms, and atoms of the same kind are called elements. Different elements can be combined together into compounds or mixtures. Compounds are those aggregates of different elements within which various kinds of bonds⁴ have formed. Mixtures are mere aggregates, within which bonds have not formed.

Bonded atoms are called molecules. There are some homonuclear molecules—molecules formed out of atoms of only one kind of element—like O₂, which consists of two bonded⁵ oxygen atoms. But many other molecules are heteronuclear—they are composed of atoms of different kinds of element—like H₂O, which is composed of two hydrogen atoms bonded⁶ to one oxygen atom. Heteronuclear molecules are more commonly called compounds. Especially large molecules and compounds are called macromolecules. Proteins, for instance, are an example of this kind of chemical kind—they are macromolecules.

In the introduction to “Elements, Compounds, and Other Chemical Kinds,” Hendry divides all chemical substances into the three groups of elements, compounds, and mixtures (2006, p. 864).⁷ In section 2 of his paper, Hendry explains the gradual

⁴ Some of the most interesting questions in the philosophy of chemistry have to do with how bonding works, whether bonds are real, etc. For an introduction to these issues, see the SEP's entry on “Philosophy of Chemistry” (Weisberg, Needham, & Hendry, 2011) or Weisberg & Needham (2010).

⁵ The ground state of O₂ is a triplet state. This means that the oxygen atoms in O₂ are usually bonded together in a spin triplet electron configuration. In simpler terms, it's a double covalent bond.

⁶ Again in simple terms, covalently.

⁷ Throughout this section, I'll continue to use Hendry (2006) as representative of the received

convergence of chemists on the account of elements as chemical kinds defined by the microstructural property of nuclear charge.

It's a good story: in 1923, the International Union of Pure and Applied Chemists (IUPAC) met in order to settle the question of what elements were. Following Dalton, chemists had long thought that atoms of the same kind constituted elements. Mendeleev's periodic table was formed on the basis of atomic weight—a macroscopic property thought to pick out atoms of the same microstructural kind.

But it was eventually determined that some of Mendeleev's elements contained sub-populations of atoms with slightly variant atomic weights, rather than purely uniform ones. As it turns out, atomic weight tracks the number of protons and neutrons that an atom has, whereas atomic number tracks simply the number of protons, or nuclear charge of an atom.

So chemists had to decide: what are elements? Are atoms of “the same kind” ones that share both number of protons and neutrons, or just number of protons? And after some very heated debate, chemists settled on just the number of protons as sufficient—a decision that still holds today. Throughout chemical practice, atoms of an element always share number of protons, but they might diverge in number of neutrons (these are called isotopes of an element).

But chemists settled on that microstructural property (number of protons) instead of the other (number of protons and neutrons) because of the former's correlation with what they really cared about: the macroscopic behavior of atoms grouped in one or the other way. In other words, chemists at the time thought that chemical reactivity, stability,

and other behavior was determined far more by number of protons than by number of protons and neutrons, so they chose number of protons as the relevant property.

Of course, this doesn't hold true for all groups of atoms (the isotopes of hydrogen being the most notable exception), but for our purposes the point is simply that chemists had a choice. Microstructure itself doesn't settle the question of what elements are—they could have been atoms with the same number of protons and neutrons, or just the same number of protons.⁸

As should already be clear, the historical story has philosophical import—in particular, it is highly relevant to my claim (i). For, although there is now general agreement about how to classify atoms into the kinds called elements (of which there are currently 118⁹), this classification is not at all straightforwardly dictated by microstructure. I'd like to put it thusly: atomic classification into element kinds is not straightforwardly dictated by microstructure because it is neither strictly nor solely

⁸ And in fact they might have been, in another possible, not-so-distant world. For instance: the chemist who discovered the divergent chemical reactivity of the isotopes of hydrogen was Ida Noddack, a woman. In the 1930s she argued, on the basis of these and others' discoveries, for the redefinition of element kinds and corresponding reconstitution of the periodic table (for more, please see Kraugh 2000). Perhaps in a less misogynistic world, her proposal would have been taken more seriously? Regardless, the point is simply that it is obviously possible that the membership conditions in element kinds might have been different than what they now in fact are.

Just to be clear, this puts me at odds with Kripke (1980/1972)—particularly his account of the *a posteriori* necessity of the particular microstructural conditions of kind membership. My position here is much closer to that of Kitcher (2001). Kitcher writes: “in the biological sciences, actual practice reveals examples of how different purposes demand different categories” (Kitcher 2001, p. 51). His discussion of chemical cases is much more speculative—Kitcher suggests that other creatures with different perceptual apparatuses might derive distinct physical and chemical kinds on the basis of their own interests, “bizarre as they might be by our own lights” (ibid.). Part of what I'm doing in this chapter is showing that we don't have to speculate about aliens with bizarre, alternative perceptual apparatuses and interests in order to make the point that chemical kinds, like biological kinds, are partially dependent on scientific interests and practice.

⁹ Though not all 118 of these are officially recognized by IUPAC, the chemical review body that determines priority of discovery and right to naming of elements. For example, in 2011 various claims to have “discovered” elements 113–116 and 118 were evaluated. The IUPAC Joint Working Party credited the discovery of elements 114 and 116, but decided that elements 113, 115, and 118 “have not met the criteria for discovery” (Barber et al. 2011, p. 1994).

dictated by microstructure.

Sub-Section 2.1: Neither Strictly

I'll explain each of these qualifications (neither strictly, nor solely) in turn. First things first: classification of atoms into elements is not strictly dictated by microstructure *simpliciter* because the atoms within elements are not microstructurally identical (numerically or qualitatively). In other words, it would be a mistake to think that atoms of the same element-kind necessarily have the same microstructure.¹⁰ Rather, atoms of the same element-kind are necessarily similar with respect to only one aspect of their microstructure (nuclear charge, or number of protons). The existence of isotopes within element kinds shows that atoms of the same element-kind are, despite this dimension of similarity, also dissimilar with respect to other aspects of their microstructure (such as atomic weight, or number of neutrons).

Many elements contain various stable isotopes. For example, carbon (which has atomic number 6, for 6 protons) comes in three standard types, or stable isotopes: carbon-12, carbon-13, and carbon-14. These atoms have 6 protons and then (respectively) 6, 7, and 8 neutrons each. This means that within the kind of element known as 'carbon' there are various kinds of atoms that are similar in some microstructural respects (e.g., number

¹⁰ I'm not accusing anyone of making this mistake. I'm merely trying to be extremely precise, so that we can distinguish among the various ways kindhood might be conferred by microstructure. Being precise has advantages: for example, one of the claims I'll make later in this chapter is that, although both element kindhood and protein kindhood are conferred by microstructural properties, they are conferred in different ways and to a different degree. What I've said here about strict conferral will help me to distinguish these cases.

However: it might also be worth noting here that, following Dalton (who died in 1844), for decades many chemists did believe that elements were composed of identical atoms. The fact that isotopes existed meant they were not, and that was a strong enough reason for IUPAC, in 1923, to propose to "redefine 'element' in light of isotopy" (Hendry 2006, p. 867).

of protons) but different in others (e.g., number of neutrons).

Another way of putting the point is that sharing values with respect to the microstructural property of number of protons puts atoms in the same element kind, while sharing the same number of neutrons doesn't necessarily do the same. Having different numbers of neutrons doesn't necessarily make two atoms different kinds of element; having different numbers of protons does. Or, element kinds contain various atom kinds—the isotopes of that element. Atoms of the same element but different isotopes will converge in terms of at least one microstructural property (number of protons) but will diverge in at least one other (number of neutrons).

Because of this aspect of divergence, I'm claiming that atomic classification is not strictly conferred by microstructure *simpliciter*. But there is an important sense in which microstructure—or, to be more precise, a microstructural property—does strictly confer membership of an atom into an element-kind. Atomic classification is strictly (in this second sense) conferred, because although atoms of the same element can have different numbers of neutrons, they necessarily have the same number of protons. In other words, despite divergence along some aspects of microstructure, along the aspect of microstructure on which atoms of an element are sorted and do converge, this convergence is complete.

I'd like to distinguish these two senses of the term 'strict' with respect to how kind membership can be conferred by microstructure. The term 'microstructuralism' is generally used to denote any claim that membership of a kind is conferred by

microstructural properties.¹¹ So, we can now distinguish between the former and latter senses of ‘strict’ when it comes to microstructuralism:

ABSOLUTELY STRICT MICROSTRUCTURALISM—members of a kind completely converge with respect to all aspects of their microstructure.

SELECTIVELY STRICT MICROSTRUCTURALISM—members of a kind completely converge with respect to those aspects of their microstructure which denote the kind.

Though element kinds do not display absolutely strict microstructuralism, they do display selectively strict microstructuralism. Other chemical kinds display neither absolutely nor selectively strict nor microstructuralism, and this difference between elements and other chemical kinds is one example of disunity among the interests that govern chemical kinds—a point that is relevant to my claim (ii).

But returning to claim (i), which was that membership in chemical kinds is not, even in the case of elements, strictly or solely conferred by microstructure. I’ve just explained why membership in element kinds isn’t strictly (in the absolute sense) conferred by microstructure; an important consequence of this fact is that membership in element kinds is not solely conferred by microstructure either.

Sub-Section 2.2: Nor Solely

Element kinds are not solely dictated by microstructure because, though the

¹¹ Definitions of the term ‘microstructuralism’ often include the term ‘microstructure’ in the definition (e.g., Hendry 2006; Tobin 2010; but see Goodwin 2011 for an exception). This trend shows that what is being defined is not quite microstructure itself, but rather what microstructure is taken to entail philosophically. It’s a definition of the ‘ism’ more than anything else.

Accounts of what is meant by ‘microstructure’ itself are usually confined to lists of examples. As in, here are some chemically relevant microstructural properties: nuclear charge, internuclear distance, atomic charge, atomic composition, chemical connectivity through bonds, angles between bonds, and stereochemistry.

microstructural property of nuclear charge is a crucial component of classificatory determination, so is the chemists' vested interest in and agreement upon this property as the relevant one. That aspect of microstructure was singled out as the (selectively strict) criterion for membership in element kinds for a reason that goes beyond microstructure itself.

In other words, both of these factors—microstructural property and scientific agreement—are vital elements (pardon the pun) in the classification of atom kinds into elements. Even more crucially, this scientific agreement (that nuclear charge should supplant atomic weight as the relevant microstructural property for atomic classification) is based on an appeal to chemical behavior—i.e., on macroscopic features rather than microstructural properties.¹²

To begin to see the role that scientific agreement plays here, consider that instead of using just the number of protons to fix atomic number, in 1923 IUPAC could have decided to make the paired number of protons and neutrons an atom has the basis of their classification of atoms into kinds. To use carbon as an example again, instead of having just ${}^6_6\text{C}$, we might have had ${}^{6:6}_6\text{C}$, ${}^{6:7}_6\text{C}$, and ${}^{6:8}_6\text{C}$ —along with ${}^{7:7}_7\text{N}$ and ${}^{7:8}_7\text{N}$.¹³ Then we would have a Periodic Table of the Isotopes rather than a Periodic Table of the Elements.

Of course, atoms have lots of different microstructural properties that chemists could pick to use as the basis for their primary atomic classification system. And some properties are going to be more plausible candidates than others. A classification system

¹² The terminology for this distinction, of microstructural properties and macroscopic features, is adopted from Weisberg, Needham, & Hendry (2011).

¹³ Keeping track of the number of protons and neutrons rather than just the total number of protons plus neutrons would be required in order to discriminate between what is currently carbon-14 and nitrogen-14, for example.

based on number of electrons, for example, would be tricky to handle. Just about all atoms would be “decaying” all the time, and very few of the “elements” would be stable.

The point of the isotope case, however, is to show that there is a plausible alternative to the status quo. There is more than one microstructural property that could be used as the basis for an effective classification system; choosing different microstructural properties would produce different classification systems; and thus chemists have to choose which of these properties to use. Additionally, deciding what makes a candidate plausible is not a solely microstructural determination.

There is certainly good microstructural reason for chemists to use number of protons as the basis for atomic classification. The fact that this property supports selectively strict microstructuralism is one such reason, and it is a good one. But it is not decisive. Because element kinds are not absolutely strictly microstructuralist, there are other candidate microstructural properties that could be used as the basis for atomic classification. The property of number of protons and neutrons is one such candidate; and this property is one that could also support selectively strict microstructuralism.

So, the fact that element kinds are selectively strictly microstructuralist kinds does not provide decisive reason for preferring a Periodic Table of the Elements to a Periodic Table of the Isotopes. Isotope kinds are also selectively strictly microstructuralist kinds. In deciding, as IUPAC did in 1923, whether to continue to classify atoms into elements or to refine atomic classification to handle isotopy, chemists appealed to chemical behavior,¹⁴ and agreed that this was relevant. Together these aspects of the decision—the appeal and the agreement—demonstrate that membership in element kinds is not solely

¹⁴ I’ll elaborate on what ‘chemical behavior’ means in the next sub-section.

conferred by microstructure.

To put the point another way, the 1923 IUPAC decision to classify atoms into elements using the microstructural property of nuclear charge rather than atomic weight on the basis of macroscopic considerations like chemical behavior shows that membership in element kinds is conferred by both microstructural and macroscopic considerations. The microstructural property (nuclear charge, or number of protons) that selectively strictly confers membership in an element kind does so because of macroscopic features of that element kind. If other macroscopic features had been deemed more relevant by the IUPAC chemists of 1923, then a different microstructural property (atomic weight, or number of protons and neutrons) might today strictly (in the selective sense), but not solely, confer membership in element kinds.¹⁵

Sub-Section 2.3: Hendry on Elements

Perhaps surprisingly, I don't think that Hendry would disagree with my account of the 1923 IUPAC decision.¹⁶ As he puts it, chemists at the time argued that:

Even though isotopes are not identical in their behavior—there are thermodynamic differences consequent on their different masses—they are chemically interchangeable, in the sense that they undergo the same chemical reactions. (Hendry 2006, p. 867)

Of course this claim, that isotopes were chemically interchangeable, turned out to be false as well. By the 1930s it was known that two isotopes of hydrogen, protium (^1H) and

¹⁵ Similarly, had Noddack's proposal been accepted in the 1930s (see footnote 8), today we would indeed have a Periodic Table of the Isotopes rather than the Elements. In fact, chemists could change their mind tomorrow, advocating a switch to number of protons and neutrons rather than just protons in the determination and layout of the Periodic Table. And this new Periodic Table would be microstructuralist to the same extent (selectively strictly, etc.) that the current one is.

¹⁶ Though I could obviously be wrong about this. I barely know what I'm thinking most of the time; I couldn't possibly be sure of what Hendry thinks.

deuterium (^2H), could be separated by electrolysis and were thus not chemically identical.¹⁷

Regardless of whether isotopes are chemically interchangeable, just the fact that this matters to the question of atomic classification is revealing. It shows that the reactions in which a substance participates play a role in the determination of that substance's chemical classification. In other words, macroscopic considerations are a crucial component of the choice between using one microstructural property and another as the basis for element kinds.

The point is that today, when we say things like “atoms of the same kind are called elements” (as I did near the beginning of this section), we are saying something that is not *determined* by microstructure—because microstructure could also support alternative claims. For instance: chemists might have privileged a different kind of microstructure (atomic weight rather than charge, say), and decided that atoms of the same kind were isotopes, rather than elements. On that reading,¹⁸ elements actually

¹⁷ And this is what Noddack used as the basis for her proposal (see footnotes 8 & 15).

¹⁸ “In that possible world,” if you prefer. Interestingly, little of this debate in the philosophy of chemistry is conducted in the formalized ways now standard to related discussions in metaphysics. In this footnote I will briefly try to connect the two conversations (as I did in footnote 8).

Again, let's distinguish two questions with respect to these kinds: the logical question and the historical question. The logical question is: what determines membership of the kind? (Or: for any chemical kind $K(x)$, what is $K(x)$'s extension?) The historical question is: what determines what determines membership of the kind? (Or: what is the interpretation function ‘I’ that fixes $K(x)$'s extension?)

The answer to the logical question is, in the case of elements: the microstructural property of number of protons. (Or: all members of the extension of $K(x)$ possess a certain microstructural property #P(x), and everything in possession of that property is in the extension of $K(x)$.)

While the answer to the historical question is: chemists at a particular time and place with a certain set of conceptions and facts underwent a sort of pragmatic epistemic process that presupposed something beyond a mere phenomenological grip on the kind. That process resulted in their agreeing that the groups formed by the microstructural property of number of protons sufficiently tracked the macroscopic chemical-behavioral properties of interest for that microstructural property to be used as the membership condition for element kinds. (Or: the interpretation function ‘I’ is fixed by the constrained stipulation of scientists, via their interest in certain macroscopic chemical-behavioral regularities, $\text{CBR}(x)$, and the

contain atoms of different kinds.

To claim that today's elements contain the same kind of atoms—to say that they are “natural” kinds, despite their microstructural differences—requires making an appeal to the relevance of macroscopic features. This is concealed in Hendry's statement:

So the isotope effect in hydrogen is an extreme case: a monster, not a paradigm. In other cases, atomic charge is the overwhelming determinant of chemical behavior, and atomic weight is a negligible factor. Now this establishes that, given the global facts about isotopic variation, IUPAC's decision to count deuterium as hydrogen was by far the most natural one. (Hendry 2006, p. 868)

Notice his use of the term ‘natural’ here. The reasons to choose nuclear charge over atomic weight might be *natural* ones, but they aren't purely *microstructuralist*. Element kinds are united by certain chemical behaviors, and that is the reason for discounting the microstructural differences between isotopes.

Hendry's exposition shows that, to chemists, chemical behavior is an important consideration in deciding whether elements should be understood to contain the same kind of atoms, or just isotopes do.¹⁹ And this chemical behavior is a macroscopic feature that is neither absolutely nor selectively strictly conferred by the microstructural property of nuclear charge, or number of protons (as the case of hydrogen demonstrates).

decision that one particular microstructural property, #P(x), tracks those macroscopic features better than another, #P&N(x).)

So, there is some agreement here with the classic Kripkean position—coming from the broadly epistemic fixing of element kinds, and because of the microstructuralist answer to the logical question. But there is also a clear disagreement with Kripke—since here there's no *a posteriori* necessity to the membership conditions of said kinds, because things could have turned out differently.

Finally: other chemical kinds are similarly complicated, and not simply microstructuralist. To use proteins as an example, these chemical kinds have still different membership conditions as a kind (obviously), not all of which conditions are microstructural, even in the logical sense, and these membership conditions are conferred neither absolutely nor selectively strictly. More on all this in the next section of the paper.

¹⁹ Weisberg, Needham, & Hendry (2011) stresses the importance of both microstructural and macroscopic properties throughout chemistry. For additional demonstration, please see their article.

To conclude my argument for claim (i): membership of atoms in element kinds is neither strictly (in the absolute sense) nor solely conferred by microstructure. Rather, it is certain microstructural properties and not others that are used to divide atoms into elements, and the choice to use some rather than others depends on a scientific agreement that takes macroscopic features into account. Hence my claim that even in the case of elements, membership in chemical kinds is neither strictly nor solely determined by microstructure.

I view this claim as more of a refinement than a dispute with Hendry and others who might adopt the received microstructuralist view in philosophy of chemistry. But I think this refinement is an important one, because it tends to get glossed over in statements like: “in most cases membership of [chemical] kinds is conferred by microstructural properties” (Hendry 2006, p. 865). What the case of isotopes makes clear is that—even with respect to the stalwart chemical kinds known as elements—membership of chemical kinds is *only partially* conferred by *some and not other* microstructural properties.

Another way of putting the point is that, without elaboration, the microstructuralist claim to confer chemical kind membership can seem to imply that it is all and only microstructural properties that confer kind membership—or even that microstructure somehow straightforwardly dictates kindhood to scientists. In contrast, when I say that membership in chemical kinds is neither strictly nor solely conferred by microstructure, I am saying that membership of chemical kinds is not conferred by all microstructural properties (absolutely strictly) nor is it conferred only by microstructural properties (solely). Membership in chemical kinds—including elements—is conferred by

some and not other microstructural properties, as well as certain macroscopic features, on the basis of scientific agreement about which properties and features are the appropriate ones.

Sub-Section 2.4: Hendry on Compounds

Now I want to turn to another aspect of Hendry's characterization of chemistry, which is that membership of chemical kinds is conferred by microstructural properties *in most cases*. As I mentioned earlier in this section Hendry divides chemical substances into three types: elements, compounds, and mixtures. And Hendry spends section 2 of his paper explaining how element-membership is (by scientific agreement) conferred by microstructural property (of nuclear charge).

Section 3, entitled "Compounds," is devoted to water. Here, Hendry draws a distinction between two kinds of parts—ingredients and components—within his account of molecular composition. *Ingredients* go into but are used up in the creation of a compound substance. *Components* persist in the product. Hendry argues that "the H₂O molecule is characteristic of water" (2006, p. 873), and the distinction between ingredients and components permits him to deal with complications resulting from molecular disassociation, oligomer formation, and theoretically possible decay chains.

Interestingly, Hendry's response to these complications appeals to Lavoisier's original *intention* with respect to the definition of molecular composition. Hendry argues that because Lavoisier's definition stressed ingredients rather than components, it is possible to maintain that the H₂O molecule is characteristic of water—since although H₂O is not the only component within bodies of water (or even necessarily the majority

of components), it is the primary ingredient.

So this is another case where, if we separate the question of what confers membership conditions for the kind into the logical and the historical, we are going to get two very different answers, only the first of which is purely microstructuralist. Here the logical question is: what confers membership in water kinds? And the answer is: having the molecule H_2O as an ingredient.²⁰ But the historical question is: what determines what confers membership in water kinds? And here the answer, even for Hendry, appeals to decidedly non-microstructural factors: in this case, to the supposed mental state of a long-dead, very famous chemist.

Regardless of the result in that particular case, it is really a stretch to argue from the one case of water to all compounds in general. The Chemical Abstracts Registry (CAS), a division of the American Chemical Society (ACS), keeps track of all known chemical compounds. There are currently more than 70 million registered organic and inorganic compounds and almost 65 million registered protein and nucleic acid sequences in CAS.²¹

Water is not somehow representative of all compounds—certainly not all 70,000,000 plus. Water is not even particularly representative of some smaller class of compounds. If anything water is a rather unique compound, as far as chemists can tell.²² But Hendry realizes that he cannot simply infer from one case to tens of millions. He says:

²⁰ Along with meeting a general necessary (though insufficient) sameness of elemental composition condition for sameness of substance (Hendry 2006, pp. 872–3).

²¹ As of August 12, 2013, when CAS RN 1448220–62-9 was the most recent CAS registry number.

²² Water has many unusual properties. For example: water molecules are polar and can form an unusually high number of intermolecular bonds; water exists quite stably on Earth in its solid, liquid, and gaseous states; in its liquid form it is the universal solvent; it has high surface tension; etc.

Other compound substances present fewer problems to the microstructuralist. Some are simply more uniform as collections of molecules, while others are typically encountered as components of mixtures, rather than as (relatively) pure macroscopic samples. In either case there is less opportunity for the extensions of the substance terms to come apart from the extensions of the names of their characteristic molecular species. (Hendry 2006, p. 873)

I have no doubt that some other compound substances present fewer problems than water does to the microstructuralist. I am equally certain that still further compound substances present more problems than water does to the microstructuralist. There are literally millions of possibilities, after all. And Hendry provides no reason (beyond the above quote) for thinking that most compounds are going to be easier rather than harder for the microstructuralist to handle than water. In the absence of any argument that generalizes from water to “most cases” it is probably better to remain agnostic about how microstructuralism fares with respect to compounds in general.

Quite reasonably, Hendry moderates his conclusion about compounds and microstructure in the fourth and final section of “Elements, Compounds, and Other Chemical Kinds.” He writes that although “the microstructuralist case for the elements seems strong,” compounds “are more problematic and complex, so that microstructuralism about compounds will need to be assessed case by case, and water is only one” (Hendry 2006, p. 873). I agree with Hendry on this point, that water is only one compound, and that other compounds should be investigated on a case-by-case basis with respect to microstructure and classification. Though of course with more than 70 million cases that is going to be quite a task.

One thing that could be done is to investigate groups of similar compounds together, rather than individually. That might help make a systematic review of chemical

compounds and their classification more tractable. And chemists are certainly fond of grouping compounds, so philosophers have ample groups to choose from. One might study hydrocarbons, or cyclic compounds, or coordination compounds, or polycrystals... The list goes on and on.

It is also worth noting here that Hendry's initial division of chemical substances was into elements, compounds, and mixtures, but so far we've only discussed elements and compounds. So there is another whole category of chemical substance still to be analyzed with respect to microstructuralism. Alloys are one fascinating kind of mixture, for example. Perhaps a philosopher might assess that case.

Hendry himself never quite gets to mixtures, though he does briefly discuss chemical kinds at other levels than substances, such as: atoms, ions, molecules, metals, and acids. Hendry admits that acids are an example of a chemical kind not united by any microstructural feature(s),²³ but Hendry claims that "while this discontinuity is an interesting phenomenon and shows that not all kind terms can be understood in the same way, it does not reflect on other chemical kinds" (2006, p. 874).

This is bad news for Hendry's more general characterization of chemistry as a discipline whose kinds are governed by microstructure—and it's not just that acids are a counter-example to Hendry's characterization. It's the claim that acids don't reflect on other chemical kinds that really does the damage. Hendry's microstructuralist characterization of chemistry is built on the generalization of one case of chemical kind (elements), plus one instance (water) of another chemical kind. And if acids don't generalize, why should elements? Hendry doesn't say.

²³ As shown in Stanford & Kitcher (2000); also discussed in Hacking (1983).

Although this might be bad news for Hendry's argument I think it is good news for the philosophy of chemistry. It means that there is lots of work to be done. If we can't generalize from one or two cases to a comprehensive characterization of chemistry, then we've got to conduct a more thorough review, and really pursue Hendry's case-by-case recommendation. I'll begin to do that in the next section, by looking at how microstructuralism fares with respect to the classification of proteins. Perhaps unsurprisingly, it doesn't fare too well.

Section 3: A Different Kind of Chemical Kind

Proteins have gotten just a fraction—approximately one-tenth—of the scholarly attention that genes have from philosophers.²⁴ There are undoubtedly many reasons for philosophers to focus on genes, such as: the gene is the standard unit of heredity in evolutionary theory²⁵; the concept of genes as information is an important part of the “central dogma” of molecular biology²⁶; genes have seemed to fit with philosophically popular reductive and determinist accounts of science²⁷; and the concept of ‘gene’ turns out to be notoriously hard to define²⁸.

But it is time for proteins to become the subject of substantial philosophical study as well, and as more than just the products of gene expression. There are many reasons why philosophers ought to take a closer look at proteins, but two should suffice for the

²⁴ Searching the Philosopher's Index for the terms ‘gene’ and ‘protein’ in all fields except full text produces 1104 entries, while searching for ‘protein’ instead of ‘gene’ produces only 100 results. Searching the title of entries produces 390 results for ‘gene’ and only 20 results for ‘protein’. (Data obtained on February 28, 2013.)

²⁵ An idea developed by Thomas Hunt Morgan and his student, Herman J. Muller, in the early to mid twentieth century.

²⁶ See Crick 1958.

²⁷ For a presentation and critique of these ideas, see Mazzocchi 2008.

²⁸ See, for example: Griffiths 2002; Portin 2002; Perini 2011.

purposes of this discussion of chemical kinds. First, proteins are a particularly large group of compounds. If we're going to try and understand how chemical kinds fare with respect to the microstructuralist thesis, on a case-by-case basis, then proteins are an especially important case, simply in virtue of sheer numbers.

This huge group is not even included in the 70,000,000 plus compounds standardly registered in CAS. This is because CAS separates protein and nucleic acid sequence registration from other compound registration. And as I already mentioned, there are almost 65,000,000 protein and nucleic acid sequences registered in CAS. Of course, this is an immense and still open set of molecules: new sequences are registered with CAS every day.²⁹

CAS' sequence database also contains redundancies, so the number of registered sequences does not necessarily accurately reflect the actual number of currently known proteins. Rather, we should consider the number of registered sequences in CAS as something of an upper bound for estimates of the number of currently known proteins. We're going to have to settle for a bounded estimate because it is actually very difficult to create a non-redundant database of known proteins.

There are various databases that do try to be non-redundant, like PDB (which stands for Protein Data Bank), and SwissProt. PDB contains only solved protein structures, so it affords a lower bound for estimates of currently known proteins. The number of proteins in PDB is currently 93,252.³⁰ SwissProt is a database of protein sequences rather than structures, but it manually annotates its entries in order to avoid

²⁹ You can watch this happen in real time at www.cas.org.

³⁰ As of August 23, 2013.

redundancy, so it too provides a low estimate of the number of currently known proteins. Databases often issue “releases” that tabulate their overall numbers, and the 2013_08 SwissProt release contained 540,732 sequence entries.³¹

Other databases, like PIR (Protein Information Resource) are much less meticulous. PIR recently issued a release that put the total number of its entries at 45,846,921.³² And the Tenth Anniversary RefSeq release included 32,504,738 proteins.³³ So, the number of currently known protein types is somewhere between 30 and 50 million, according to non-meticulously non-redundant sequence identification. Obviously, it is difficult to say precisely how many proteins have been identified as yet.

It is also difficult to say how many types of protein exist in total, among human and other organisms. Some estimate 100 million, others more than 100 billion.³⁴ Because of the way proteins are generated (by chaining amino acids), there are more potential proteins than there are atoms in the universe.³⁵ Regardless, as it currently stands proteins and nucleic acid sequences comprise nearly half of all formally registered chemical compounds. A systematic account of chemical kinds must include the immense group of proteins.

But that was just the first point. The second point is that proteins are incredibly complex macromolecules, and this complexity is a barrier to unified microstructuralist

³¹ Dated July 24, 2013. Swiss-Prot is a division of UniProt (Universal Protein Resource).

³² Also dated July 24, 2013.

³³ RefSeq is a National Center for Biotechnology Information (NCBI) database. The Tenth Anniversary release is dated July 25, 2013.

³⁴ For example: there are approximately 20,000 different genes in the human genome (Pennisi 2012), yet estimates of the number of different human proteins range from one hundred thousand to one million (Hinz 2010). Obviously, this is far more than expected according to Beadle and Tatum’s (1941) original one gene – one enzyme hypothesis (which is often glossed as one gene – one protein).

³⁵ Anecdotal, it is also quite difficult to estimate the number of token proteins in any given individual organism. For example: half of the dry weight of an *E. coli* cell is protein, whereas a fifth is RNA, and only about 3% is DNA (Voet & Voet 2004).

classification. It's not that proteins don't have microstructure, or that microstructure isn't important to the individuation of proteins. They do, and it certainly is. It's just that protein structure is incredibly complex, and includes both micro- and macro-structural components. Microstructure in particular does not simply or straightforwardly dictate the individuation of proteins. To reapply some terminology from the earlier discussion of elements: membership in protein kinds is neither strictly nor solely conferred by microstructural properties. I'll explain why in what remains of this section—but to do so I'll first have to explain some of the basics of protein science.

Sub-Section 3.1: Basic Protein Science

Proteins are macromolecules: extremely large molecules that might be composed, but are not simply aggregates, of many smaller molecules. Among the various kinds of macromolecules, proteins are polymers: long chain molecules, created by sequential bonding of many repeating units of smaller molecules, or monomers. In the case of proteins, the monomers in question are amino acids, and they are joined together via peptide bonds. There are 20 standard amino acids.

Proteins are built out of amino acids by joining them together in chains, each amino acid bonded to its neighbor on either side, and then, usually, looping and folding the chain into a globular structure. Between two and about 50 amino acids bonded together is called a 'peptide'; more than 50 and the macromolecule is called a 'polypeptide'. So, proteins are polypeptides: long and often clumped together chains of amino acids each linked to its two neighbors via covalent peptide bonds.

The clumping is important. Despite their basic chaining, the native state or

natural formation of most proteins is not simply that of a straight line.³⁶ Although some fibrous proteins approach this kind of one-dimensionality, many proteins are globular and thus have a complicated superstructure that cannot simply be captured by the linear sequence of amino acids of which it is composed. The overall shape of a protein is called its ‘conformation’.

Overall protein shape or conformation is tracked by dividing protein structure into a hierarchy of four levels: primary structure (amino acid sequence, producing chains); secondary structure (local bond formation between amino acids, producing shapes like turns and sheets); tertiary structure (additional interactions between elements scattered throughout the molecule, completing overall conformation); and quaternary structure (multiple-molecule shape formation, producing complexes).³⁷

The primary structure of proteins is what is encoded in genes. Very quickly: genes (sequential strands of nucleic acids) are transcribed, by RNA polymerase, into messenger RNA (also made of nucleic acids), which is then translated, by ribosomes, into a sequence of amino acids (which are not nucleic acids). This is only the primary structure of the protein, however. The strand then contorts, often with assistance and only in the right context, according to the sorts of interactions mentioned above. This part of the production of proteins is called ‘protein folding’³⁸ and it finally results in a

³⁶ Although the folding of a protein into its native state is a process that often requires a particular context and even assistance from other molecules. See Slater (2009) for more detail.

³⁷ For more information please see, for example, *Lehninger's Principles of Biochemistry*, most recently in the 5th edition (Nelson & Cox 2008). I will also elaborate on the details on protein science in later chapters of the dissertation.

³⁸ Sandra D. Mitchell is currently doing philosophical work on the topic of protein folding. I have seen her give numerous talks on this subject, though I am not currently aware of any published work by her on this matter as yet.

completed protein in its overall shape or conformation.³⁹ A protein in this form is said to be in its ‘native state’. A protein that is no longer in this state—after being exposed to some agent of destruction of its superstructure, like high heat—is said to be ‘denatured’. Denatured proteins are nonfunctional.

By now it should be clear that amino acid sequence, or primary structure, is not equivalent to overall protein conformation, or superstructure.⁴⁰ References to a protein’s “structure” are often ambiguous with respect to which of these things is being referred to. But we need to distinguish these:

PRIMARY STRUCTURE – a sequence of amino acids chained together via peptide bond into the polypeptide base of a protein; a microstructural property of proteins.

SUPERSTRUCTURE – the overall shape or conformation of a protein, from primary structure to tertiary or quaternary (if applicable) structure; a macrostructural property of proteins.

And this is not merely a semantic point. In the next sub-section I’ll explain how different kinds of relationships between the primary structure and superstructure of a protein (as well as its function) are tracked in the individuation of proteins.

Sub-Section 3.2: Individuating Proteins

In the next two sub-sections I will discuss the individuation of proteins, showing first what doesn’t (in this sub-section), and then what does (in the next sub-section), confer membership in protein kinds. By the end of sub-section 3.3 I’ll have shown that

³⁹ The overall shape or conformation of proteins is somewhat flexible, however—more so in some cases rather than others. More on this topic in sub-section 3.3.

⁴⁰ This is a better term than either tertiary or quaternary structure because it can encompass both. In other words, reference to a protein’s superstructure picks out its overall confirmation in either case—up to tertiary structure, if it’s just a single-stranded protein, or including quaternary structure, if it has that because it’s a complex of multiple strands.

my claim (i) applies to membership in protein kinds just as it did to element kinds. I'll also finally make my argument for claim (ii) explicit. Again, however, some scientific information is required in order to make and understand the philosophical arguments.

By now it is common knowledge that genes vary in sequence; perhaps it is less widely known that proteins do too. In fact, protein sequence (or primary structure) often varies precisely because of variation in the sequence of the gene that codes for it. There are many different ways in which tokens of the same gene type can vary: by substitution, deletion, duplication, insertion, indel (deletion plus insertion), inversion, conversion, and translocation (just to name a few). But in general, genes that produce slightly different tokens of the same protein type are called variants. Again, there are many different ways in which these tokens of the same protein type can vary: by substitution, deletion, duplication, insertion, indel, and extension (again, among others).

Substitutions, for example, replace one amino acid with another in the primary structure of a protein. This does not—at least according to overwhelmingly consistent scientific practice—make the token protein of a different type. To be really clear: what biochemists refer to as ‘a’ protein, a token protein of one type, does not necessarily have the same primary structure or amino acid sequence as another token protein of the same type.⁴¹ In other words, membership in proteins kinds is neither absolutely strictly conferred by microstructure nor selectively strictly conferred by the microstructural property of having a particular primary structure or amino acid sequence.

Within protein types, token proteins with different primary structure but similar

⁴¹ One interesting implication of this fact is that many protein kinds are thus in violation of Hendry's necessary (though insufficient) sameness of elemental composition condition for sameness of substance (2006, pp. 872–3).

function are called ‘alleloforms’. Token proteins with both different primary structure and different function are called ‘isoforms’. To relate this discussion of proteins to the earlier discussion of elements, the alleloforms and isoforms of a protein are something like the isotopes of an element. These are token molecules of the same type that differ microstructurally—in the case of proteins, along the microstructural property of amino acid sequence. Sometimes this variation affects their chemical behavior, but even these isoforms are still classified as different tokens of the same protein type.

So, proteins are kind of like elements—they have “isotopes”. However, proteins are also like many other, much simpler and smaller compounds—they can have “isomers” as well. Isomers are compounds with the same (molecular) composition but different (chemical) structure. For example, pentane (C_5H_{12}) has three isomers (normal pentane, isopentane, and neopentane). In each case the five carbon and twelve hydrogen atoms are differently arranged with respect to one another (normal pentane is an unbranched alkane; isopentane is a branched-chain alkane; and neopentane is a double-branched-chain alkane).

Analogously, token proteins with the same molecular composition and primary structure might have different superstructure. This can happen because, for example, the same amino acid sequence might be shaped into various overall conformations during the process of protein folding. Changes in conformation might depend on the presence or absence of certain compounds, such as chaperonins, or on various environmental conditions during folding. Just as token compounds with the same molecular composition might be arranged into different chemical structures (forming isomers), token proteins with the same primary structure might be arranged into various

superstructures (or, have different chemical superstructure).

Again, this does not—according to overwhelmingly consistent scientific practice—make these token proteins of a different type. This is in contrast with the case of isomers, which are often treated as different compounds.⁴² In other words, token molecules with the same molecular composition are often sorted into different molecular type if they have different chemical structure. This is how it works for isomers and basic compounds; but it is not how it works for proteins and complex macromolecules. Token proteins with the same primary structure but different overall conformation or superstructure are not generally considered different protein types.

Somewhat unhelpfully, token proteins that differ in this way are also referred to as ‘isoforms’. So the term ‘isoform’ is ambiguous: it applies to token proteins with different primary structures and different functions, as well as to token proteins with the same primary structures but different superstructures. However, these superstructural variations are often inferred via functional differences, as it is very hard to directly track superstructural variation of proteins. That means that the term has a third sense: it also refers to token proteins with the same primary structures but different functions. Basically, the term ‘isoform’ refers to any variant of a protein that differs functionally, whether because of different primary structure, different superstructure, or both.

With regard to these functional differences: slight changes in either the primary structure or superstructure of a protein can lead to varying degrees of difference in its function. Slater (2009) discusses a nice example of slight changes in structure producing

⁴² Different isomers of a compound often behave very differently from one another—performing in different chemical reactions, etc.

slight changes in function. The enzyme alcohol dehydrogenase (ADH) is a protein that, in *Drosophila*, undergoes slight functional modification corresponding with slight structural change. There are “slow” and “fast” version of the protein that are produced by slightly variant gene sequences, which code for different versions of the enzyme.

In other cases a single amino acid swap can drastically alter protein behavior. For example, in humans the gene ABCA1 normally codes for Cholesterol efflux regulatory protein. Variants of ABCA1—altering only one amino acid in the protein sequence—can produce proteins that are severely altered functionally. Some variants result in the absence of HDL cholesterol. Others result in moderately low LDL cholesterol. Many of these variants can lead to (different versions of) high density lipoprotein deficiency disorder. Some have been detected in colorectal cancer samples.⁴³ Again, all of these tokens count as variants of the same protein type.

Amidst all of the scientific detail, there are three very important, and importantly distinct, philosophical points being made here. In what remains of this subsection, I will reiterate these three points while also connecting them to one another in significant ways.

Point one: proteins have isoforms and alleloforms, like elements have isotopes.⁴⁴ This means that proteins can vary in primary structure with and without varying in superstructure (or function). In the case of isoforms, primary structure varies along with variation in superstructure (or function)—there is a one-to-one relationship between the primary structure and the superstructure (or function) of these proteins. In the case of alleloforms, primary structure varies without variation in superstructure (or function)—

⁴³ See the UniProt database of protein variants (<http://www.uniprot.org/docs/humsavar>) for more detail on this and many other examples.

⁴⁴ One could even think of alleloforms as behaviorally similar isotopes (like the stable carbon isotopes), whereas isoforms are more like behaviorally distinct isotopes (e.g., protium, deuterium, and tritium).

there is a many-to-one relationship between the primary structure and the superstructure (or function) of these proteins.

Point two: proteins also have isoforms of another kind, like basic compounds have isomers. This means that proteins can vary in superstructure (or function) without varying in primary structure. In the case of this kind of isoform, primary structure doesn't vary while superstructure (or function) does vary—there is a one-to-many relationship between the primary structure and the superstructure (or function) of these proteins.

Putting these first two points together: there is a many-to-many relationship between the primary structure and superstructure (or function) of proteins in general.

There's one final point to make, and then to incorporate with the others. It is now time to disassociate a protein's superstructure—its overall shape or conformation—from its function. It is often very difficult to determine a protein's superstructure. Doing so is called 'solving the structure' of a protein. As I mentioned earlier, there are currently just 93,252 solved protein structures recorded in PDB, in contrast with the tens of millions of protein sequences stored in other databases.

However, there are at least a few known cases of proteins with identical superstructure performing different functions in different contexts. Crystallins are a popular example of this (see Tobin 2010). In these cases, there is a one-to-many relationship between protein superstructure and function. There are also lots of cases of proteins with different superstructures performing similar functions. One example: there are many structurally dissimilar proteins that perform the function of DNA-binding—even binding differently to the same stretches of DNA. In these cases, there is a many-to-

one relationship between protein superstructure and function. Again, this means that there is a many-to-many relationship between protein superstructure and function.

Putting all three of these points together: there is a many-to-many-to-many relationship between the primary structure, superstructure, and function of proteins. Unsurprisingly, this has consequences for the viability of the microstructuralist thesis with respect to proteins. I will address this issue directly in the next sub-section.

Sub-Section 3.3: Proteins, Microstructuralism, and Unification

With respect to any kind of kind, to be microstructuralist about that kind is to say that membership in said kind is conferred by a microstructural property or properties. So, to be a microstructuralist about proteins is to say that membership in protein kinds is conferred by a microstructural property or properties. Of course, the candidate microstructural property for conferring membership in protein-kinds is primary structure.⁴⁵

I have already shown that membership in proteins kinds is neither absolutely nor selectively strictly conferred by the microstructural property of having a particular primary structure or amino acid sequence. But as discussion in the previous sub-section shows, neither is it strictly (in either sense) conferred by the macrostructural property of having a particular overall conformation or superstructure. Finally, membership in protein kinds isn't strictly conferred by the macroscopic feature of function either.

I propose that membership in protein kinds is mostly—but not solely—conferred by the property of origination, in the right context, from a particular gene. This is

⁴⁵ Certainly there are other candidates. However, this is by far the most plausible one.

evidenced by the undeniable fact that scientists call all of the token proteins encoded from a particular species-specific gene of the same protein type.

With respect to this aspect of scientific practice: within humans, token proteins of one protein type, which differ in primary structure or amino acid sequence, are called ‘variants’. Variants that function normally⁴⁶ are simply called ‘polymorphisms’ (most of these are probably alleloforms, although it’s possible that future discoveries will turn some of them isoforms instead). Variants whose function differs from the standard somehow are labeled ‘disease variants’ (these are pathological isoforms).

The UniProt database on just single amino acid variants in humans contains 67,878 total variants.⁴⁷ 37,782 of these variants are classified as polymorphisms. 23,846 of these are classified as disease variants. The remaining 6,250 are unclassified. Each token variant is grouped according to its gene (and corresponding protein) type. For example (having picked one at random), gene CYP4F22 codes for the protein Cytochrome P450 4F22. There are 7 known variants: 2 are classified as polymorphisms, and 5 are classified as disease variants (they cause a form of ichthyosis).⁴⁸ All of these are tokens of the type Cytochrome P450 4F22.

So, scientists keep track of important differences within protein types—among primary structure, superstructure, and function—but these are all presented as differences within one kind of protein. Just as elements have isotopes and compounds have isomers, so too do proteins have alleloforms and isoforms.

But I say that membership is only *mostly* conferred by genetic origination

⁴⁶ This doesn’t mean anything more than “as scientists expect this protein to function given previously obtained results.”

⁴⁷ As of July 24, 2013.

⁴⁸ Visit the database at <http://www.uniprot.org/docs/humsavar> for more.

because other considerations also matter. For example, a protein's native state also includes its superstructure—a protein is denatured without it—so membership in protein kinds is also partly conferred by its macrostructure. Just as an atom can decay, undergoing nuclear transmutation and resulting in an atom of a different element, so too can proteins lose their superstructure and thus their identity as a token of a particular type of protein.

And I say that membership is mostly conferred by genetic origination rather than the primary structure that gene encodes because protein identity incorporates strict-species-specificity in addition to sequence-somewhat-specificity. It's 'sequence-somewhat-specificity' because, remember, since the nucleic acid sequence of a particular gene can vary within a somewhat limited range, so can the amino acid sequence or primary structure of the corresponding protein. And it's 'strict-species-specificity' because regardless of amino acid sequence identity, token proteins from different species are not considered to be proteins of the same type.

Within the human genome, for example, there are certain 'ultraconserved elements'.⁴⁹ Within these elements there are genes that code for certain types of protein—like Polypyrimidine Tract Binding Protein 2 (PTBP2)—for which there are orthologous proteins in other species—in the case of PTBP2, in *Callithrix jacchus* (the common marmoset). Tokens of these two different types of protein—human PTBP2 and marmoset PTBP2—have identical amino acid sequence or primary structure. But they are considered orthologous proteins, rather than tokens of the same protein type.

Another way to put this point is that it is important to distinguish genetic

⁴⁹ Originally presented in Bejerano et al. (2004).

origination from the particular microstructural property of amino acid sequence or primary structure—that any given protein type has because of its genetic origination—in order to accommodate irrelevant difference while discounting irrelevant similarity. Irrelevant difference occurs among different amino acid sequences, or primary structures, within tokens—alleloforms—of one protein type. Irrelevant similarity occurs among identical amino acid sequences, or primary structures, among different protein types—orthologs—of different species.

The details of my proposed account of how membership in protein kinds is conferred—by an assortment of properties including genetic origination, macroscopic superstructure, somewhat-specific-sequential microstructure, and strict-species-specificity—will require elaboration elsewhere. Here the focus is on claims (i) and (ii). My proposal is related to these claims because it presents an alternative to microstructuralism, or any view of chemical kinds that supposes membership in these kinds to be conferred by only one kind of property—microstructural, macrostructural, functional, what have you. But my negative arguments with respect to what doesn't work—i.e., microstructuralism—should go through regardless of my positive proposal for what does work.

Recall my claim (i): membership in chemical kinds is neither strictly nor solely conferred by microstructure. In the previous section I focused on elements, arguing that that membership in element kinds is neither strictly nor solely conferred by microstructural properties. In this section I have focused on proteins, showing that membership in protein kinds is neither strictly nor solely conferred by microstructural properties. Here I have shown that neither primary structure (a microstructural property)

nor superstructure (a macrostructural property) strictly or solely confers membership in protein kinds. The fact that the microstructuralist thesis fails in both of these extremely significant cases should raise significant doubts about the standard presumption that chemical kinds are generally microstructuralist. It's also worth remembering that function doesn't do the trick either.

One last point: I have also said a lot in this section that pertains to my claim (ii), which was that chemistry does not have a special status (relative to, say, biology) as an exemplary science with unified interests that govern its classification. This is a good place to finally make the argument explicit. There is an incredible amount and diversity of chemical kinds. Elements are one kind of chemical kind with diverse interests—microstructural and macroscopic—governing their individuation. Basic compounds are another case with different interests—chemical structure over molecular composition—governing their individuation, or what little of it philosophers have examined. Proteins, a kind of macromolecule, are yet another example of a case with still different classificatory interests—microstructure and macrostructure and function and origination—all playing a role in their individuation.

Hendry implies that what unifies classificatory interests in chemistry is microstructure. But not even elements are strictly or solely microstructuralist kinds. And without surveying all the more than 70 million currently known compounds, the 30 to 50 million currently known proteins, and the countless mixtures, it's difficult to say what all the various interests that might play a role in chemical classification are. But it's safe to say that these interests are diverse. We already know from what we've surveyed here that these diverse classificatory interests span microstructure, macrostructure, chemical

behavior, biological function, and genetic origin, for example.⁵⁰ Chemistry's special status is undeserved.

Sub-Section 3.4: Slater, Tobin, and Goodwin

As I mentioned earlier, other philosophers of science have recently turned their attention to proteins. In fact, three of them conducted a somewhat direct, if extended, conversation with one another throughout the pages of *Philosophy of Science* (Slater 2009), *Foundations of Chemistry* (Tobin 2010), and *Biology & Philosophy* (Goodwin 2011).

In what remains of this section I'll discuss this recent work by Matthew H. Slater, Emma Tobin, and William Goodwin. In each case I'll first describe the position the author takes with respect to protein kinds and microstructuralism, and then show how their account fails to comprehensively and/or consistently account for important aspects of protein science. Putting the cart rather before the horse, I'll argue that Slater (2009) overemphasizes functional concerns while failing to consider the possibility of joint membership conditions in protein kinds; Tobin (2010) misunderstands the nature of the indeterminacy of the proteins at the heart of her argument; and Goodwin (2011) doesn't take isoforms and alleloforms into account.

But taking each in turn: Slater (2009) uses the case of proteins to argue that philosophers ought to be "pluralists about macromolecular classification" (p. 851).

Throughout the article the author focuses more on the issue of monism versus pluralism

⁵⁰ It might also be worth explicitly stating here that this discussion should make clear the folly of the standard generalization that chemistry is about structure whereas biology is about function. Chemistry takes into account both (micro)structure and (chemical) function (in reactions, say). Biology pays attention to (underlying) structure and (biological) function (in evolutionary explanations, say).

rather than microstructuralism with respect to the individuation of proteins. Regardless, the discussion is highly relevant—and Slater’s position is an interesting one. He argues that proteins are individuated by function rather than structure, and that the functional individuation of proteins supports pluralism about protein kinds rather than monism.

I’m going to dispute the former claim, that proteins are individuated by function, thus blocking the latter claim, that functional individuation supports pluralism. Slater’s claim that proteins are individuated by function rather than structure has two parts: first, that individuating proteins by primary or secondary structure would “count too few” proteins, and second, that tertiary structure is too dynamic for the purposes of individuation. I disagree with both. In the next few paragraphs I’ll show that in the former case, Slater indirectly contradicts himself; and in the latter case, the point is untenable when broadly considered.

With respect to the first part of Slater’s claim, that primary or secondary structure is insufficient, he writes:

Thus, just as individuating chemical kinds by their ingredients significantly undercounts the variation introduced by different structures (hence ‘isomers’, from ‘same parts’), counting proteins by primary or secondary structure is thus apt to count too few. (Slater 2009, p. 853)

This remark takes into account the second kind of isoform discussed in sub-section 3.2: those functionally distinct tokens of the same protein type that have identical primary structure. According to Slater, this should make these tokens of a different protein type. He doesn’t say why the current scientific approach—of tracking this sort of variation via isoforms and disease variants while still counting these as tokens of the same protein type—is insufficient.

The remark also fails to take into account the existence of alleloforms. This is strange, as Slater discusses alleloforms later in the article. Regardless, his position here is that because there are isoforms (proteins with similar primary structure and dissimilar function), and because protein individuation should demarcate functionally distinct proteins, individuating proteins on the basis of primary structure counts too few. But because of the existence of alleloforms (proteins with dissimilar primary structure and similar function), if protein individuation should demarcate functionally distinct proteins, then individuating proteins on the basis of primary structure also counts too many. It's inaccurate just to say that structural individuation simply counts too few.

We should also resist the second part of Slater's claim—that dynamism is an obstacle to individuation. He writes that “proteins are structures in motion” (2009, p. 854).⁵¹ This I agree with; however, I don't see why that should be considered an obstacle to individuation on the basis of what I have been calling superstructure.⁵²

As I've said earlier in the chapter, proteins must be folded, and the same primary structure can be folded into different superstructures. Proteins with the same token sequence can have different overall shape. But the same token protein can also change its shape. Proteins can bend, twist, flatten, swell, and more. They can also pick things up, change their shape, and drop them. For instance, many proteins are enzymes—they catalyze biological reactions. Enzymatic proteins interact with other compounds in the body. They “absorb” them into particularly shaped “pockets” and change their shape before releasing them. Sometimes this also changes the overall shape or stability of the

⁵¹ Slater credits this phrase to Moira Howes.

⁵² Of course, I don't think that proteins can be individuated purely on the basis of superstructure. But it's for a different reason than Slater's (I explained my reasons in sub-sections 3.2 and 3.3).

protein itself. So interacting with such proteins alters other compounds in the body, and sometimes the proteins emerge from these interactions also affected.⁵³

Another way of putting this point is that proteins tend to have a flexible rather than a fixed superstructure. So, Slater and I agree on this fact, that proteins are dynamic structures. According to Slater, this makes superstructure a poor candidate property for the basis of individuation. However: proteins are not unique in being structures in motion; many crucial chemical kinds are. Atoms are structures in motion. Bonds are structures in motion. And both of these chemical entities can be individuated on the basis of dynamic features. We can't pinpoint the position of electrons within atoms, but we can still individuate them—and we have developed the concept of orbital shells in order to do so. Bonds also have extremely dynamic structure—and similarly, we have developed ways of individuating them throughout their various ranges and shapes of motion.

So it's a mistake to reason from superstructural dynamicity to inability to individuate proteins on the basis of this dynamic property. It's also a mistake for Slater to insist that proteins be individuated on the basis of function because neither microstructure nor superstructure suffices. As I've shown throughout section 3, this conclusion is utterly incommensurate with scientific practice. More theoretically, Slater reaches it by treating and then dismissing each candidate property *da solo*. He presents three options: microstructure, superstructure, and function. Then he argues that microstructure alone doesn't suffice. Neither does superstructure. So, we're left with

⁵³ The role of enzymes in physiology is an important one and has long been the subject of biological study. The early history of proteins is in large part a history of enzymes. This history is explored in Appendix A of the dissertation.

function. But kind membership is not necessarily conferred by just one property. Just because microstructure can't do it alone, and superstructure can't do it either, doesn't mean that function has to do it by itself.

It's only because microstructuralism stresses that membership in microstructuralist kinds is conferred by microstructural properties and not others that the failure of microstructuralist properties to solely confer kind membership is debilitating to the microstructuralist thesis. But if we're not being microstructuralist about protein kinds, the fact that microstructure can't solely confer kind membership doesn't mean it isn't contributing at all to kind membership. Microstructure can still contribute to protein kinds, as can other properties like superstructure and function. None of them have to do it all on their own, in fact. And it shouldn't be surprising that with complex objects like proteins, more than one property is required for their individuation.

Tobin (2010) uses one particular way in which proteins are complex in order to issue a direct challenge to anyone hoping to be microstructuralist about proteins. And though I agree that microstructuralism fails with respect to proteins, again it's for different reasons. As I see it, Tobin's argument rests on a misunderstanding of how structurally similar proteins perform different functions.

Tobin (2010) discusses the case of "moonlighting proteins"—proteins that do different things in different contexts. Again, these are what I've called isoforms, of various kinds. Some token proteins of the same protein type perform different functions despite having the same primary structure or superstructure; other token proteins of the same protein type vary in primary structure or superstructure and correspondingly vary in function. But Tobin also claims that some moonlighting proteins are "intrinsically

structurally disordered”—what Goodwin (2011) calls ‘intrinsically unstructured proteins’ (IUPs).

Goodwin and I agree that Tobin’s (2010) paper mistakes the variation that exists between the primary structure and/or superstructure of token proteins of the same protein type for structural indeterminacy in token proteins. In other words, I think that Tobin thinks that some token moonlighting proteins have indeterminate primary structure, when actually it’s that certain types of moonlighting proteins have various, determinate token superstructures. As Goodwin (2011) puts it, it seems “as if Tobin thinks that IUPs have no fixed primary structure, and this is why they undermine microstructuralism. If this is the argument, then it is based on a misunderstanding of IUPs” (p. 543, footnote #12).

This is a place where the distinction between tokens and types really comes in handy. If Tobin simply means that certain protein types do not have fixed primary structure, well then she’s right—but it’s true of more than just moonlighting proteins. It applies to all proteins with alleloforms, for example. However, if Tobin means that certain token proteins do not have fixed primary structures, I don’t know of a single case in which that’s true. Of course proteins are “structures in motion,” as was just discussed, and token proteins can certainly change their secondary, tertiary, and quaternary superstructure. But I don’t know of any proteins that preserve their token identity through changes in primary structure.⁵⁴

Then again, if such proteins do exist that’s just another reason to dismiss microstructuralism with respect to proteins. And I think that case has already been made.

⁵⁴ There are such things as conjugated proteins, of course, which add certain elements (like a metal) to their configuration at times, and then lose it again (like an enzyme can pick up and drop things). But these additions aren’t part of the primary structure of the proteins, and to say that this is “intrinsically disordered” would be strange.

Goodwin, however, does not. *Contra* Slater and Tobin, Goodwin (2011) argues that “primary structure should be regarded as the fundamental distinguishing characteristic of protein taxonomy” (p. 533). This is because, he concludes, “proteins are individuated, most fundamentally, by primary structure” (p. 544).

I agree with much of what Goodwin has to say in critique of Slater (2009) and Tobin (2010). But I must dispute Goodwin’s conclusion. Proteins are not individuated by primary structure—certainly not in such a way that membership in protein kinds is fundamentally conferred by this particular microstructural property.

Much of Goodwin’s argument consists in an appeal to scientific practice. And it is true that proteins have, traditionally, been identified by their primary structure. But this is similar to the way in which atomic weight was once used to identify the elements. Like elements, protein types have turned out to contain microstructural sub-populations. Samples of any given element may contain different isotopes—atoms with different atomic configurations. Likewise, tokens of any given protein type may exhibit variation—these tokens of a type may differ in terms of their amino acid sequences.

Recall the discussion, throughout this section, of the many-to-many-to-many relation between the primary structure, superstructure, and function of protein types. With respect to the genes that code for proteins, variant sequences can produce alleloforms or isoforms. In humans, alleloforms and harmless isoforms are called polymorphisms. Physiologically disruptive isoforms are called disease variants. But all of these—alleloforms, isoforms, polymorphisms, and disease variants—are considered and called tokens of the same protein type.

This shows that proteins are not individuated by primary sequence—certainly not

by primary sequence alone or “fundamentally.”⁵⁵ As I have already said, I think that proteins are individuated by a combination of ranging functions, conformations, and amino acid sequences. Regardless, it is certainly a mistake to think that a unique primary structure or amino acid sequence picks out or differentiates one protein from another.

It might be worth pointing out there that there is also an open-endedness to the set of sequences that picks out any particular type of protein. As in, given the failure of any one sequence as uniquely or fundamentally picking out a particular protein, one might be tempted to appeal instead to a fixed disjunction of sequences to individuate said protein. But that’s not an option, either—tomorrow another variant could be discovered, and then it would need to be added to the set. The set of individuating sequences is not closed, and must remain open.

To summarize: Goodwin (2011) does not discuss alleloforms, isoforms, variants, or polymorphisms in his piece—he somehow misses this aspect of biochemical practice. He also misinterprets the significance of other laboratory practices. For although biochemists might use primary structure to identify or manipulate particular proteins, they do not think that this is what makes a particular protein the kind of protein it is—just as they might use atomic weight or number of electrons to pick out or manipulate a particular atom, but wouldn’t make the mistake of thinking those are the properties that make the atom the kind of atom it is.

⁵⁵ It is unclear what Goodwin means by the term ‘fundamentally’. But if primary sequence were fundamental to protein individuation, I presume that distinct primary sequences would pick out distinct proteins. They certainly don’t, as a host of distinct primary sequences often pick out the same protein.

Section 4: The Argument from Iso-Ism

At this point the microstructuralist about chemical kinds might decide to stand up and say: “Enough! Who cares about practice?” For it is true that much of what I’ve said here about how both proteins and elements are individuated relies on scientific practice. I’ve argued that membership in element kinds is, according to scientific practice, conferred by scientific agreement with respect to certain microstructural and macroscopic properties. And I’ve shown that membership in protein kinds is, according to scientific practice, not conferred by any microstructural, macrostructural, or functional properties, either strictly or solely. But the microstructuralist about chemical kinds might be tempted to go “purely” microstructuralist, and simply insist that scientific practice doesn’t matter, in the end, for how membership in chemical kinds is conferred.

The pure microstructuralist, for example, might be tempted to say something like: “look, scientific practice has just got it wrong. Even though scientists aren’t strict microstructuralists about elements or proteins, they should be. Elements are individuated selectively strictly and solely by the microstructural property of number of protons; proteins are individuated selectively strictly and solely by the microstructural property of amino acid sequence, or primary structure. Scientists don’t currently recognize this but they should.” Yet this is another place where the distinction between absolutely and selectively strict microstructuralism is critical. Because without an appeal to scientific practice, the pure microstructuralist about chemical kinds has no way of explaining why certain microstructural properties and not others confer membership in chemical kinds. In the case of element kinds, why number of protons instead of number of protons and neutrons, or number of electrons? In the case of simple compounds, why chemical

structure rather than molecular composition? In the case of proteins, why primary structure rather than number of electrons or molecular composition or secondary structure?

As I've shown, what makes certain microstructural properties the ones that confer membership in their respective chemical kinds is more than just the microstructural properties themselves. In other words, because members of chemical kinds are not absolutely microstructurally identical, selection of the relevant aspect of microstructure must occur, and this is part of what dictates chemical kind-hood. Because of the failure of absolutely strict microstructural identity within chemical kinds, it's simply not an option to be solely microstructuralist about selectively strict microstructural properties conferring membership in chemical kinds. Instead, certain microstructural properties and not others confer kind membership, and determining which ones do so requires appealing to other non-microstructuralist factors. I think that an interesting and significant part of scientific practice consists in the determination of what appropriately constitutes such appeals.

But perhaps the staunch microstructuralist would make one more attempt and insist on adopting an absolutely pure microstructuralism. In other words, the microstructuralist might propose that, in order to maintain that membership in chemical kinds is solely conferred by microstructure, membership in chemical kinds ought to be conferred by absolutely strict microstructural identity. However, this is not a viable option, due to what I'm calling 'iso-ism'. Here is the argument from iso-ism: there are isotopes, isomers, isoforms, and relevant equivalents in every kind of chemical kind—no chemical kinds are identical with respect to all aspects of their microstructure. In fact,

because electrons are constantly moving around the atoms that they are a part of, not even an atom at time t_0 and t_1 is identical with respect to all aspects of its microstructure. And since everything is made of atoms, tying kind-hood to absolutely strict microstructural identity would collapse kinds in their entirety. Everything would be its own unique kind in every instant.

So, adopting absolutely pure microstructuralism doesn't do chemical kinds any favors. Going this route with respect to what confers membership in chemical kinds collapses all of chemical kindhood into an infinity of kind sets with exactly one instantaneous member each. Absolutely pure microstructuralism is not an option; neither is a less radical but still "pure" option based on solely selectively strict microstructuralism.

In sum, membership in chemical kinds is neither strictly nor solely conferred by microstructural properties. The selection of some microstructural properties and not others, as well as engagement with other non-microstructural factors, is an ineliminable part of what confers membership in chemical kinds. There's simply no avoiding an engagement with scientific practice.

Section 5: Conclusion

I have argued that membership in chemical kinds is not—in either the case of elements, or proteins—conferred strictly or solely by microstructural properties. In the case of elements, kind membership is conferred selectively strictly by the microstructural property of nuclear charge or number of protons. But it is not solely conferred by this property. The microstructural property of nuclear charge or number of protons that

selectively strictly confers membership in an element kind does so because of macroscopic features of that element kind. If other macroscopic features had been deemed more relevant by the IUPAC chemists of 1923, then a different microstructural property (atomic weight, or number of protons and neutrons) might today selectively strictly, but not solely, confer membership in element kinds.

I have also shown that membership in proteins kinds is neither absolutely nor selectively strictly conferred by the microstructural property of having a particular primary structure or amino acid sequence. But neither is it strictly (in either sense) conferred by the macrostructural property of having a particular overall conformation or superstructure. And yet the macroscopic feature of function doesn't strictly confer membership in protein kinds either.

I think that membership in protein kinds is conferred by an assortment of properties including genetic origination, macroscopic superstructure, somewhat-specific-sequential microstructure, and strict-species-specificity. But more importantly, the fact that the microstructuralist thesis fails in both of these extremely significant cases should raise significant doubts about the standard presumption that chemical kinds are generally microstructuralist.

Finally, I have argued that there is significant diversity with respect to the interests that govern chemical kinds. Elements are one kind of chemical kind with diverse interests—microstructural and macroscopic—governing their individuation. Basic compounds are another case with different interests—chemical structure over molecular composition—governing their individuation, or what little of it philosophers have examined. Proteins are yet another example of a case with still different

classificatory interests—microstructure and macrostructure and function and origination—all playing a role in their individuation.

So, chemistry is not a science with unified interests governing its classification; nor is membership in most of its kinds straightforwardly conferred by microstructure. Those who think either of these things are mistaken, as are any who think that we can simply go microstructuralist “all the way down,” or that chemistry is about structure whereas biology is about function. Things are simply more complicated than that.

REFERENCES

- Barber, R. C., Karol, P. J., Nakahara, H., Vardaci, E., & Vogt, E. W. 2011. Discovery of the elements with atomic numbers greater than or equal to 113 (IUPAC Technical Report). *Pure and Applied Chemistry* 83: 1485–1498.
- Beadle, G. W., & Tatum, E. L. 1941. Genetic control of biochemical reactions in *Neurospora*. *Proceedings of the National Academy of the Sciences* 27: 499–506.
- Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., & Haussler, D. 2004. Ultraconserved elements in the human genome. *Science* 304: 1321–1325.
- Bird, A. & Tobin, E. 2008. Natural Kinds. *The Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/natural_kinds/
- Boyd, R. N. 1991. Realism, anti-foundationalism, and the enthusiasm for natural kinds. *Philosophical Studies* 61: 127–148.
- Boyd, R. N. 1999. Kinds, complexity and multiple realization. *Philosophical Studies* 95: 67–98.
- Chakravartty, A. 2007. *A metaphysics for scientific realism: Knowing the unobservable*. Cambridge: Cambridge University Press.
- Crick, F. 1958. On protein synthesis. *The Symposia of the Society for Experimental Biology* 12: 138–163.
- Ellis, B. 2001. *Scientific essentialism*. Cambridge: Cambridge University Press.
- Goodwin, W. 2011. Structure, function, and protein taxonomy. *Biology & Philosophy* 26: 533–545.
- Griffiths, P. 2002. Lost: One gene concept. Reward to finder. *Biology and Philosophy* 17: 271–283.
- Hacking, I. 1983. *Representing and intervening*. Cambridge: Cambridge University Press.
- Hendry, R. F. 2006. Elements, compounds, and other chemical kinds. *Philosophy of Science* 73: 864–875.
- Hinz, U., & The UniProt Consortium. 2010. From protein sequences to 3D-structures and beyond: The example of the UniProt Knowledgebase. *Cellular and Molecular Life Sciences* 67: 1049–1064.

- Kitcher, P. 2001. *Science, truth, and democracy*. Oxford: Oxford University Press.
- Kragh, H. 2000. Conceptual changes in chemistry: The notion of a chemical element, ca. 1900–1925. *Studies in History and Philosophy of Modern Physics B* 31: 435–450.
- Kripke, S. 1980. *Naming and necessity* (revised edition). Cambridge, MA: Harvard University Press. (Originally published in 1972.)
- Mazzocchi, M. 2008. Complexity in biology: Exceeding the limits of reductionism and determinism using complexity theory. *EMBO Reports* 9: 10–14.
- Millikan, R. G. 1999. Historical kinds and the “special sciences.” *Philosophical Studies* 95: 45–65.
- Nelson, D. L., & Cox, M. M. 2008. *Lehninger’s Principles of biochemistry* (5th ed.). W. H. Freeman: New York.
- Pennisi, E. 2012. ENCODE Project writes eulogy for junk DNA. *Science* 337: 1159–1161.
- Perini, L. 2011. Sequence matters: Genomic research and the gene concept. *Philosophy of Science* 78: 752–762.
- Portin, P. 2002. Historical development of the concept of the gene. *Journal of Medicine and Philosophy* 27: 257–286.
- Putnam, H. 1973. Meaning and reference. *The Journal of Philosophy* 70: 699–711.
- Slater, M. H. 2009. Macromolecular pluralism. *Philosophy of Science* 76: 851–863.
- Stanford, K. & Kitcher, P. 2000. Refining the causal theory of reference for natural kind terms. *Philosophical Studies* 97: 99–129.
- Tobin, E. 2010. Microstructuralism and macromolecules: The case of moonlighting proteins. *Foundations of Chemistry* 12: 41–54.
- Voet, D. & Voet, J. G. 2004. *Biochemistry*, Volume 1 (3rd ed.). Hoboken, New Jersey: Wiley.
- Weisberg, M., & Needham, P. Matter, Structure, and Change: Aspects of the Philosophy of Chemistry. *Philosophy Compass* 5: 927–937.
- Weisberg, M., Needham, P., & Hendry, R. F. 2011. Philosophy of Chemistry. *The Stanford Encyclopedia of Philosophy*: <http://plato.stanford.edu/entries/chemistry/>

CHAPTER 2: PROTEIN TAXA

In so far as a natural classification is grounded on real Kinds, its groups are certainly not conventional; it is perfectly true that they do not depend upon an arbitrary choice of the naturalist.

John Stuart Mill's *A System of Logic* (1843)

Abstract: The previous chapter dealt with the individuation of proteins. This chapter focuses on protein classification at higher levels—in other words, on protein taxa. Given the extreme complexity of proteins and the myriad interests that govern their classification, there is actually a remarkable degree of consensus regarding how proteins are individuated. But this does not hold for protein classification more generally—for the way in which individuated proteins are organized. As this chapter shows, there are many different ways to relate individuated proteins to one another, and these different ways each produce different protein taxonomies. So at least as it currently stands, protein individuation is overwhelmingly monist, while protein organization is thoroughly pluralist. Due to the complexity of proteins and the practical advantage granted by tracking different aspects of this complexity in different classification systems, it is argued that such pluralism is here to stay.

Section 1: Introduction

Elsewhere I have argued that membership in protein kinds is conferred by an assortment of properties including genetic origination, macroscopic superstructure, somewhat-specific-sequential microstructure, and strict-species-specificity.¹ Although this makes protein kinds rather complicated, it does not—*contra* Slater (2009) or Tobin

¹ Please see section 3.3 of the previous chapter.

(2010)—make them pluralist. And though this monist account of protein kinds does incorporate the microstructural property of primary structure—à la Goodwin (2011)—it does not give that property undue status.

My view of protein kinds also correlates nicely with scientific practice. The view explains the importance of primary structure—again, as stressed by Goodwin (2011)—but it appropriately locates that importance in a host of practices in protein science. This includes the identification, manipulation, and aspects of the individuation of proteins, but it does not mistake primary structure for the “fundamental” individuating property of proteins. If primary structure strictly or solely conferred membership in protein kinds, there would be even more protein kinds than there actually are.²

And there are already an incredible number of protein kinds. Estimates of the number of currently known proteins range from 30 to 50 million.³ Given the staggering number and complexity of these macromolecules, there is actually a remarkable degree of consensus among scientists with respect to the individuation of proteins. There is not—as Slater (2009) seems to indicate—a plurality of ways to individuate proteins. Scientists have developed a system of concepts (including variants, alleloforms, isoforms, and more) with which to manage and track the microstructural, macrostructural, and functional variation among tokens of a single protein type. So there aren’t separate systems for typing proteins by microstructure, or macrostructure, or function; nor is any token protein molecule a member of various protein types at various points in its molecular life. There is just one system for typing proteins; each token protein molecule

² Argued in section 3.4 of chapter 1.

³ More in the introduction to section 3, chapter 1.

is a member of one and only one protein type, from the point of folding throughout its dynamic life until denaturation.

This monism does not hold at higher levels of protein taxa. As I will show throughout this chapter, the complexity of proteins—along with the variety of uses to which knowledge about proteins can be put—has led to a proliferation of ways to organize individuated proteins. By the ‘complexity’ of proteins I simply mean something deflationary, like: these macromolecules have a lot of properties, and these properties interact and cluster in different ways. As a result of this kind of complexity, there are many ways to group proteins according to (different sets of) their properties—or, there are various ways to relate proteins to one another. This means that there is more than one way to classify proteins at higher levels—or, that there is more than just one protein taxonomy.

In this chapter I will both present and account for the existence of this plurality of protein taxonomies, despite the monist individuation of proteins into kinds. Just to be clear, I’m saying that, although everyone agrees on the individuation of the basic entities being sorted, there is a variety of ways of arranging these entities into hierarchical classification schemes, or taxonomies.⁴ And this is all within the relevant scientific field. This is not a case—à la Dupré (1993)—of differences among the projects and purposes of different scientific fields, or at different levels of scientific inquiry, generating different classification schemes. This is a case of the same scientists in the same labs within the

⁴ It’s as if everyone in biology agreed on what the species were, but then disagreed about how to relate said species to one another. Like: sure there’s an evolutionary species tree, but there’s also a species-by-niche tree (functional), and a species-by-geography tree (locational), and maybe also a species-by-body-size-and-shape tree (macrostructural), or even a species-by-entry-onto-Noah’s Ark tree (biblical). In other words, the two cases are not actually all that different. But in the protein case, the multiple taxonomies are serious scientific equals.

same field classifying proteins in a variety of ways, and using these various classification systems interchangeably and even simultaneously.

In the next section of the chapter, section 2, I'll present some relevant information about the complexity of proteins. This will help to explain how the various general systems for classifying proteins, discussed in section 3, are generated. Most of these general classification systems are rather haphazard—they are not meant to be exhaustive, and when they are, they're certainly not meant to be exclusive. So, in subsections 3.1–3.3, I present potential candidates for an exclusive and exhaustive classification system of individuated proteins. In other words, I consider the possibility that the current plurality of protein taxonomies is merely temporary. I examine plausible candidate classifications, which might one day underwrite a comprehensive and complete “one true” taxonomy of proteins. And then I consider obstacles to this possibility.

Since there are literally millions of proteins, it would be impossible to look at all of them, and then determine what the obstacles to each of these potentially monist, unified classification systems might be. So I look at just a subset of proteins here, in section 4—at the nuclear hormone receptor superfamily. But even taking a careful look at just this group of gene regulatory proteins sufficiently demonstrates that the plurality of protein classificatory systems is here to stay.

This pluralism is permanent because all proteins—not just nuclear receptors—are macromolecules with a complex mix of component, microstructural, macrostructural, mechanistic, locational, functional, physiological, and evolutionary properties, just to name a few. All of these properties are of scientific interest, some more than others, and in ways that vary by context. Classifying proteins according to one or another of these

properties, or sets of properties, produces different groupings. Or: different ways of classifying proteins track different properties or sets of properties; the resulting classification systems are not co-extensive.

And none of these (sets of) properties is obviously more right, or real, or helpful than the others. Though biologists or philosophers of biology might tend to assume that evolutionary considerations trump all others when it comes to questions of classification,⁵ at least in this case, the science is riddled with other, non-evolutionary concerns. Proteins are biologically crucial objects but they are also chemical entities; they are biochemical kinds. As I'll show here, even the evolutionary properties of proteins are tracked via microstructural properties. And each of the systems designed to classify proteins via different sets of properties are indispensable to protein science. This is because different sets of properties of proteins correspond with different powers, and tracking different powers of proteins grants researchers different affordances in protein science, all of which are crucial to investigation.

To summarize: in section 4 I use the nuclear hormone receptor superfamily in order to demonstrate that no single protein taxonomy is likely to ever do the classificatory job alone. Looking carefully at this protein superfamily shows that pursuing only one framework for protein classification—even one based on evolutionary history—would impoverish rather than assist protein scholarship. Just looking at this case is sufficient, I propose, to demonstrate that this practical advantage is likely to indefinitely sustain the current plurality of protein classification systems.

In the conclusion of the chapter I reiterate the preceding argument and summarize

⁵ Please see debate re: the species concept.

the philosophical implications of the antecedent discussion. But first: a brief introduction to proteins. The various dimensions of these complex macromolecules (component, contingent, superstructural, functional, evolutionary, and more), discussed next in section 2, will help to explain the variety of protein classification systems, discussed later in section 3.

Section 2: A Primer on Proteins

There are more potential proteins than there are actual atoms in the universe. Here's why: proteins are macromolecules, or very large molecules. Macromolecules are composed, and are not simply aggregates, of many smaller molecules.⁶ Among the various kinds of macromolecules, proteins are polymers: long chain molecules, created by sequential bonding of many repeating units of smaller molecules, or monomers.

In the case of proteins, the monomers in question are amino acids, and they are joined together via peptide bonds. There are 20 standard amino acids, and they were discovered over more than a century: leucine (1819); glycine (1820); tyrosine (1846); serine (1865); glutamic acid (1866); aspartic acid (1869); asparagine (1873); glutamine (1873); alanine (1875); phenylalanine (1881); lysine (1889); cysteine (1890); arginine (1895); histidine (1896); valine (1901); proline (1901); tryptophan (1901); isoleucine (1903); methionine (1922); threonine (1936).⁷

⁶ This difference between composition and aggregation is worth noting because of some debate, during the early decades of the twentieth century, over whether proteins were 'true' macromolecules or simply 'colloids'—large molecules that were simply aggregates of smaller, 'truer' molecules called 'crystalloids'. For more, please see Olby's *The Path to the Double Helix: The Discovery of DNA* (1964).

⁷ There are two additional amino acids much more recently found in rather specific contexts: selenocysteine (1976) and pyrrolysine (2004). See Rother & Krzycki's "Selenocysteine, Pyrrolysine, and the Unique Energy Metabolism of Methanogenic Archaea" (2010). Sometimes selenocysteine gets included in the list

Just over half of the 20 standard amino acids can be created within the human body from other compounds; the other 9 are called the ‘essential’ amino acids because they must be ingested in order for humans to have them available for building the proteins that contain them. Proteins are built out of amino acids by joining them together in chains, each amino acid bonded to its neighbor on either side, and then, usually, looping and folding the chain into a globular structure. Between two and about 50 amino acids bonded together constitute what is called a ‘peptide’; when more than 50 are bound together the macromolecule is called a ‘polypeptide’. So, proteins are polypeptides: long and often clumped together chains of amino acids each linked to its two neighbors via covalent peptide bonds.

The largest known protein is titin,⁸ with a sequence of approximately 30,000 amino acids. Using that as an upper bound on the number of amino acids that might be joined together in sequence to form a protein, and given that there are 20 different amino acids, there are $20^{30,000}$ conceptually possible amino acids sequences within that space of potential proteins. Even given that large swaths of this space might be deemed physically impossible due to molecular restrictions on viable chaining, this is still an astronomical space. $20^{30,000}$ is equivalent to $2^{30,000} \times 10^{30,000}$. Just 2^{50} is close to a quadrillion. One of the largest numbers in common usage in the googol, which is 10^{100} and does not come even close to $10^{30,000}$. And those two terms are supposed to be multiplied together. For another point of comparison, the estimated number of atoms in the universe is something close to 10^{80} . So there are far, far more potential proteins than there are actual atoms.

of standard amino acids, bringing their number to 21.

⁸ ‘Titin’ is the general term for any of the species-specific versions of this protein, which is found in many organisms. There is human titin, or hTitin; mouse titin, or mTitin; etc. And yes, the name is a play on the Greek mythical race of great deities, the Titans.

The size of the space of possible proteins is worth considering because it explains why these macromolecules are, as a group, so exceptionally plastic. There are an unimaginable number of possible configurations that they can take. But individual proteins are also contextually plastic⁹—they can take different shapes in different circumstances. For example, almost all known enzymes—highly selective catalysts for metabolic reactions—are proteins. The catalytic power of specific enzymes varies greatly in the presence or absence of certain activators or inhibitors. These molecules change the shape of the enzyme so that it is more or less conducive to the substrate of the catalytic reaction. So, these proteins take various shapes, and differences in shape determine differences in activity.

This plasticity of configuration comes from the fact that, despite their basic chaining, the native state or natural formation of most of these molecules is not simply that of a straight line. Although some fibrous proteins approach this kind of one-dimensionality, many proteins are globular and thus have a complex superstructure that cannot simply be captured or conveyed by primary amino acid sequence.

As a result of this complexity, the structure of proteins is separated into four distinct levels: primary, secondary, tertiary, and quaternary. See figure 2.1 below:

⁹ I am borrowing this term and its particular usage from Sandra D. Mitchell's *Unsimple Truths* (2009).

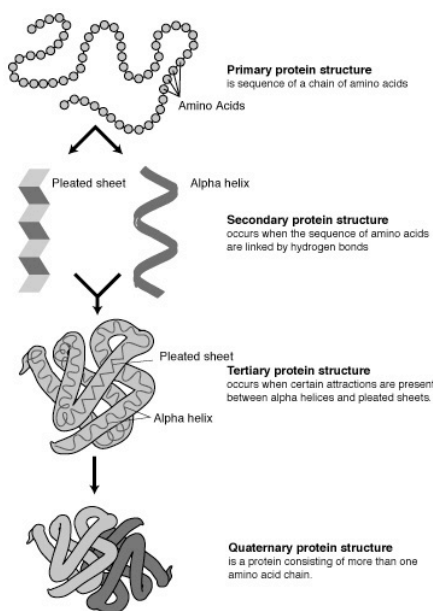


Figure 2.1: Four levels of protein structure (image from the National Human Genome Research Institute).

The linear sequence of amino acids is known as the primary structure of a protein.

Amino acids all have a central or alpha carbon, with four groups bonded to this tetravalent center: an amino group, a carboxyl group, a hydrogen atom, and a side chain or R-group. See figure 2.2:

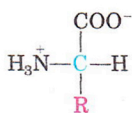
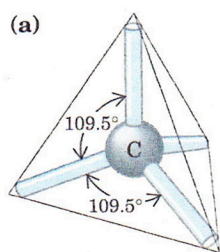


FIGURE 3-2 General structure of an amino acid. This structure is common to all but one of the α -amino acids. (Proline, a cyclic amino acid, is the exception.) The R group, or side chain (red), attached to the α carbon (blue) is different in each amino acid.

FIGURE 1-14 Geometry of carbon bonding. (a) Carbon atoms have a characteristic tetrahedral arrangement of their four single bonds.

Figure 2.2: A helpful pair of diagrams from *Lehninger's Principles of Biochemistry* (Nelson & Cox 2008). On the left is a depiction of the tetravalent nature of carbon (p. 12); on the right is a schematic of the basic structure of amino acids (p. 72). [Color figure available online.]

The amino and carboxyl groups of different amino acids are what join together to form

the peptide bond linking each amino acid in sequence to the ones on either side of it. The R-group is what varies with each amino acid. The R-groups of the 20 standard amino acids can be divided into five groups with characteristic properties: there are nonpolar aliphatic, aromatic, polar uncharged, positively charged, and negatively charged R-groups. Which kinds of R-groups are next to each other and how they are clustered along the amino acid sequence of the protein is what determines the secondary structure of the protein.

For example, the turn known as an alpha helix is a common secondary structure of proteins. An alpha helix is formed when the linear sequence of an amino acid coils itself by forming a hydrogen bond between each amino group and the carboxyl group four amino acids away. The sequence turns towards itself in order to put these groups, which are not otherwise next to each other on the chain, within reach.

The alpha helix structure is a common one, and is quite stable, because it allows for polar groups along the sequence to form hydrogen bonds, making the exterior surface of the structure hydrophilic, while clustering the hydrophobic groups in the interior of the structure, in an energetically favorable state. However, not all polypeptide sequences can form alpha helices, because of unwieldy side chains. The R-groups affect whether alpha helices or other secondary structures emerge from the primary structure of the protein—providing a nice example of just one of the many ways in which interactions among amino acids—the parts—affect the composition and structure of the entire protein—the whole.

How these secondary structures interact with each other and with particular R-groups determines the tertiary structure of the protein. Tertiary structure interactions

include hydrogen bonding (often due to polar R-groups), ionic bonding (due to charged R-groups), hydrophobic interactions (because of the thermodynamics of hydrophilic R-groups), and even covalent bonding (of disulfide linkages between R-groups with a sulfur atom, like cysteine). Adding the interactions from this tertiary level of protein structure to the primary and secondary structures completes the overall conformation of any protein composed of a single amino acid sequence.

But multiple polypeptide sequences with their own secondary and tertiary structure sometimes interact to form single proteins known as multiple-molecule complexes or multisubunit proteins, and this generates quaternary structure. In other words, a protein with quaternary structure is composed of two or more amino acid chains, each with their own primary, secondary, and tertiary structures, but also engaging in (multipart causal) interaction with each other. The same interactions that characterize the tertiary structure of proteins characterize quaternary structure, but in this case the interactions are between the R-groups, shapes, and overall conformation of different (token, not necessarily type) amino acid sequences. For example, hemoglobin, possibly the most thoroughly studied protein, has quaternary structure. It is composed of four separate polypeptide sequences of two different kinds (two tokens each of two types).¹⁰

To summarize: proteins are large molecules composed of amino acids, covalently linked into a chain via peptide bonds, but also contorted due to a host of other chemical

¹⁰ The case of hemoglobin is quite an important one. For example, one might wonder why proteins with quaternary structure (i.e., proteins composed of distinct amino acid sequences) are even considered single proteins at all. This, I believe, has to do with the fact that the crystallization of hemoglobin in the middle of the 19th century was a key event in the history of protein science, as was the discovery of its structure by X-ray crystallography nearly 100 years later. Hemoglobin has always been considered a prominent and paradigmatic protein, and upon the discovery of its quaternary structure, it might just have seemed less radical to simply add a fourth level to the structure of proteins than to banish hemoglobin from the category entirely. But further scholarship is required to support this claim.

interactions that occur among the variable R-groups, the shapes that clusters of these R-groups form, and even the various shapes formed by separate sequences. The incredible complexity of this kind of macromolecular conformation is managed (somewhat) by dividing protein structure into a hierarchy of four levels: primary structure (amino acid sequence, producing chains); secondary structure (local bond formation between amino acids, producing shapes like turns and sheets); tertiary structure (additional interactions between elements scattered throughout the molecule, completing overall conformation); and quaternary structure (multiple-molecule shape formation, producing complexes).¹¹

Section 3: The Classification of Proteins

In this section I'll use what I've just presented about proteins in order to explain the various systems that have arisen for organizing the (incredibly large) group of currently known proteins.

Again, the number of currently known proteins is estimated to be somewhere between 30 and 50 million.¹² One way to bring order to such a large and haphazard heap of proteins is to develop systems for classifying them—so perhaps unsurprisingly, there is no shortage of such systems. To start, the entire group of proteins can be initially divided into the two (aforementioned in section 2) structural classes of fibrous and globular. Enzymes (often labeled with the suffix “ase”) can also be and often are singled out for discussion as a distinct group, in part because of their importance to metabolic biochemistry.

¹¹ For more information please see, for example, *Lehninger's Principles of Biochemistry*, most recently in the 5th edition (Nelson & Cox 2008).

¹² Chapter 1, section 3.

Alternatively, proteins can be divided into the classes of simple and conjugated. Many proteins are composed solely of amino acids—these are the simple proteins—but others contain additional chemical components (not amino acids) that are permanently associated with the protein—these are the conjugated proteins. Conjugated proteins can be further classified according to the kind of chemical, known as the prosthetic group, attached to the protein. For example, there are the classes of: lipoproteins (prosthetic group—lipids); glycoproteins (prosthetic group—carbohydrates); phosphoproteins (prosthetic group—phosphate groups); hemoproteins (prosthetic group—heme, or iron porphyrin); flavoproteins (prosthetic group—flavin nucleotides); and metalloproteins (prosthetic group—iron, zinc, calcium, copper, and others).

However: none of these classificatory systems are meant to be exclusive, or even (in some cases) exhaustive. There are some candidate classification systems that attempt to be more rigorous, however. One current example of a more systematic and thorough kind of attempt is based on the two most common secondary structures found in proteins: alpha helices (discussed in section 2) and beta sheets (sets of beta strands, each approximately 3–10 amino acids long, linked to each other via hydrogen bonds). In other words, this attempt to systematically classify proteins is based primarily on considerations of secondary structure.¹³

In the rest of this section I will discuss various classification systems for proteins, each of which uses a different primary feature as the basis for classification. One of the classification systems is a fully actual one, which already exists; the other two are

¹³ In what remains of the chapter I'll use the term 'macrostructural' to refer to any considerations of secondary, tertiary, or quaternary structure, aka superstructure. I'll restrict use of the term 'microstructural' to considerations of primary structure.

possibilities, still in development. I will discuss the actual, superstructural case first, and then I will examine two other significant considerations—functional and evolutionary—that could each be used as the basis for a systematic classification of proteins. In the section that follows this one I'll use details from the nuclear hormone receptor superfamily, and its various organizational systems, in order to argue that no one classification system for proteins is ever likely to battle for supremacy, emerge triumphant, and replace all others.

Sub-Section 3.1: Superstructural Considerations

Basic information about protein structure is stored in the Protein Data Bank (PDB).¹⁴ Other databases take information from that depository and sort it more selectively, like the Structural Classification of Proteins (SCOP) database.

SCOP takes information from PDB and divides all known proteins into four classes: (i) proteins whose secondary structures are entirely alpha helices; (ii) proteins whose secondary structures are all beta sheets; (iii) proteins whose secondary structures include both alpha helices and beta sheets, interspersed or alternating; and (iv) proteins whose secondary structures include both alpha helices and beta sheets, but which are somewhat segregated from each other. Alpha helices were described in the previous section; the figure below illustrates the way in which both alpha helices and beta sheets are depicted in protein diagrams:

¹⁴ Introduced in the previous chapter (section 3).

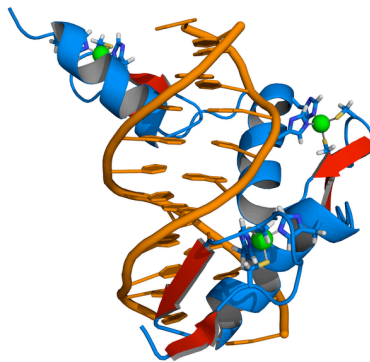


Figure 2.3: Diagram of a (zinc finger) protein with both alpha helices and beta sheets. The curly-cues or telephone-cord shapes represent the alpha helices; the arrows represent the beta sheets (from http://en.wikipedia.org/wiki/File:Zinc_finger_DNA_complex.png).

The proteins within these four classes are then further demarcated according to the presence of recognizable supersecondary¹⁵ structures, or folds. These folds consist of stable arrangements of several secondary structure elements (like alpha helices or beta sheets) and assorted connecting regions. Of some folds there is only one known exemplar; others are found in a variety of proteins, as recurring structural patterns, and these are called motifs:

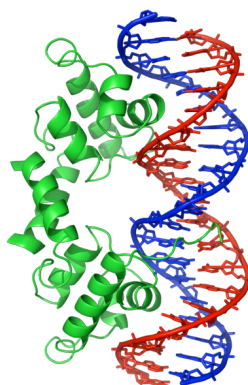


Figure 2.4: On the left, the DNA-binding helix-turn-helix motif next to a strand, on the right, of DNA (from http://en.wikipedia.org/wiki/File:Lambda_repressor_1LMB.png).

¹⁵ This is another kind of protein macrostructure, one that can include multiple secondary structures and occasionally tertiary structure. Basically, a supersecondary structure is a larger chunk of a protein than just one secondary structure, which may or may not include tertiary elements. For instance, imagine a segment of a protein that consists of a certain combination of alpha helices and beta sheets—that segment could be a supersecondary structure.

After these superstructural discriminations, the proteins in each category are arranged according to amino acid sequence alignment (i.e., according to microstructural properties).

Some motifs, or commonly found folds, are also associated with a particular function. This functional association introduces the concept of domains. Domains are regions of proteins that are identified as independent functional units and are thus considered distinct superstructural units of a polypeptide. They usually have a stable, compact globular structure and are composed of one or more motifs.

Sub-Section 3.2: Functional Considerations

There is no comprehensive classification system for proteins based primarily on domains. But they are worth discussing here in a bit more detail, because looking at domains is one way of seeing the evolved contingency¹⁶ of proteins. And the concept of evolved contingency is crucial to the combination of component, microstructural, macrostructural, mechanistic, locational, functional, physiological, evolutionary, and other properties that significantly complicates any proposal to classify proteins purely along just one of those lines (as will be shown in section 4).

Genes, like proteins, display evolved contingency, though in a certain limited way. This is because every gene that has evolved could have evolved differently, at least by a little bit.¹⁷ And proteins are similarly contingent, since in many cases swapping one

¹⁶ As with ‘contextual plasticity’, the particular use and meaning of this term is borrowed from Mitchell (2009).

¹⁷ Genes encode proteins by indicating primary amino acid sequence with trios of nucleobases. There are four different kinds of nucleobase in DNA—adenine, cytosine, guanine, and thymine—and thus there are 4^3 or 64 possible arrangements of three of the four bases in sequence. But there are only 20 amino acids.

amino acid for another at any given point in the sequence will not substantially alter higher-level structure or overall configuration of the protein. But proteins also display a much more radical kind of evolved contingency, perfectly conveyed by the distinction between a fold and a domain. Folds track protein structure; domains track function. In other words, a protein has a certain fold insofar as it has certain *properties* of amino acid sequence and overall configuration, whereas the protein contains a certain domain insofar as it has certain *powers*, of substrate, ligand, or DNA binding, say.¹⁸

For example, assorted proteins contain various folds and motifs arranged in ways that enable them to bind to DNA, so these arrangements are called DNA binding domains. Since there are various arrangements that confer this ability, there is also a variety of DNA binding domains, and they are not necessarily structurally or evolutionarily related. The helix-turn-helix DNA binding motif (depicted in figure 2.4) is one kind of DNA binding domain found in many bacterial and eukaryotic regulatory proteins, for example, whereas the zinc finger DNA binding motif (shown in figure 2.3) is another kind of DNA binding domain, found in many eukaryotic regulatory proteins. These two motifs, despite both being classified as DNA binding domains, are quite distinct, structurally and evolutionarily.

But to return to the issue of comprehensive attempts to organize proteins: SCOP, which emphasizes protein structure, is one actual, already developed classification system. The idea of domains, though important for understanding protein function, does

All but two of them can be indicated by more than one trio, and there are no proteins made only of methionine and tryptophan. As a result, every known gene that codes for a protein could have had a slightly different nucleobase sequence than it actually does, while still producing the same amino acid sequence.

¹⁸ The importance of this distinction should be apparent by the conclusion of the paper.

not as yet fundamentally underwrite any comprehensive classificatory attempts at protein organization.

Sub-Section 3.3: Evolutionary Considerations

There is, however, a plausible historical and conceptual basis for attempting to classify proteins primarily evolutionarily. This is because proteins are often sorted into superfamilies, families, and subfamilies based on the extent of their evolutionary relationships (which are inferred from algorithmic measurements of percent sequence alignment—or, microstructure¹⁹).

The nomenclature of superfamilies, families, and subfamilies is taken from species classification and its use began decades ago, with the development of Margaret O. Dayhoff's *Atlas of Protein Sequence and Structure*. The *Atlas* was first produced in the mid sixties (Dayhoff & Eck 1965) and production continued for little over a decade. During its development, Dayhoff (the recurring chief editor of the *Atlas*)—as well as others involved in that and related projects—began searching for an apt nomenclature for the organization of proteins.²⁰ At the time it was presumed that evolutionary relationships would dictate appropriate protein- as well as organism-classification schemas, so molecular biologists looked to the evolutionary taxonomy of species, which had inherited the nomenclature of the original, Linnaean taxonomy.

¹⁹ Yes: in protein science evolutionary considerations are primarily tracked via microstructural variation. As I said in the last chapter, it's a mistake to think that chemistry is about structure and biology is about function. Functional and macroscopic considerations, as well as microstructure, play a role in chemistry. And microstructural considerations play a role in the study of evolution, that oh-so-biological subject.

²⁰ For a wonderful account of Dayhoff's career and the development of the *Atlas*, see Strasser's "Collecting, Comparing, and Computing Sequences" (2009).

Linnaeus proposed his taxonomy of organisms in the middle of the 18th century.²¹ He was not the first to offer a rank-based classification of living things, but his contribution formed the basis for the system still used today. The Linnaean hierarchy did not originally include the terms ‘superfamily’, ‘family’, or ‘subfamily’,²² but Linnaeus eventually used the term ‘family’ in discussion of plant groups.²³ It was often used as an equivalent of ‘order’, until the end of the eighteenth century, when introduced as a rank between order and genus by Pierre André Latreille.²⁴ Further division into sub- and super-families began nearly a century later, at the end of the 19th century.

The nomenclature of the Linnaean taxonomic system was preserved even through the transition from morphological to evolutionary taxonomies of biological species following the Darwinian revolution in biology. So, when Dayhoff looked to evolutionary taxonomies of species for relevant taxonomic nomenclature, she ended up with what were actually the terminological remnants of a morphological, rather than evolutionary, system. She was not the only one.

For example: at the beginning of the 20th century, in the field of linguistics, the evolutionary tree of life was used as a metaphor for the evolution of languages. Terms from the biological taxonomy were used in the organizing and naming of linguistic groups, beginning in the 1920s,²⁵ and this is the origin of the terminology of language families and superfamilies.

²¹ In his *Systema Naturae*, first published in 1735.

²² It was limited to ‘kingdom’, ‘order’, ‘genus’, ‘species’, and then one rank below species, with a different name in each kingdom. For animals, ‘subspecies’; in plants, it was ‘varieties’; and in Linnaeus’ mineral kingdom, ‘forms’.

²³ In his *Philosophia botanica* (1751).

²⁴ In his *Précis des caractères génériques des insectes, disposés dans un ordre naturel* (1796).

²⁵ Tracked via citations in ISI’s Web of Science.

It was about forty years after the linguists that Dayhoff similarly appropriated terminology from species classification in the process of constructing the *Atlas*. She summarized the protein classification project in 1976: “in order to deal with the totality of protein sequence information in all living forms, some system of classification must be agreed upon... we have found useful a hierarchical division into superfamilies, families, subfamilies, and *Atlas* entries that is based on the degree of similarity among sequences” (Dayhoff 1976, p. 2132). The following figure shows the number of superfamilies, families, subfamilies, and *Atlas* entries that had been collected and compared by 1983:

SUPERFAMILY ORGANIZATION			
Number in group	Group	Criteria for clustering	Number of residues in first sequences of each group
512	Superfamilies	Probability of similarity by chance $< 10^{-6}$	98,394
774	Families	$< 50\%$ different	142,045
1161	Subfamilies	$< 20\%$ different	192,097
1667	Entries	$< 5\%$ different	258,126

Figure 2.5: A snapshot that shows the superfamilies, families, subfamilies, and entries identified by the Atlas project by 1983, as well as the criteria of establishment for each division (Dayhoff, Barker, & Hunt 1983, p. 529).

Dayhoff’s terminology, as well as her methods, continue to be applied. In the paper that contains the table shown above, the authors speculated that there might be only about 1,000 superfamilies in total (Dayhoff, Barker, & Hunt 1983, p. 529). And as the figure shows, only 774 families had been identified at that time. Current estimates of the number of protein superfamilies hovers around 2,000; the estimated number of families ranges from four thousand to sixty thousand.

Section 4: Classificatory Complications

It is now time to consider the possibility that the current plurality—of actual and potential systems for classifying proteins—is merely temporary. Despite the current diversity, might there be a one true protein classification system waiting to emerge from the wings? Let's pick one contender to consider for the hypothetical monistic throne.

Because evolutionary considerations are somewhat paramount in biology, the evolutionary superstructure for sorting proteins (discussed just above in subsection 3.3) might suggest—perhaps to a philosopher of biology, or a philosophically-minded biologist—the possibility of sorting proteins not just into families and superfamilies but rather into a full, hierarchical evolutionary taxonomy of proteins.

There is, for instance, some relevant terminology from evolutionary genetics that our hypothetical philosopher-biologist could apply to protein classification and use to fill out the lower levels of a taxonomy of proteins that has superfamilies, families, and subfamilies as its upper taxa. Perhaps this is a candidate classification system for proteins that might—eventually at least—supplant all others. In other words, might evolutionary considerations underwrite a monist classification of individuated proteins?

Evolutionarily related gene sequences are often identified as homologs—meaning, the gene sequences share a common ancestor and are thus related by descent. There are two different kinds of homologs: orthologs and paralogs. Paralogs are gene sequences differentiated by duplication events, whereas orthologs are gene sequences differentiated by speciation events. For example, imagine a gene sequence GENE in the genome of an evolutionarily distant ancestral population of the made-up species *Latinus latin*. A genetic duplication event results in two sequences, GENE α and GENE β —these

are paralogs. Then a speciation event divides the ancestral population into *Latinus latin* and *Splatinus splatin*. There are now four sequences: lGENE α , lGENE β , spGENE α , and spGENE β . lGENE α and spGENE α are orthologs, as are lGENE β and spGENE β . Both lGENE α and spGENE α are each still paralogs of lGENE β and spGENE β . All are homologs of each other.

Because proteins are the product of gene expression, the homology of gene sequences that produce certain proteins can be extended (and often are, by protein scientists) to the proteins themselves. In other words, if one gene sequence that produces a certain protein is a paralog or an ortholog of another gene sequence that produces a different protein, scientists can and do say not just that the gene sequences are paralogs or orthologs but that the proteins are too.

Many proteins found in one species have related orthologs in other species; and some groups of orthologs can be related to other groups of orthologs all as paralogs. So, within subfamilies proteins could be stored into paralogs and then orthologs. Moving down the hypothetical hierarchy, at the next level would be individuated, species-specific proteins, such as human titin.²⁶ There are many orthologs of human titin, such as mouse titin. The titins also have several known paralogs, such as the Death-Associated Protein Kinase (DAPK) proteins.

But to return to the level of species-specific proteins: within a species or even within an individual, different versions of the “same” protein type can often be found in various alleloforms and isoforms. Isoforms are functionally distinct versions of a single species-specific protein; alleloforms are sequentially distinct but supposedly functionally

²⁶ Titin was already mentioned in section 2. It’s the largest currently known protein.

identical versions.²⁷ There are 10 known isoforms of human titin, for example, and 19 alleloforms of DAPK1. Finally, there are many instances of a single titin isoform in any given human individual. So at the bottom of the hypothetical, evolutionary taxonomy of proteins would be particular, token proteins—just as there are particular, token organisms at the bottom of the evolutionary taxonomy of biological organisms.

To summarize, our imaginary philosopher-biologist might wish to compare today's evolutionary classification system(s) for biological organisms with a potential, also evolutionary, system for the classification of proteins. In both cases the structure of the classification system is classically taxonomic—in other words, these are hierarchical classification schemes, composed of nested groups that are ordered together as progress is made up the hierarchy:

ORGANISMS	PROTEINS
KINGDOM	SUPERFAMILY
PHYLUM	FAMILY
CLASS	SUBFAMILY
ORDER	PARALOG
FAMILY	ORTHOLOG
GENUS	PROTEIN (TYPE)
SPECIES (TYPE)	ISOFORM
SUBSPECIES	ALLELOFORM
ORGANISM (TOKEN)	PROTEIN (TOKEN)

Figure 2.6: On the left, the taxa of standard biological classification,²⁸ progressing up the hierarchy from least to most inclusive. On the right, a prospective taxonomy for protein classification, similarly structured from least to most inclusive categories.

In the taxonomy of proteins pictured above, superfamilies contain families, which contain

²⁷ Gene sequences altered by mutation are called 'variants' and these variants produce alleloforms and isoforms (or, to use the terminology applied specifically for humans, polymorphisms and disease variants).

However, protein isoforms can also be generated in non-mutational ways. Alternative splicing and translational modification are two ways, for example, in which "one" species-specific protein might end up in different forms.

For a much more detailed discussion of these issues, please see chapter 1.

²⁸ Domains, though a currently popular taxonomic rank, have not yet been admitted into the official nomenclature.

subfamilies, then groups of paralogs, orthologs, species-specific proteins (the primary type into which token proteins are individuated), isoforms, alleloforms, and finally, token proteins, or particulars. In the taxonomy of organisms, kingdoms contain phyla, which contain classes, then orders, families, genera, species (similarly noted as the primary type into which token organisms are individuated) subspecies, and finally, token organisms, or particulars.

But the temptation to impose such a neat and tidy classificatory structure—to base the organization of all proteins on evolutionary considerations alone—should be resisted. Throughout the rest of this section I will discuss one particular protein superfamily (the nuclear hormone receptor superfamily) and show how prioritizing different sets of properties (among the microstructural, macrostructural, mechanistic, locational, functional, physiological, and evolutionary) produces three different classification systems of the superfamily. As I will show, each of these classification systems tracks important information about the nuclear receptors, and affords different research strategies. None of them are disposable; even emphasizing one over the others would misrepresent or skew research in the field.

Sub-Section 4.1: Nuclear Receptors

The nuclear hormone receptor superfamily is an excellent case for discussion because nuclear receptors are simultaneously complex and crucial molecules. Identification of the nuclear receptors and collection of them into a superfamily began in the 1980s. Actually, two very important (for molecular and biomedical research) protein superfamilies were identified in that decade.

The first of these two was the immunoglobulins, and there was quite a fight over how to classify these proteins.²⁹ They didn't meet the standards in terms of sequence alignment for a family, but they were clearly related and were eventually established as a superfamily in the early eighties.³⁰ The term was still in circulation, and had retained some of its novelty in application to proteins, when the group of nuclear receptors began to take shape in the mid eighties. Various terms like 'class', 'family', 'paradigm', and 'system' were all used before 'superfamily' was settled on near the end of the decade.³¹

The nuclear hormone receptor superfamily groups together a set of molecules that act both as receptors and transcription factors. Like many other kinds of receptor, the presence of an appropriate signaling molecule induces a conformational shift in the structure of the nuclear receptor, activating it. Unlike other kinds of receptors, activated nuclear receptors bind to response elements located directly on the genome, within the nucleus of cells. Informally, these response elements are called 'genetic activation switches'. These switches control the action of, not just singular, but often whole sets of genes. Thus nuclear receptors are influential molecules that can, when they are themselves activated, up- and down-regulate the action of many different genes at once. In other words, these gene regulatory proteins are not only receptors, but also powerful multiple-transcription factors.

²⁹ Traces of this conflict are visible today in a series of articles, news bits, and even impassioned letters scattered throughout the journals *Immunology Today*, *Nature*, and *Science*.

³⁰ The argument is summarized by Alan F. Williams in a letter (1984) to *Immunology Today*.

³¹ As the group began to take shape there was initial uncertainty about what to call it. In the presentation abstract of one of the initial talks on the subject of a potential "family of eukaryotic transcriptional regulatory factors" (Evans, Weinberger, Hollenberg, & Rosenfeld, 1986, p. 63), the terms 'paradigm', 'class', 'mechanism', and 'family' were all used. The term 'superfamily' was first used publicly in another talk later that year, entitled "Human Steroid-Receptors and erbA Protooncogene Products: Members of a New Superfamily of Enhancer Binding-Proteins" (Weinberger, Giguere, Hollenberg, Rosenfeld, & Evans, 1986).

The very multiplicity and power of the transcription factor ability of nuclear receptors makes them a simultaneously enticing and frustrating target for intervention. Nuclear receptors often have different targets in different cell types, so activating one can mean activating not just one response element, but many. When each response element in turn activates a variety of genes, there are myriad cascading effects.

On the one hand, receptors seem like master switches that can be used to control complex genetic networks whose malfunction can be physiologically devastating. On the other hand, the fact that these receptors control so many things—often, different pathways in different cell types—means that it is very difficult to intervene on one malfunctioning pathway without also targeting some other normally functioning pathway.³²

But the nuclear receptors are not simply scientifically and commercially arresting—they are philosophically provocative as well. This is because they are an incredibly well studied group of molecules, especially in mice and humans, and there are still at least three major systems for classifying these gene regulatory proteins. There are phylogenies of the superfamily derived primarily from both evolutionary and microstructural properties; there are ways of typing the nuclear receptors according to a variety of superstructural, mechanistic, and (intracellular) locational properties; and there is one final way of organizing the superfamily by expression pattern, of those genes which code for the nuclear receptors. This last is a way of organizing the superfamily that prioritizes (intercellular) locational, functional, and physiological properties.

³² Much to the chagrin of scientists around the world, and to various pharmaceutical outfits, most of the drugs that have so far been developed as ligands for various nuclear receptors have turned out to generate side effects just as devastating as the original reason for treatment. Many of the rest have simply been ineffective.

Sub-Section 4.2: A Brief Note about Reference

I will explain each of these three classification systems in turn. But first, for the sake of clarity in the ensuing discussion, it should be noted that a reference to “the” nuclear hormone receptor superfamily (NHRSF) is often ambiguous, in at least three different ways. For one, use of the term does not necessarily refer to all the known nuclear receptors (NRs) found throughout the animal kingdom. The term often refers only to those nuclear receptors found in mice and humans, or just humans, or mice, humans, and rats. Of the hundreds of known nuclear receptors throughout the animal kingdom, only about 50 of them are found within mice, humans, and rats. So the size of the superfamily can vary quite a bit—from hundreds to just under 50.

For another, reference to even just the human nuclear hormone receptor superfamily may include between 48 and 52 nuclear receptors. There are 48 commonly recognized members of the nuclear hormone receptor superfamily in humans. Once it was thought that there were 49, but the FXR receptor has two known forms, FXR α and FXR β , and it turns out that in humans the FXR β gene is a pseudogene (Zhang et al. 2008). And then there are three gene sequences that may or may not code for receptors—the sequences are related to those of the nuclear receptors, but appear to have coding regions for two DNA binding domains each (Wu et al. 2007). Since there is widespread uncertainty regarding how to classify these receptors, sometimes they are included in the superfamily. Other times they are not.

It should be clear by now that the common practice of referring simply to “the nuclear hormone receptor superfamily” is somewhat ambiguous, often exclusionary, and occasionally misleading. A third and final note of confusion: the nuclear hormone

receptor superfamily is also occasionally called ‘the nuclear receptor superfamily’, ‘the nuclear hormone receptor family’, or even just ‘the nuclear receptor family’. I will continue to use ‘the nuclear hormone receptor superfamily’ to consistently refer to the 48 commonly recognized nuclear receptors found in humans, unless otherwise noted.

Sub-Section 4.3: Phylogenies of the NHRSF

Here is a classification of the nuclear hormone receptor superfamily based on a combination of evolutionary, functional, and microstructural considerations. It is what is standardly called a ‘phylogeny’³³ of the nuclear hormone receptor superfamily:

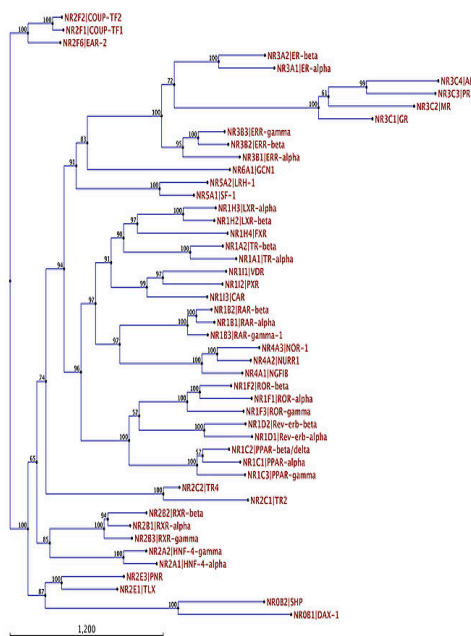


Figure 2.7: A phylogeny of the nuclear hormone receptor superfamily (altered from http://en.wikipedia.org/wiki/File:Nr_alignment_tree.jpg).

This phylogeny includes only the 48 nuclear receptors specific to humans (so, these are species-specific protein types). It was generated via sequence alignment, because this is

³³ Though please note that this is a gene, not a species tree.

how evolutionary relationships among proteins are standardly inferred.³⁴ Within the method of sequence alignment, however, there are actually several different ways of generating a phylogeny of even just these 48 species-specific types, or individuated proteins.

There are different ways to generate nuclear receptor phylogenies via sequence alignment because nuclear receptors are multiple-domain proteins: they have five distinct functional regions (more on this in the next subsection). The various domains in multiple-domain proteins have not necessarily evolved at the same rates. Since nuclear receptors are multiple-domain proteins, their evolutionary history is potentially fragmented in this way, and evolutionary relationships are often tracked among the domains within the receptors rather than among the receptors in their entirety.

Figure 2.7 depicts a phylogeny of the nuclear hormone receptor superfamily generated via whole sequence alignment. But this is unusual for the superfamily: most nuclear receptor phylogenies (and there are other several candidates on offer), although purportedly of the superfamily itself, are actually phylogenies of the DNA-binding domain, ligand-binding domain, or both of these together without the other three domains. In other words, it is not whole-sequence alignment that is standardly used to establish evolutionary relationships among the nuclear receptors; rather it is only partial-sequence alignment. And choosing one domain over another for alignment purposes can produce different phylogenies.

But there are not only various phylogenies of the nuclear hormone receptor

³⁴ Again, this is a blatant case of evolutionary considerations being tracked almost entirely with structural markers.

superfamily; there are also entirely non-phylogenetic ways of organizing the superfamily.

Sub-Section 4.4: Types of NRs

One very prominent, and in fact the oldest, way of dividing the nuclear hormone receptor superfamily separates the nuclear receptors into general types³⁵ based on various features—superstructural (of the protein), mechanistic (in terms of receptor activation and binding mechanisms), and locational (within the cell)—which determine the method of transcriptional regulation by these gene regulatory proteins.

All nuclear receptors have five domains:

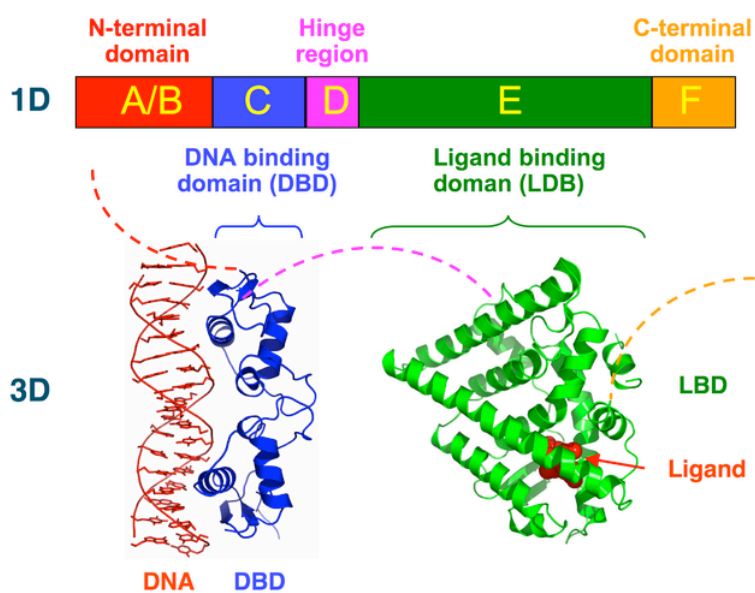


Figure 2.8: The five domains of a nuclear receptor (altered from http://en.wikipedia.org/wiki/File:Nuclear_Receptor_Structure.png). [Color figure available online.]

The A/B and F domains on either end are highly variable in this group of proteins and therefore not very well studied; for nuclear receptors the traditional areas of focus are the

³⁵ Since I've been using the type/token distinction to unambiguously refer to individuated proteins (types) and particular proteins (tokens), this is a rather unfortunate term with which to name these groups of NRs. That, alas, is the term of art.

C and E domains; and the D domain is sort of a flexible hinge between those two. The E domain is the Ligand-Binding Domain (LBD); it is a moderately conserved in sequence and highly conserved in function “sandwich fold” domain; so, a nuclear receptor remains in the cytoplasm or the nucleus until a ligand arrives and that molecule gets “sandwiched” as the filling in the hydrophobic core of the molecule. This begins a cascade of events that involves the C domain, also known as the DNA Binding Domain (DBD), a highly conserved in both sequence and function “zinc fingers” domain with which a nuclear receptor binds to a response element on the genome.

Differences in these domains affect the mechanisms by which the nuclear receptors act, and so these differences can be used to organize the superfamily into four distinct types: type I nuclear receptors reside in the cytoplasm, form homodimers,³⁶ and bind at inverted repeat response elements; type II nuclear receptors reside in the nucleus, form heterodimers,³⁷ and bind at inverted repeat response elements; type III nuclear receptors are similar to type I but bind to direct repeat response elements; and type IV nuclear receptors can be either monomers or dimers but they use just a single DNA binding domain to bind to a half-site response element.

Many receptors have been unambiguously identified as type I or type II. However, classification of receptors as type III or type IV is not as common, for several reasons. For one, there are not that many known type III or IV receptors. For another, the type IV category is so similar to that of type I that not everyone agrees such receptors ought to be in a separate group. Finally, many of the receptors have not yet been

³⁶ Meaning that, when activated, each receptor pairs up with another copy of the same molecule.

³⁷ Meaning that, when activated, each receptor pairs up with a different molecule—in the case of these nuclear receptors, usually with another nuclear receptor, RXR.

identified as any particular type.

Some of the nuclear receptors initially placed into the superfamily by sequence alignment have since been identified as nuclear receptors more thoroughly, beyond simple alignment. Ligands and even response elements for these receptors have been found. But this is not true for all the nuclear receptors. Even within the extensively studied group of 48 nuclear receptors found in humans, there are still at least 10 ‘orphans’—meaning these receptors don’t yet have a known ligand. Without a ligand, it is very difficult to study a receptor, since ligands are what activate it, and activation is what generates the receptor’s effects.

Sometimes the nuclear receptors are simply divided into three groups: type I, type II, and orphans. So, as was the case with phylogenetic superfamilies, there are also actually several ways to “type” the nuclear hormone receptor superfamily. In other words, within both of these systems for classifying the nuclear receptors there are a variety of actual classifications on offer.

Sub-Section 4.5: The NHRSF by Expression Pattern

There is one final method of organizing the superfamily. It is the youngest of the three classification systems, and so far just one actual classification has been developed within this system, which takes (intercellular) locational and physiological effects into consideration. This way of organizing the nuclear hormone receptor superfamily separates all of the nuclear receptors by expression pattern³⁸ into certain physiologically linked suites of cell types. It is designed to detect and display the relationships between

³⁸ As in, by tracking the expression of the genes that code for these proteins, the nuclear receptors.

particular nuclear receptors (like the androgen receptor) and certain spheres of physiological influence (such as steroidogenesis). So, this way of organizing the superfamily tracks the physiological effects of nuclear receptor action by nuclear receptor location in various cell types along with the typical functional roles of these cell types—or, how and where the nuclear receptors variously manage the assorted tasks involved in the creation and maintenance of physiology.

In this case the units into which the superfamily is divided are called ‘clusters’. Here each nuclear receptor is grouped with others expressed in a similar suite of cell types:

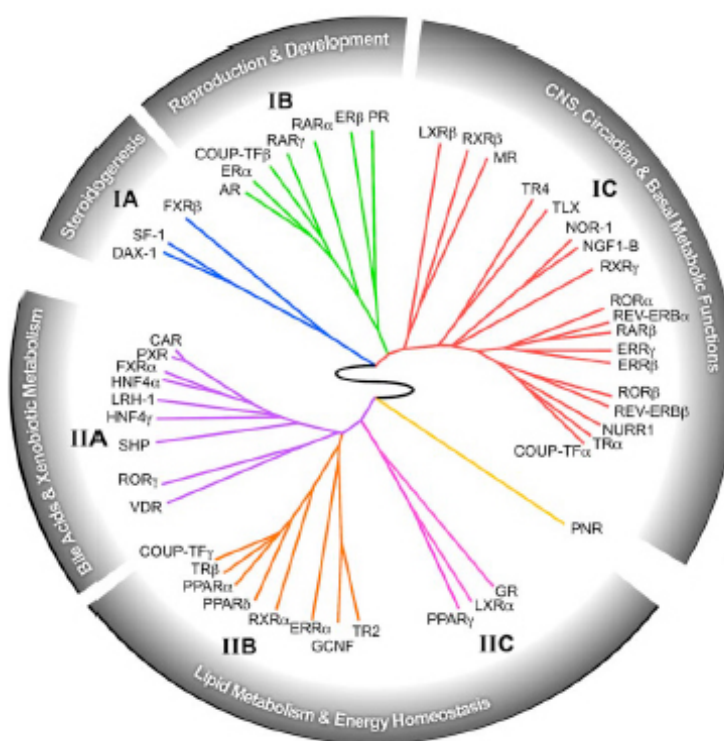


Figure 2.9: A physiologic nuclear hormone receptor superfamily (altered from Bookout et al 2006).³⁹ [Color figure available online.]

³⁹ This depiction is actually of the mouse *and* human nuclear hormone receptor superfamily, since it includes all 49 receptors found in mice—the 48 also produced in humans plus FXRβ.

This produces, first, two (very general) clusters or “paradigms”: (I) reproduction, development, and growth; and (II) nutrient uptake, metabolism, and excretion (Bookout et al. 2006). These two are then each divided into three for a total of six clusters within the superfamily and one lone receptor, PNR, which is only expressed in one type of cell⁴⁰ and thus cannot be part of a suite of cell types. The clusters are labeled by the physiologic pathways that they affect: steroidogenesis (IA), reproduction and development (IB); central nervous system with circadian and basal metabolism (IC); bile acids and xenobiotic metabolism (IIA); and lipid metabolism energy homeostasis (IIB and IIC).

Sometimes this particular classification system is called the ‘Ring of Physiology’ in nuclear receptor research (Bookout et al., 2006, p. 796), because it purports to provide information about the sphere of physiological influence upon which each nuclear receptor acts.

Sub-Section 4.6: Implications

Before moving on to the conclusion of the piece, I’d like to finish out this section by discussing the implications of the nuclear hormone receptor superfamily’s diverse set of classification systems. Each of the three ways of organizing the nuclear hormone receptor superfamily documented here diverges from one another along at least three dimensions: (1) extension; (2) taxonomic structure; and (3) primary considerations. These three facts about classificatory divergence combine to generate the chief conceptual point of the chapter, which is reiterated in the next and final section.

First: extension. Each of the classification systems generates different actual

⁴⁰ At least, so far expression has only been detected in one type of cell.

classifications whereby the nuclear receptors are grouped within the superfamily in different ways. To explain by way of example: one of the nuclear receptors included in the figure above is Retinoic Acid Receptor alpha ($\text{RAR}\alpha$). $\text{RAR}\alpha$ is involved in retinoid signaling; disruption of the gene sequence that codes for the $\text{RAR}\alpha$ receptor correlates with almost 100% of Acute Promyelocytic Leukemia (APL) cases. APL is, if untreated, a deadly form of leukemia—a cancer of the blood and bone marrow. Here is a rendering of $\text{RAR}\alpha$:

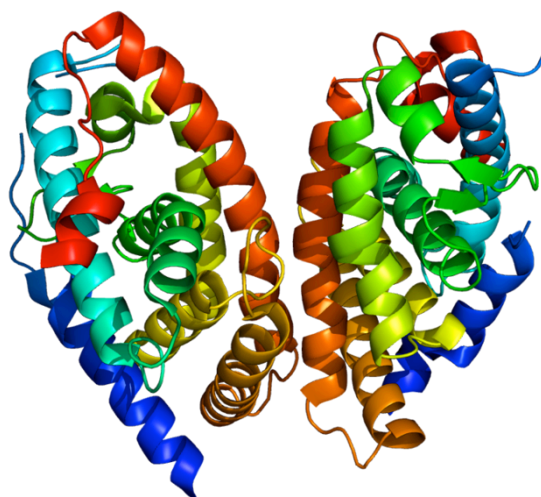


Figure 2.10: The PDB graphic $\text{RAR}\alpha$ (altered from http://en.wikipedia.org/wiki/File:Protein_RARA_PDB_1dkf.png).

All three of the classification systems discussed in this section place $\text{RAR}\alpha$ in distinct collectives with new neighbors. According to the ‘Ring of Physiology’, $\text{RAR}\alpha$ belongs in the reproduction and development cluster IB along with the steroid receptors (estrogen, progesterone, and androgen). But when typing the superfamily, $\text{RAR}\alpha$ is a type II molecule—it resides in the nucleus, forms heterodimers, and binds to inverted repeat response elements—while the steroid receptors are type I—they reside in the cytoplasm and form homodimers. Finally, according to the phylogeny of the nuclear hormone

receptor superfamily shown in figure 2.7, the most closely related sequences to the RARs are the NR4s or nerve-growth-factor-like nuclear receptors, a set of type III or orphan receptors—they reside in the cytoplasm and bind to direct repeat response elements—whose assigned functional cluster in the Ring of Physiology is IC (central nervous system along with circadian and basal metabolism).

Moving on to the second dimension of difference: taxonomic structure. Each of the three classification systems has a slightly different taxonomic structure that follows from differences in extension like the one detailed above. These structures can be compared as follows:

FAMILY	TYPE	PARADIGM
SUBFAMILY	PARALOG	CLUSTER
PARALOG	ORTHOLOG	ORTHOLOG
ORTHOLOG	PROTEIN (TYPE)	PROTEIN (TYPE)
<u>PROTEIN (TYPE)</u>	<u>ISOFORM</u>	<u>ISOFORM</u>
ISOFORM	ALLELOFORM	ALLELOFORM
ALLELOFORM	PROTEIN (TOKEN)	PROTEIN (TOKEN)
PROTEIN (TOKEN)		

Figure 2.11: From left to right, the taxonomic structures of the nuclear hormone receptor superfamily by phylogeny, typing, and expression pattern.

As this comparison shows, the individuation of proteins is not at issue here (and the taxonomic ranks below the line in the figure above simply move intact with the individuated protein types of which they are a part). In fact (as the figure shows), the individuated protein types—the species-specific nuclear receptors—even stick together, throughout all three systems, with their orthologous counterparts in other species. But paralogs do not necessarily travel through the various classificatory permutations.

In other words, there is widespread agreement about the individuation of proteins (into species-specific types), as well as how to track the variation of tokens within these

protein types (into alleloforms and isoforms). But there is quite a bit of variation among the higher taxa of these structures. Even the paralog taxon falls out of one of the classification systems (by expression pattern), because organizing the nuclear receptors by expression pattern breaks up the groups of paralogs. In other words, paralogous nuclear receptors do not necessarily have similar expression patterns or physiological implications.

Shared structure resumes at the ortholog taxon. In other words, all of the classification systems group orthologs, species-specific proteins, isoforms and particulars together.⁴¹ Philosophers might be tempted to label these taxa as the “real” ones, since they are robust across all three classification systems. But that would ignore the fact that the taxa at higher levels are also based on “real” properties of the proteins being classified. It’s just that selecting different bundles of properties—from among the many

⁴¹ In this subsection I’m focusing on three dimensions of divergence that combine to generate the overall conclusion of the piece. But there is a fourth point worth discussing here, about ambiguities of reference in protein science. I’ve already mentioned the use of vague terms specifically in nuclear receptor research (in subsection 4.2). But there’s also a more general point worth making, and it’s clearly indicated by the figure above: all three taxonomies depicted above include the term ‘protein’ at two different taxonomic levels. This is because protein scientists use the term ‘protein’ to refer to both individual molecules and to groups of molecules.

This referential ambiguity can be confusing—it’s as though biologists denoted the entities at both taxonomic ranks of organisms and species with just one term, either ‘organism’ or ‘species’. As in, “wow, look at that tiger on the prowl! What an organism” but also “we must save the tiger organism from extinction.” Or “we must save the blue whale species from extinction” but also “blue whales are the largest species on Earth.”

Another way of putting the point is that when protein scientists say something like “X is a protein,” they might be talking about a token protein molecule (“look under the microscope—there’s a protein!”), or about a certain group of token protein molecules found in a certain species (“RAR α is a protein that can cause APL in humans”). And in fact the situation is even worse than that. Again, as I’ve already mentioned (in a footnote in section 2), a group of orthologous proteins can also be referred to as ‘a protein’. As in, “titin is the largest currently known protein.”

Probably the best way to think about the term ‘protein’ here is rather like the term ‘animal’. Individual organisms, species, and even broader groups are all “animals.” For example, a monkey is a kind of animal, and ‘animal’ here can mean: a particular organism; a group of organisms that form a species; or even a group of species, spanning many genera, all of which are examples of the kind of animal called ‘monkey’.

There is more that could be said about ambiguities of reference and vague terms here, but that will have to be a topic for another day.

properties that these complex macromolecules have—sorts the proteins in different ways.

This brings me to the third dimension of difference: primary considerations. The different ways in which the nuclear receptors are sorted in these different classification systems reflects different sets of considerations: spanning the evolutionary, functional, microstructural, superstructural, mechanistic, locational, and physiological properties of proteins. Differences among the primary considerations within each of the various classification systems correlate with different sets of properties, the powers conveyed by those sets of properties, and the affordances granted researchers by tracking these disparate powers.

Emphasizing evolutionary, functional, and microstructural properties produces a classification, using sequence alignment, that purports to document patterns of shared history or common descent among the nuclear receptors. I'll call these the 'relational powers' of the nuclear receptors. Emphasizing superstructural, mechanistic, and (intracellular) locational properties produces a classification, using various detection techniques, that attempts to establish the binding and activation patterns of the receptors. I dub these their 'chemical powers'. Emphasizing (intercellular) locational and physiological properties produces a classification, using expression pattern, that strives to detail the various contributions of nuclear receptors to the overall development and maintenance of the organism—in other words, their 'biological powers'.

Emphasizing any of these sets of properties and powers over other sets when examining or comparing nuclear receptors grants different affordances to researchers in the field—each of which affordances are incredibly useful. Primarily tracking the evolutionary, functional, and microstructural properties of nuclear receptors—their

relational powers—in order to generate phylogenies of the nuclear hormone receptor superfamily affords certain similarity inferences and suggests particular comparisons among related nuclear receptors. Primarily tracking superstructural, mechanistic, and (intracellular) locational properties to type the nuclear receptors—by chemical powers—affords differential access to nuclear receptors as, for example, high-affinity⁴² or low-affinity⁴³ targets. Primarily tracking (intercellular) location and physiological properties in expression pattern—or, biological powers—affords inferences as to the intersection of particular nuclear receptors with various phenomena of interest in both physiology and pathology.

Each of these affordances—to ancestry, targetability, and applicability, say—granted by tracking different powers—relational, chemical, and biological, respectively—are crucial to the successful design and pursuit of questions and answers in the conducting of nuclear receptor research. And the classification systems, which grant these affordances by tracking different sets of properties and powers, cannot be collapsed into one another. The classification systems do not co-refer to the same taxonomic arrangements—they have both different intension and extension; or, the systems group the nuclear receptors differently. The different affordances granted by the different classification systems arise because the different classification systems track different sets of properties and powers of the proteins; the classificatory divergence is ontologically derived, but epistemically useful.

⁴² ‘High-affinity’ describes a very strong relationship between a ligand and nuclear receptor.

⁴³ ‘Low-affinity’ describes a weak relationship between a ligand and nuclear receptor.

Section 5: Conclusion

In *The Disorder of Things* (1993), John Dupré argues that differences among the projects and purposes of different scientific fields, or at different levels of scientific inquiry, generate different classification schemes. But this discussion of nuclear receptors exposes the use of different classification systems within the same scientific field. In the study of nuclear receptors, the same scientists in the same labs within the same field classify nuclear receptors in a variety of ways, and use these various classification systems interchangeably and even simultaneously.

As has been shown, this classificatory pluralism arises from the complexity of the nuclear receptors and from the different affordances granted by consideration of different sets of properties and powers with respect to the proteins being classified. Nuclear receptors have lots of properties: such as component, microstructural, macrostructural, mechanistic, locational, functional, physiological, and evolutionary ones. Different sets of properties correlate with different relational, chemical, and biological powers. Tracking these different powers grants researchers different affordances to ancestry, targetability, and applicability.

The classifications cannot be collapsed into one another because, among the many properties of each nuclear receptor, different properties cluster together in different ways in order to generate different powers relevant to different affordances. The distinction between properties and powers is required because different properties can also cluster together in order to generate the same power.⁴⁴ And no one classification scheme can

⁴⁴ For example, among DNA binding proteins there are a variety of structures (Helix-turn-helix, Helix-loop-helix, Zinc finger, Leucine zipper, etc.) with which proteins bind to DNA. So this one power (DNA binding) comes, in different proteins, from different structural properties. (More on the implications of this

replace the others because each affordance tracked by a different classification scheme is crucial to the ability of scientists to apply, conduct, develop, direct, fund, and make progress in their research.

Critically, nuclear receptors are not the only complex proteins. As detailed in section 2, proteins in general are a complex mix of component, microstructural, macrostructural, mechanistic, locational, functional, physiological, and evolutionary properties. These properties combine in different ways to generate different powers. Throughout protein science, different powers are tracked via different classification systems in order to grant different affordances. Because these affordances are each so useful, there's no monist unification looming on the pluralist horizon of protein classification.

I'll finish with one final point: it would be a mistake to infer total conventionalism from this intractable pluralism. In other words, the diverse systems of protein classification are only partly conventional. The affordances being granted are of relative scientific interest, certainly. But these affordances are granted via the tracking of various sets of properties and powers of the proteins being classified. These sets of properties and powers are features of the proteins themselves. In terms of the pluralist horizon: different constellations of properties form different shapes, but everybody's looking at the same stars in the protein-filled sky.

In less fanciful terms, and returning to the lead quote from Mill's *A System of Logic* (1843): different systems for classifying proteins do, in a sense, depend on the

choice of the protein scientist—but it's not an arbitrary one.⁴⁵ There is a nuanced exchange here, between convention and nature, when it comes to the generation of these diverse but grounded classification systems in protein science. This exchange is possible because of the complexity of proteins. Complexity admits of different characterizations and resulting classifications in science. There is a non-arbitrary conventionalism—a sort of selective naturalism⁴⁶—to the diverse classification of these complex biochemical kinds.

⁴⁵ For ease of reference, I'll reproduce the quote here: "in so far as a natural classification is grounded on real Kinds, its groups are certainly not conventional; it is perfectly true that they do not depend upon an arbitrary choice of the naturalist" (Chapter VII, Section 4, pg. 503).

⁴⁶ More on this concept in the next chapter.

REFERENCES

- Bookout, A., Jeong, Y., Downes, M., Yu, R. T., Evans, R. M., & Mangelsdorf, D. J. 2006. Anatomical profiling of nuclear receptor expression reveals a heirarchical transcriptional network. *Cell* 126:789–799.
- Dayhoff, M. O. 1976. The origin and evolution of protein superfamilies. *Federation Proceedings* 35: 2132–2138.
- Dayhoff, M. O., Barker, W. C., & Hunt, L. T. 1983. Establishing homologies in protein sequences. *Methods in Enzymology* 91: 524–545.
- Dayhoff, M. O., & Eck, R. V. 1965. *Atlas of protein sequence and structure*. Silver Spring, MD: National Biomedical Research Foundation.
- Dupré, J. 1993. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press
- Evans, R. M., Weinberger, C., Hollenberg, S., & Rosenfeld, M. G. 1986. Structure and expression of human glucocorticoid receptor cDNA: A trans-acting factor related to the c-erbA protooncogene family. *Journal of Molecular and Cellular Biology* 5: 63.
- Goodwin, W. 2011. Structure, function, and protein taxonomy. *Biology & Philosophy* 26: 533–545.
- Mill, J. S. 1843. *A System of Logic, Ratiocinative and Inductive*. The Project Gutenberg Ebook: January 31, 2009.
- Mitchell, S. D. 2009. *Unsimple truths: Science, complexity, and policy*. Chicago: University of Chicago Press.
- Nelson, D. L., & Cox, M. M. 2008. *Lehninger's Principles of biochemistry* (5th ed.). W. H. Freeman: New York.
- Olby, R. 1974. *The path to the double helix: The discovery of DNA*. New York: Dover.
- Rother, M., & Krzycki, J. A. 2010. Selenocysteine, pyrrolysine, and the unique energy metabolism of methanogenic archaea.” *Archaea* 2010: 1–14.
- Slater, M. H. 2009. Macromolecular pluralism. *Philosophy of Science* 76: 851–863.
- Strasser, B. J. 2009. Collecting, comparing, and computing sequences: The making of Margaret O. Dayhoff's *Atlas of protein sequence and structure*, 1954–1965. *Journal of the History of Biology* 43: 623–660.

- Tobin, E. 2010. Microstructuralism and macromolecules: The case of moonlighting proteins. *Foundations of Chemistry* 12: 41–54.
- Weinberger, C., Giguere, V., Hollenberg, S., Rosenfeld, & Evans, R. M. 1986. Human steroid-receptors and erbA protooncogene products: Members of a new superfamily of enhancer binding-proteins. *Cold Spring Harbor Symposia on Quantitative Biology*, 51, 759–772.
- Williams, A. F. 1984. Molecules in the immunoglobulin superfamily. *Immunology Today* 5: 219–221.
- Wu, W., Niles, E. G., Hirai, H., & LoVerde, P. T. 2007. Evolution of a novel subfamily of nuclear receptors with members that each contain two DNA binding domains. *BMC Evolutionary Biology* 7:27.
- Zhang, Z. D., Cayting, P., Weinstock, G., & Gerstein, M. 2008. Analysis of nuclear receptor pseudogenes in vertebrates: How the silent tell their stories. *Molecular Biology and Evolution* 25: 131–143.

CHAPTER 3: COMPLEX OBJECTS

It is as if the five blind men of the legend not only perceived different aspects of the elephant, but, conscious of the tremendous difficulties of reconciling their views of the same object, decided to treat their views as if they were of *different* objects.

Wimsatt's *Complexity and Organization* (1974)

Abstract: The first chapter of the dissertation looked at, among other things, the classificatory individuation of proteins—and declared it monist. The second chapter looked at the classificatory organization of proteins—and pronounced it pluralist. This third chapter seeks an explanation for this particular kind of diverse yet grounded, enduring classificatory pluralism—and finds it in complexity. When complex objects are being classified, even scientifically, both a variety of possible classification schemes and a plurality of actual ones can arise. These diverse classificatory possibilities arise because complex objects have myriad properties, which combine to generate various powers, and researchers have incentive to generate any and all those classification systems that grant them reliable inferences. The inferences are to affordances, which are granted by tracking relations among those properties or powers that regularly generate the affordances. The process by which combinations of properties, powers, and affordances interact with scientific research to generate diverse yet grounded classification systems is called ‘selective naturalism’.

Section 1: Introduction

Part of the previous chapter of the dissertation focused on a group of gene regulatory proteins called the nuclear hormone receptor superfamily. In that particular

segment,¹ I discussed various ways of scientifically classifying the 49 nuclear receptors found in mice (48 of which are also found in humans). I showed that, within the study of these nuclear receptors, there are three main systems for classifying them: one each based on phylogeny, typology, and expression pattern. At the close of the discussion, I made three observations about these distinct classification systems: first, within each system the nuclear receptors are organized differently relative to one another; second, the organization of the systems is distinct enough to include different taxonomic levels and numbers of levels; and third, each system is based on a different primary consideration (ancestry, targetability, or applicability).

In this chapter I offer an underlying explanation for all three of these observations. To put the main point quite succinctly: these observed features of scientific classification are all derivative on the complexity of the objects being classified. And I will illustrate, by example and throughout this chapter, precisely what makes an object complex in this sense.² Most importantly, when complex objects are being classified, even scientifically, both a variety of possible classification schemes and a plurality of actual ones can arise. These diverse classificatory possibilities occur because complex objects have myriad properties, which combine to generate various powers, and researchers have incentive to generate any and all those classification systems that grant them reliable inferences. The inferences are to affordances, which are granted by

¹ Section 4 of chapter 2.

² Hint: the term will be used to mean something simple and deflationary, like “has lots of interacting parts.” I understand that there are alternative definitions of ‘complex’ available for use here. For example, from dynamical systems theory: complex systems are those that exhibit non-linear behavior. The problem with employing this or any other more precise but domain-specific sense of the term is that it restricts the scope of the ensuing discussion. I want to use an extremely liberal sense of ‘complex’ so that my corresponding account of the effect of complexity on classification will be widely applicable.

tracking relations among those properties or powers that regularly generate the affordances.

In the field of nuclear receptors, for example, researchers are especially interested in knowing three particular things about each nuclear receptor: one, its evolutionary history (what other receptors it's related to); two, its accessibility (whether and how it is targetable by natural ligands or designer drugs); and three, its involvement in particular physiological phenomena (its relevance to disease and/or enhancement). I call these things affordances, because knowing about the evolutionary history or ancestry, the ligand or drug targetability, and the physiological applicability of any particular nuclear receptor allows researchers to infer and apply their research in valuable ways. Knowledge of these things *affords* research. So, nuclear receptors are classified in ways that group together those receptors that are similar to one another in ways relevant to the ability to make inferences to those affordances.

In the case of the nuclear hormone receptor superfamily, the inferences to affordances being granted by the different classification systems are as follows: by phylogeny, to ancestry; by typology, to targetability; and by expression pattern, to applicability. Inferences to these affordances are granted by classification through the organization, within the systems, of information regarding which of the various instances and combinations of powers and properties tend to make a nuclear receptor evolutionarily-relatable, ligand-targetable, or disease-applicable.³

³ Just a reminder: phylogenies of the superfamily track evolutionary, functional, and microstructural properties—loosely, their relational powers—and the tracking of these properties and powers is what grants reliable inferences to ancestry. Typologies of the superfamily track superstructural, mechanistic, and (intracellular) locational properties—what I've called their chemical powers—and the tracking of these properties and powers is what grants reliable inferences to targetability. Finally, the classification by

Researchers have incentive to track ancestry because it allows for homology inferences to be made; targetability because it allows for ligands to be discovered or designed, and experiments to be run; and applicability because it helps to apply and market the research, among many other things. Each of these investigative abilities are valuable to research in the field of nuclear receptors; and from certain perspectives on the research, some of these affordances are more crucial than others.

Now, I'll say much more about these ideas throughout the chapter—specifically, I'll talk about complexity in detail throughout section 2; I'll discuss properties and powers in section 3; and I'll explain perspectives and affordances in section 4. In what remains of this introductory section, however, I'd like to briefly situate this discussion within the broader philosophical debate.

The subject of my case study, nuclear receptors, is a new one for the philosophy of science. And it is a case in which a pluralist variety of systems for the classification of the complex objects under study is not just possible, but actual. But I am by no means the first philosopher to discuss a case of classificatory diversity and/or pluralism. So let me state what I see as my contribution relative to others located somewhat nearby in the conceptual field.

As I mentioned at the close of the last chapter, in *The Disorder of Things* (1993) John Dupré argues that differences among the projects and purposes of different scientific fields, or at different levels of scientific inquiry, generate different classification

expression pattern tracks (intercellular) locational and physiological properties—or, the biological powers of the nuclear receptors—and the tracking of these properties and powers is what grants reliable inferences to applicability.

schemes.⁴ But my discussion of nuclear receptors exposes the use of different classification systems within the same scientific field. In the course of studying the nuclear receptors, the same scientists in the same labs within the same field classify these receptors in a variety of ways, and use these various classification systems interchangeably and even simultaneously.

So, although Dupré and I are both offering case-based accounts of classificatory pluralism, his focus is chiefly on cases of what I would call *interdisciplinary classificatory pluralism*—cases in which different disciplines use different classification systems for the same objects—whereas my focus is on *intradisciplinary classificatory pluralism*—the use of different classifications within a discipline. One of Dupré’s classic cases comes from a particular family of plants (Liliaceae). Chefs and botanists classify the plants in this family very differently. The botanist has many divisions within the family—into genera and species, for example; but the chef will principally divide these plants into edible and non-edible—serving some members of the family (onions and garlic) but not others (it’s rare to find a cooked lily or a tulip).

So there is interdisciplinary classificatory pluralism when it comes to this group of plants; there is (at the very least) both a culinary and a botanical classification system, each residing within their own respective domains.⁵ In contrast, my case study of nuclear receptors documents intradisciplinary classificatory pluralism: there are phylogenies,

⁴ See also his more recent “In Defense of Classification” (2001).

⁵ It is also worth noting that the three things I noticed about the distinct classification systems of nuclear receptors—summarized in the last sentence of the first paragraph of this chapter—are also true of these culinary and botanical classification systems for herbs. As in, they have different relative organization, taxonomic levels, and primary considerations.

typologies, and a “ring of physiology”⁶ all at use within one particular sub-field, which exists at a nexus of protein science, molecular biology, and gene expression.

So, although Dupré and I each examine the issue of classificatory pluralism, his cases tend to be interdisciplinary whereas I have focused on the intradisciplinary. Two other philosophers who have also looked at intradisciplinary classificatory pluralism are P. Kyle Stanford and Philip Kitcher. In their “Refining the Causal Theory of Reference for Natural Kind Terms” (2000), Stanford and Kitcher discuss two cases in detail: acids (in chemistry) and species (in biology).

Like many other chemical kinds, acids were initially characterized macroscopically, by their chemical behavior and reactivity. As Stanford and Kitcher explain it: “Boyle first characterized the members of the natural kind *acid* phenomenologically, by their sour taste, corrosiveness, their ability to precipitate sulfur from sulfide solutions and to redden blue plant dyes, and by the loss of these properties on contact with alkalies (bases)” (2000, p. 115). But as the theories, techniques, and tools of chemistry advance, chemists often develop microstructural enhancements or even alternatives to the familiar macroscopic accounts.

Almost two hundred years after Boyle died, Arrhenius offered the first alternative, microstructural account of acids. This was the first account of its kind, but not the last, because it turns out that acids as a group fail to share one unique microstructural property that can be used as a membership condition for the group.⁷ So now there are also

⁶ This is a nickname for the expression pattern classification system of the nuclear receptors, devised by those scientists who first proposed to classify the nuclear receptors in this way (Bookout et. al 2006).

⁷ This is somewhat analogous to the situation, famously explained by Putnam (1975), of jadeite and nephrite. As it turns out, jade as a group lacks a unifying microstructural property with which membership in the group is conferred. The group actually consists of two different minerals with distinct elemental

Brønsted-Lowry acids⁸ and Lewis acids,⁹ in addition to Arrhenius acids.¹⁰ Stanford and Kitcher explain the situation thusly: “when chemists used stereotypical features to point to underlying structures that they were unable to specify, Boyle and his followers were partially referring to Arrhenius acids, partially referring to Brønsted-Lowry acids, and partially referring to Lewis acids” (Stanford & Kitcher 2000, p. 119–20).

As this quote demonstrates, however, the puzzle posed by the case of acids is primarily one of reference—of how the referent of natural kind terms is fixed. In other words, this is a matter of *individuation* as opposed to *organization*. Recall the distinction between these two separate classificatory activities.¹¹ Questions about classificatory individuation are questions about how entities are picked out and how kind terms refer. But questions about classificatory organization are questions about how select entities are related to one another and how similarity judgments and inferences are made.

The case of acids is one of individuation—the question in that case is about how the term ‘acid’ refers, and the answer is plural, since the term can refer in any one of three different ways. The case of acids is one of *intradisciplinary classificatory (individuating) pluralism*. In contrast, the case study under discussion here, of nuclear receptors, is one of *intradisciplinary classificatory (organizing) pluralism*. Although nuclear receptors are monistically individuated,¹² there is a question about how to relate them to one another, and again the answer is plural, since there are three major systems

composition and microstructure.

⁸ The Brønsted-Lowry definition (circa 1923) defines acids as compounds able to lose or donate a proton.

⁹ The Lewis definition (circa 1938) defines acids as electron pair acceptors.

¹⁰ The Arrhenius definition (circa 1887) defines acids as hydrogen compounds dissolved in aqueous solution.

¹¹ Introduced in chapter 2—briefly, the former relates to category membership and the latter to relations between categories.

¹² Like proteins in general—see chapter 1, section 3.

for doing so (the phylogenetic, the typological, and the expression-patterned).

Interestingly, the other case that Stanford and Kitcher discuss—that of species—is one that comes quite a bit closer to being a case of intradisciplinary classificatory (organizing) pluralism. But the authors' discussion of the case focuses primarily on the question of what the kind term 'species' refers to—again, on intradisciplinary classificatory (individuating) pluralism. Once more, the answer is a pluralistic one, since the term can be used to refer to, among other options, a reproductive concept,¹³ an ecological concept,¹⁴ a phylogenetic concept,¹⁵ or a cohesion concept.¹⁶

Stanford and Kitcher do, however, mention the implications of this case for intradisciplinary classificatory (organizing) pluralism in passing. They write:

Now the explanatory projects of various fields within biology are dependent not simply on classifications of organisms at the species level: at least in some cases, one wants to identify broader kinds and to use notions drawn from a hierarchy of divisions. This means that the judgments about what organisms count as significantly similar, the critical scrutiny and refinement of talk of “the same type”, is not just responsive to the work that the division can do in explaining properties of the organisms falling under a species concept (the kind of processes to which we alluded in discussing the modification of Boyle's original list of stereotypical properties [of acids]) but also to broader systematic considerations. (Stanford & Kitcher 2000, p. 123)

But as the quote shows, their interest in classificatory organization—in a “hierarchy of divisions,” as they put it—quickly turns back into a focus on individuation—simply, on how other terms for “broader kinds” at higher taxonomic levels, such as genus and phylum, might refer. And that's all they say on the subject of taxonomy, turning their

¹³ According to which members of a species are defined in terms of interbreeding (i.e., Mayr 1957).

¹⁴ According to which members of a species are defined in terms of ecological niche (i.e., Van Valen 1976).

¹⁵ According to which members of a species are defined in terms of historical connectedness and individuating characteristics (i.e., Cracraft 1983).

¹⁶ According to which members of a species are defined in terms of demographic and genetic cohesion (i.e., Templeton 1989).

attention quickly back to questions of reference. Stanford and Kitcher never return to the issue of intradisciplinary classificatory (organizing) pluralism.¹⁷

In what remains of this chapter, I'm going to stay focused on multiple, diverse taxonomies of monistically individuated objects. In particular, I want to explore the question of how this sort of classificatory pluralism might arise even within thoroughly intradisciplinary and rigorously scientific endeavors. As I mentioned earlier in this introduction, I think that the answer is complexity.

In the next section I'll look at two influential—one old and one new—philosophical accounts of complexity. Then I'll show, in section 3, how one of the accounts, designed primarily for complex systems, can be extended to complex objects. Finally, in section 4, I'll demonstrate how this account of complex objects can be applied to proteins in general, and nuclear receptors in particular, in order to explain the generation of diverse yet grounded classification systems in protein science. The conclusion summarizes the discussion as well as explores some of its implications.

Section 2: Complexity

Perhaps the definitive treatment of complexity—in the philosophy of biology, at least—is William C. Wimsatt's "Complexity and Organization" (1974). In typical Wimsatt style, the paper does not contain a surplus of direct exposition, designed to precisely explain and define the concept of complexity. Instead, Wimsatt exposes what complexity is by discussing where it's hidden—by explaining how and what

¹⁷ Though Kitcher also discusses (intradisciplinary classificatory) individuating pluralism in his *Science, Truth, and Democracy* (2001), again there is little pursuit of the topic of organizational pluralism.

simplification conceals.

Wimsatt calls our tendencies to simplify ‘the reductionist orientation’. He writes early in the paper that “a number of features of the reductionist orientation contribute to a point of view which is ill-suited to an adequate treatment of the concept of complexity” (p. 67). Wimsatt then examines three such features of the reductionist orientation: (1) monistic bias; (2) fractured, isolated analysis; and (3) divorced perspectives in need of reconciliation. According to Wimsatt, these features impede or obscure our attempts to address and understand complexity. In other words, our reductionist orientation has encouraged us to emphasize monism and discount pluralism, and to adopt a divide-and-conquer approach to complex phenomena, which then requires a complex reunification of perspectives that we generally fail to pursue.

On this indirect account, complexity itself is something that: (1) may call for heuristic pluralism rather than monistic bias; (2) includes multiple components of analysis—each of which may be isolable or at least localizable; but which (3) then requires reconciliation in order for understanding. Renouncing the reductionist orientation and embracing complexity requires considering potential pluralism, reconfiguring deconstructed phenomena, and reunifying isolated perspectives.

We often fail to pursue reunification, despite its obvious potential for enhanced understanding, because it can seem like a fiendishly difficult—perhaps impossible—task to accomplish. Wimsatt makes this tension explicit in his own approach to the topic of complexity. The desire is there: “it thus would be profitable to see how we tend to relate our different views or theoretical perspectives of objects and in particular of complicated objects” (Wimsatt, 1974, p. 69). But so is the difficulty: “this would be an enormous task

for even two views” (ibid.). That we ought to confront complexity is easier said than done.

And so, nearly four decades later, the call to acknowledge and embrace complexity is still being sounded. In her recent book, *Unsimple Truths: Science, Complexity, and Policy* (2009), Sandra D. Mitchell writes that “the world is indeed complex; so, too, should be our representations and analyses of it. Yet science has traditionally sought to reduce the ‘blooming, buzzing confusion’ to simple, universal, timeless underlying laws to explain what there is and how it behaves” (p. 11). The remarkable successes of strategies of simplification in certain areas has encouraged scientists to emphasize these strategies; to focus on those phenomena that are most amenable to such tactics; perhaps even to mistakenly characterize all phenomena of interest as simplifiable in these ways.

According to Mitchell, philosophers of science can be just as guilty of this kind of oversimplification as scientists. The way she tells it, nineteenth-century philosophers of science like John Hershel, William Whewell, and John Stuart Mill based their accounts of scientific knowledge on the astoundingly successful work of seventeenth century scientists like Isaac Newton and Johannes Kepler. Newton’s laws of motion and gravitation, for example, offered a sweeping and powerful way to unify, explain, and predict an incredible range of phenomena. And this set the philosophical standard for scientific knowledge.

Mitchell does not want to replace this characterization of science as in the business of discovering universal, exceptionless generalizations that furnish unified, underlying explanations. She wants to supplement it. In her own words: “it is not the

case that the traditional framework always fails; there remain stunning successes since the time of Newton to testify to its power. The problem is that much of the world escapes its concepts and methods” (2009, 12). One might ask: what makes most of the world a fugitive from traditionally reductivist philosophy of science? Mitchell’s answer is: complexity.

Of course, it is not as though the subject of complexity has been completely ignored in the years between Wimsatt’s initial paper (1974) and Mitchell’s recent book (2009). Many scientists and philosophers have made progress in the study of complex phenomena and complexity *simpliciter* (with the scientists tending to concentrate on the phenomena and the philosophers inclined toward the concept, unsurprisingly).

Studies in the science of complexity can be found in the fields of systems biology, evo-devo, developmental systems theory, and complexity theory, for example. Philosophical accounts of instances and conceptions of complexity can also be found—for example, in William Bechtel and Richard C. Richardson’s *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research* (1993). Nancy Cartwright’s *How the Laws of Physics Lie* (1983) is a study of the difficulty deriving tidy explanations and laws from complex systems. This subject gets further treatment in her *The Dappled World: A Study of the Boundaries of Science* (1999).

But these are still just exceptions to the general reductionist rule; according to Wimsatt and Mitchell, we must continually fight to overcome our tendencies to oversimplify. The reductionist orientation is still the most popular stance, bristling with monistic bias, fractured analysis, and isolated perspectives. Particularly in those areas rife with complexity—evidenced by heuristic pluralism, deconstructed multicomponent

phenomena, and perspectives in need of reconciliation—we must be on the lookout for oversimplification.

Regardless of whether the reductionist orientation is still the dominant one, proteins are fiendishly complex molecules, and I stipulate that we ought to be wary of oversimplification in the course of their study. The subject of proteins is a difficult one, and newly of interest to philosophers. As I showed in an earlier chapter,¹⁸ initial contributions¹⁹ have tended to present only partial views of their subject—just as Wimsatt said, they engage with only one part of the elephant.²⁰

In this dissertation I have been attempting—in so far as the limited understanding provided by current science allows—to engage with, as much as possible, the full complexity—in Wimsatt’s (1974) sense—of proteins. And as it turns out, proteins display all of the features of Wimsatt’s indirect account: they are many different things to many different people (difficult to characterize monistically); they have an incredible number of different parts and properties, sometime isolable, sometimes localizable (multiple components); and the science of proteins is highly compartmentalized (divorced perspectives)—mostly in order to try and manage them.

But it is Mitchell’s (2009) account of complexity that can be adjusted and applied

¹⁸ Sub-section 3.4 of chapter 1.

¹⁹ See Slater (2009), Tobin (2010), and Goodwin (2011).

²⁰ Recall the opening quote. In it, Wimsatt refers to the Indian parable of the blind men and the elephant. There are many versions of this story in Indian tradition and lore. But the general story goes something like this: a group of blind men (usually five or six of them) come upon an elephant. Each one encounters a different part of the elephant and describes it differently. The tail is like a brush; the leg is like a pillar; the ear is like a fan; the tusk is like a pipe; the trunk is like a tree branch. The blind men squabble over what it is that they have encountered, and each insists that his description is the right one. Usually someone else must explain to them (often a king) that they are all right in some sense—the elephant is like all of the things they have described. It is one thing and they must put together their different views, in order to “see” the whole elephant. Wimsatt’s point is simply that so must we, scientists and philosophers, attempt to put together our partial views of the complex objects and phenomena that we study. To fail to do this is to remain as one of the blind men clinging to only one part of the elephant.

to proteins in a way that helps to explain the intradisciplinary classificatory (organizing) pluralism that has arisen during their scientific investigation. So it's to that account that I will now turn.

Section 3: Extending Mitchell's View

The first thing to note about Mitchell's (2009) account of complexity is that she calls a lot of different things complex. She discusses complex systems more than anything else, but she also talks about the complex world as well as complex structures, behaviors, phenomena, networks, processes, sciences, organization, interactions, sets, patterns, relationships, situations, motions, disorders, and ecosystems, properties, sets of states, and structures. All in all, she calls more than 40 different things complex in *Unsimple Truths*—but she never calls objects complex. (More on this in a bit.)

Mitchell also identifies different species of complexity—such as biological, chemical, neurological, evolved, genetic, and natural. And she provides many different examples of complex systems, behaviors, disorders, etc. Her examples of complex things include multicellular organisms, social insect colonies, eukaryotic slime mold, bird flocks, tumbling *E. coli* bacteria, global climate, and genetically modified agriculture, in addition to major depressive disorder.

Finally, Mitchell also develops what she calls a “taxonomy of complexity.” This is a list of those features which often characterize complex biological systems. In *Unsimple Truths*, Mitchell discusses four kinds of complexity exhibited by complex systems: “multilevel organization, multicomponent causal interactions, plasticity in relation to context variation, and evolved contingency” (2009, 21).

But in an earlier work, *Biological Complexity and Integrative Pluralism* (2003), Mitchell categorizes complexity in three different ways: as constitutive, dynamic, and evolved. And though she eventually moves on from these three kinds of complexity in favor of the other four, the idea of constitutive complexity is particularly relevant for this discussion. This is because it is the closest she comes to the idea of complex objects that I wish to develop here.

Mitchell's initial definition of constitutive complexity is actually just an example: "organisms display complexity of structure, the whole being formed of numerous parts in nonrandom organization" (2003, 4).²¹ When expanding on this idea of constitutive complexity, she provides additional examples, mentioning individual cells in a multicellular organism, collections of species in an ecosystem, and individual insects in a colony. She also writes, "minimally, complex systems can be distinguished from simple objects by having multiple parts that stand in nonsimple relations. That is, there is structure or order in the way in which the whole is composed of the parts" (5). In this explication of the concept of constitutive complexity, Mitchell contrasts complex systems with simple objects.

There are two distinct dimensions to this contrast, between simple and complex, system and object. But Mitchell conflates these dimensions when she associates objects with simplicity and systems with complexity, reducing the number of categories in her ontology from four to two. Her account explicitly includes simple objects and complex systems but overlooks simple systems and complex objects. There's no reason to admit

²¹ Mitchell cites Herbert A. Simon's *The Science of the Artificial* (1981) and William C. Wimsatt's "Forms of Aggregativity" (1986) after providing this account of constitutive complexity.

of only the two categories: objects aren't necessarily simple while systems are necessarily complex. And this conflation is in conflict with her own account, because Mitchell implicitly admits to the existence of simple systems.

To explain: in addition to her account of complexity, Mitchell also develops a general philosophical approach, which she calls 'integrative pluralism'. The basic idea is that there is complexity in the world—especially the biological world. Complex systems display multilevel organization, multicomponent causal interaction, contextual plasticity, and evolved contingency. Understanding and explaining the behavior of these systems often requires developing an account at each of the levels spanned by the system. Generating a comprehensive account of such a complex system means integrating these multiple layers of explanation, hence *integrative pluralism*.

According to Mitchell (2009, chapter 6), this is a moderate sort of pluralism: it is neither limitlessly pluralistic in the “anything goes” tradition²² nor ultimately monist in the “pluralist sciences are immature sciences” tradition²³. I think that her pluralism is itself multilevel: Mitchell is advocating a pluralist approach both to explanations of specific complex systems and to scientific explanation in general, and the former pluralism resides at what might be thought of as a lower level than the latter.

On Mitchell's view there is basic, low-level pluralism with respect to particular scientific explanations. To use an example of hers, explaining the emergence of major depressive disorder requires appealing to genetic as well as environmental factors. A complete account of the disorder must integrate the explanations offered at each of these

²² For a classic example of this position, see Feyerabend (1975).

²³ For examples of this view, see Kuhn (1962/1996) and Kitcher (1981).

levels. But there is also a higher-level pluralism within the view. When Mitchell says “my argument is *not* one for a wholesale replacement of the traditional views but, rather, for an expansion of traditional epistemologies of science to accommodate aspects of knowledge that do not fit the older formulations” (2009, 4), she is advocating a pluralist expansion to the previously monist account of scientific explanation. The idea is not that all explanations are multilevel and pluralist—just that some are. Certain systems (complex ones) will require integrating plural explanations; other systems (simple ones) will not. This tolerance for both monist and pluralist explanations itself amounts to a sort of pluralist philosophy of explanation, or a meta-explanatory pluralism.

In summary, Mitchell’s account promotes, at the lower level, sometimes monism and occasionally pluralism when it comes to actual scientific explanations; and the fact that she embraces two distinct kinds of explanation makes her approach pluralist in a second way, at a higher level. Mitchell is a pluralist in that she thinks understanding complex systems requires integrating multiple explanations; she is also a pluralist in that she thinks explanations can be integrated in this way or not—they can also be monist and non-integrative. So she is both an explanatory pluralist and a meta-explanatory pluralist.

Mitchell uses the existence of complex systems to motivate her doubly pluralist account. In other words, integrative pluralism is required because of complex systems; without it we cannot understand them. More importantly for our purposes here, the presence of both simple and complex systems is part of what necessitates the higher-level explanatory pluralism within Mitchell’s integrative account.

In other words, if all systems were complex, then the reductive accounts of scientific explanation and inference could be wholly dispensed with, rather than added to.

But there are systems, according to Mitchell, that do not display multilevel organization, multicomponent causal interaction, contextual plasticity, and evolved contingency—and which hence are not complex. These, I contend, are the simple systems already embedded in Mitchell's account, and they should be explicitly recognized as part of her ontology.

Additionally, I think that her ontology should also be expanded to explicitly include the fourth category of complex objects. And like simple systems, complex objects might already be there in Mitchell's account, albeit even more implicitly than simple systems. The possibility of complex objects can, I contend, be found among Mitchell's examples, at least some of which have been potentially miscategorized as systems rather than objects.

An object is typically an individual material entity, with properties and dependent parts, itself substantial and independent. A system is typically an integrated whole, with structure and behavior, but put together of interacting and interdependent components. So, although it makes intuitive sense to classify insect colonies, bird flocks, climate, and agriculture as systems, it is less obvious that individual bacterium, single-celled amoebae, and multicellular organisms should be thought of as systems rather than objects.

None of these latter cases—individual bacterium, single-celled amoebae, and multicellular organisms—are obviously or decisively objects rather than systems, however. This shows that Mitchell's account of complex systems is merely compatible with rather than already truly inclusive of complex objects. Mitchell's examples are all obviously systems, or cases that lie on the borderline between systems and objects. None of them are clearly and definitively objects.

Luckily for my argument, nothing really hangs on the status of an amoeba as a system or an object, or even on the broader ontological distinction between systems and objects in and of themselves. I think that Mitchell's ontology should be expanded regardless—to explicitly include complex objects—because thinking explicitly about complex objects, and thinking of them as distinct from complex systems, allows for previously unconsidered possibilities to emerge. In my case, thinking about proteins as complex objects helps to understand their complexity while preserving their commonsense ontology, as objects. Thinking about proteins as complex objects also suggests a new way to explain the disordered state of protein classification: the lack of a comprehensive and clear method for organizing proteins might be a product of their extreme complexity, in combination with the varied interests researchers have in tracking different dimensions of this complexity.

Thinking of proteins as complex objects, and examining how their classification handles this complexity, reveals that classification requires its own integratively pluralist approach when coping with complexity. In other words, comparing and then extending Mitchell's view: thinking about complex systems leads Mitchell to advocate for integrative pluralism of scientific explanations, models, laws, and methods; thinking about complex objects correspondingly necessitates the application of integrative pluralism to scientific classification.

Section 4: Proteins as Complex Objects

But Mitchell doesn't take this step herself. Her emphasis is chiefly on systems,

and entirely off objects.²⁴ As I have said, she never even mentions complex objects (at least not in her 2009 book). Again, I don't think that this is because things are either systems or objects, or because systems are always complex and objects are always simple.

The point, instead, is that when we think about systems we tend to think about complex, dynamic and dispersed interaction, but when we think about objects we tend to think about simple, bounded, and concrete individuality. However: there can be bounded complexity, dynamic simplicity, dispersed objects, and concrete systems. I want to explore these disassociations from the common correlates—because I think that interesting things happen outside the bounds of normality, among atypical associations.

More specifically for the purposes of this discussion, I want to focus on complex objects because I am trying to explain complicated classification. When we think about classification we tend to think of the things we are classifying as discrete, bounded individuals—as simple objects. And if we base our understanding of classification on how classification works with simple objects, we might end up with a characterization of classification that can't really handle complex objects.

And in fact I think that is about where we have ended up. Certainly we have trouble understanding the classification of complex biological kinds.²⁵ My assertion here is that with an account of proteins as complex objects, it will be possible to better

²⁴ There's a potential explanation for why Mitchell avoids talk of complex objects to be found in Wimsatt's classic paper: "the net result [of complexity] is often not to talk about objects at all, but to emphasize predicates, or the systems of predicates grouped together as theories or models. Thus, although biologists, social scientists, and others who work in areas where 'complexity' is a frequent term talk almost invariably of the complexity of systems (thereby meaning the objects, in the full-blooded sense, which they study), most analyses of complexity in the philosophical literature have been concerned with the simplicity or complexity of sets of predicates or of theories involving those predicates" (1974, p. 68).

²⁵ Again, see the species problem.

understand the messy, diverse, and complicated classification of proteins. My supposition is that, perhaps, the account of complex objects in general will help to explain the messy, diverse, and complicated classification of other kinds of chemical and biological things as well.

Despite not having considered any paradigmatic objects in her analysis of complexity, it is obvious that, on Mitchell's account, it is just as possible for complexity to be a feature of objects as of systems. This is because objects have parts. These parts can be organized in a multilevel way, and might causally interact with each other. Objects can also be contextually plastic, and may display evolved contingency. In other words—mostly Mitchell's—"complex [objects] can be distinguished from simple objects by having multiple parts that stand in nonsimple relations" (2003, 5).²⁶

So, I will now use the taxonomy of complexity that Mitchell develops in *Unsimple Truths* to present proteins as complex objects. Mitchell identifies four "kinds" of complexity: (1) multilevel organization; (2) multicomponent causal interaction; (3) contextual plasticity; and (4) evolved contingency. I will show that the first three of these features are all part of what makes individual proteins complex objects; and the fourth feature, evolved contingency, complicates attempts to relate and classify individual proteins with one another.²⁷

First is *multilevel organization*, which occurs when there are different levels of organization within one complex object. For example, a human organism has different

²⁶ This sense of 'complex' dovetails nicely with my primitive one (which was simply something like "having lots of interacting parts").

²⁷ Actually, this fourth feature of proteins and other evolved entities is what requires the addition of a level between properties and affordances in the proposed framework of complex classification. More on that near the close of this chapter.

levels of biological organization: along cell types, organs, organs systems, whole organism. There are also other ways to level the organism: into genotype and phenotype, for example.

But I want to stay focused on proteins. Remember, proteins are polypeptides—macromolecules composed of multiple amino acids linked by peptide and other bonds. Amino acids are tetravalent molecules: they consist of a central or alpha carbon with four groups bonded to the tetravalent center. These four groups always include an amino group,²⁸ a carboxyl group,²⁹ one hydrogen molecule,³⁰ and what is called a side chain or R-group. Different side chains or R-groups make for different amino acids, and there are 20 common ones.

Proteins are not just chains of amino acids, however. As I have said, they have multilevel organization, and amino acid sequence comprises just the first or primary level. Comprehensive protein structure is divided into four different levels: (1) primary structure—amino acid sequence, producing chains; (2) secondary structure—local bond formation between amino acids, producing shapes like turns and sheets; (3) tertiary structure—additional interactions between elements scattered throughout the molecule, completing overall conformation; and (4) quaternary structure—multiple-molecule shape formation, producing complexes.

Now recall the second kind of complexity from Mitchell's taxonomy: multicomponent causal interaction. The multilevel organization of proteins is intricately

²⁸ Standard chemical notation for this group is: N-H

²⁹ Standard chemical notation for this group is: C=O

³⁰ Standard chemical notation for this group is: H

linked to *multipart causal interaction* within proteins.³¹ For example, this is an alpha helix, a common secondary structure in proteins:

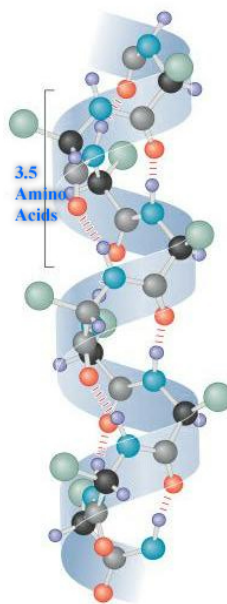


Figure 3.1: Diagram of an alpha helix (from <http://commons.wikimedia.org/wiki/File:AlphaHelixProtein.jpg>).

The corkscrew or telephone cord shape of the alpha helix is fixed by hydrogen bonds between each amino group³² and the oxygen of another carboxyl group³³ in the turn of the helix, four amino acids down the chain. The typical alpha helix is around 11 amino acids long, allowing for bonds between amino and carboxyl groups at 1 and 4, 2 and 5, 3 and 6, 4 and 7, etc.

To summarize the role of multilevel organization and multipart causal interaction in protein constitution and formation: proteins are large molecules composed of amino acids. These amino acids are covalently linked into a chain via peptide bonds, but also contorted due to a host of other chemical interactions that occur among the amino,

³¹ I will use ‘multipart’ here rather than Mitchell’s ‘multicomponent’ just to emphasize this treatment of proteins as complex *objects with parts* rather than *systems with components*.

³² Recall: N-H

³³ Again: C=O

carboxyl, and variable R-groups, the shapes that clusters of these groups form, and even the various shapes formed by separate sequences. The incredible complexity of this kind of macromolecular conformation is tracked by dividing protein structure into a hierarchy of four levels. Note the implication: in the case of proteins, multilevel organization is both an ontological product of and an epistemic way to track the multipart causal interactions that generate their organizational complications.

But proteins have another dimension of complexity—*contextual plasticity*—that is made evident by their dynamics in different contexts and in the presence of independent molecular structures that interact with them as well. For example, here are several depictions of the ligand-binding domain of the estrogen receptor (or ER):

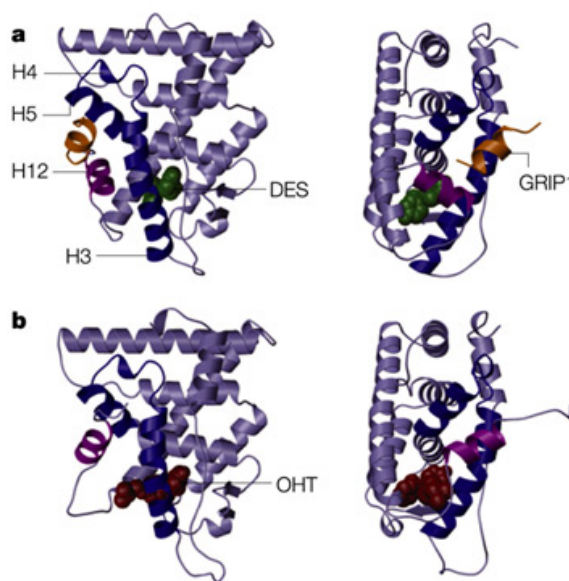


Figure 3.2: Structure of the estrogen receptor (ER) ligand-binding domain (from Simak and Coombes 2002). [Color figure available online.]

In the upper half of the diagram (labeled ‘a’), we see ER bound to an agonist DES (green sphere, labeled on the upper left-hand view), and the co-activator GRIP1 (copper ribbon, labeled on the upper right-hand view). But on the lower half of the diagram (labeled ‘b’)

we see ER binding to the antagonist OHT (red model, labeled on the lower left-hand view) rather than the agonist DES. This reconfigures the protein so that the co-activator GRIP1 cannot bind to ER in its normal position (note the absence of the copper ribbon in the lower half of the diagram, on either the left-hand or right-hand view of the protein). Comparing the upper and lower versions of the protein, we can see the way in which OHT has distorted the confirmation of the protein via interactions with secondary structures of the protein (such as the alpha helixes labeled H5 and H12 in the upper left-hand view), filling the space that the co-activator GRIP1 would normally bind to. This distortion is especially visible when comparing the left-hand views of the upper and lower versions of the protein.

More generally, note the difference between the upper and lower versions of the protein depicted on both the left-hand and right-hand views. What this shows is that proteins are complex objects that can, like us, change their shape. They can bend and turn, fold and twist, pick things up and drop them, even attract and repel certain things. This brings us to Mitchell's fourth dimension of complexity: *evolved contingency*.

One of the constraints on evolution is that it must make do with the tools at hand. And so proteins have evolved to do different things in different places. Imagine that a protein in the body is like a person in their home. The person does different things in different places: they eat in the kitchen, sleep in the bedroom, shower in the bathroom. Although persons live in houses and eat in the kitchen we don't also expect them to eat in the shower. Similarly, proteins often do one thing in one cell type and another thing in a different cell type. Their complex structure often allows them to accomplish many different functions.

Different proteins, like persons, also do similar things in different ways. Another aspect of the contingency of evolution is that different structures of biological entities can evolve to accomplish similar functions. For example, the two proteins displayed below are both DNA-binding proteins:

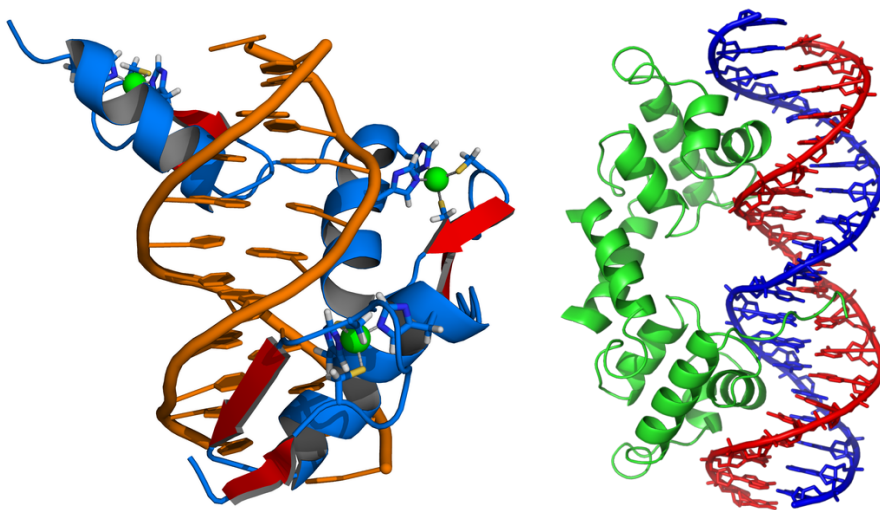


Figure 3.3: On the left, a protein with a triple zinc finger motif in complex with DNA (from http://en.wikipedia.org/wiki/File:Zinc_finger_DNA_complex.png). On the right, a repressor with a helix-turn-helix motif DNA binding domain, also in complex (from http://en.wikipedia.org/wiki/File:Lambda_repressor_1LMB.png).

In a classification of proteins by function these two molecules would be grouped together. However, one of these proteins uses a zinc finger mechanism for binding to DNA and the other uses a helix-turn-helix motif. These structures are not evolutionarily related—they are like bat wings and bird wings. In a classification of proteins by structure these two molecules would not be grouped together.

This brings us back to the main topic under discussion here: intradisciplinary classificatory (organizing) pluralism of proteins as complex objects. As the discussion of multilevel organization, multipart causal interaction, contextual plasticity, and evolved contingency shows, diverse classification systems emerge from the complexity of the

proteins being classified—from the fact that they have different features and can do different things. More thematically, as I showed in chapter 2, there are many different ways of classifying all kinds of proteins. And as I’ve shown in this chapter, proteins are complex objects. Putting these two points together: complex objects have different properties and powers on which to classify them. This leads to multiple, diverse classification systems.

But there’s a third and final thing to add to the picture: the incentive that researchers have, from distinct perspectives, to track whatever information grants them reliable inferences to certain affordances relevant to that perspective. My position is that the first datum (classificatory pluralism) is explained by interaction between the second (object complexity) and third (distinct perspectives).

Before I discuss the interaction, let me explain the claim about incentives, perspectives, inferences, and affordances. Allow me to return once more to the case of the nuclear hormone receptor superfamily.³⁴ To briefly refresh: nuclear receptors are complex multiple-domain proteins that can act as both receptors and transcription factors. Nuclear receptors tend to be expressed in many different cell types and can often do different things in each. As a result of this multi-functionality, the study of nuclear receptors exists at a junction between what might be deemed evolutionary, mechanical, and physiological perspectives.

Correspondingly, there are three different systems—phylogenetic, typological, and expressive—for classifying the proteins in the superfamily, each of which arises from

³⁴ It’s often difficult to talk precisely and accurately about all (approximately) 65 million known proteins at once.

taking any one of these different views on the same subjects—evolutionary, mechanical, and physiological. I discussed all three of these classification systems in the previous chapter.³⁵

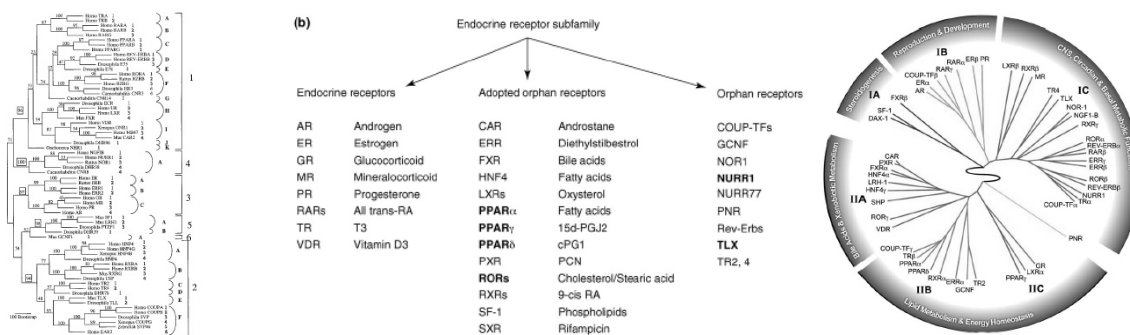


Figure 3.4: Three distinct classifications of the nuclear hormone receptor superfamily, one from each of the different classification systems. On the left, a phylogeny of the superfamily (from Nuclear Receptors Nomenclature Committee 1999). In the center, a typing of the nuclear receptors into three groups: endocrine receptors, orphan receptors, and adopted orphan receptors (adapted from Shi 2007). On the right, the ring of physiology (altered from Bookout et al 2006).

In each case the classification system grants reliable inferences to those affordances most important to the perspective from which the classification system was designed, by tracking information about whichever properties and powers regularly generate those crucial affordances.

First, emphasizing the evolutionary perspective produces phylogenies of the superfamily. This is a way of organizing nuclear receptors by alignment of primary amino acid sequence. Phylogenies of the superfamily track evolutionary, functional, and microstructural properties—loosely, their relational powers—and the tracking of these properties and powers is what grants reliable inferences to ancestry.

A second and more mechanical analysis groups the nuclear receptors into general types based on various features like the presence of “pockets” and “sticky” segments,

³⁵ Chapter 2, section 4.

binding tendencies, typical receptor action, and even primary habitat within the cell (either the nucleus or the cytoplasm). These typologies of the superfamily track superstructural, mechanistic, and (intracellular) locational properties—what I’ve called their chemical powers—and the tracking of these properties and powers is what grants reliable inferences to targetability.

Third, from the physiological perspective, we get a particularly medically relevant classification—an organization of the superfamily based on expression pattern in different suites of cell types. The classification by expression pattern tracks (intercellular) locational and physiological properties—or, the biological powers of the nuclear receptors—and the tracking of these properties and powers is what grants reliable inferences to applicability.

Researchers have incentive to track ancestry because it allows for homology inferences to be made; targetability because it allows for ligands to be discovered or designed, and experiments to be run; and applicability because it helps to direct, apply, and market basic research towards translational and clinical research, for example.

In sum, there are distinct perspectives that scientists can take on this group of gene regulatory proteins, and these distinct perspectives also contribute, along with the complexity of the objects being classified, to the generation of different classification systems. This is the site of interaction: both factors—object complexity and distinct perspectives—contribute to classificatory pluralism.

Both factors also constrain it. As a result, we should resist the temptation to go utterly relativist—this is a constrained classificatory pluralism, and the constraints

produce a set of *diverse yet grounded* classification systems.³⁶

As we've seen, proteins are complex objects—they are objects with lots of different properties. These properties interact and combine in ways that generate many and various powers. Which groups of properties and powers researchers are interested in drives the production of various classification schemes. So that's how we get diversity.

But we also get restraint. Although nuclear receptors are an object of study for many kinds of scientists—stereochemists, molecular biologists, physical chemists, evolutionary theorists, transgenomicists, gene expressionists, physiologists, and doctors, just to name a few—this does not mean that, in Mitchell's terms, "anything goes."

All of the classificatory work on nuclear receptors done by all of these different kinds of scientists clusters around three vantage points: the evolutionary, mechanical, and physiological perspectives. Each of the classification systems correlates with one of these perspectives. Any given scientist may use any one of the systems, together or apart. And they cannot be collapsed into one another because they do not co-refer—the differences among them are both intensional and extensional.³⁷

And though the classification systems are generated at least in part by researchers' interest in certain things, they are not entirely unfettered constructions. This is because

³⁶ We should also resist the temptation to declare that there are three different sets of objects being classified—there is still just one group of nuclear receptors, like there is only one elephant. This is a case of organizing rather than individuating classificatory pluralism, after all. But how to put the trunk, the tail, and the tusks back together again?

In "Complexity and Organization" (1974) Wimsatt recommends that we at least consider attempting to reconcile different views of complex objects. He also says "this would be an enormous task for even two views" (p. 69). Regardless of the enormity of the task, it should be clear by now that my proposal for integration is as follows: I think that the account of proteins as complex objects, plus the diverse set of perspectives in protein science, points to an account of the intradisciplinary classificatory (organizing) pluralism of proteins that combines object complexity (of properties and powers) with perspective diversity (of inferences and affordances).

³⁷ As I showed in subsection 4.6 of chapter 2.

these classification systems are designed, not just for codifying existing knowledge, but also for producing new knowledge. They are made for use, and this use is a kind of test (successful application is a kind of selection criterion). So, the classification systems have to work: each classification must generate reliable inferences, or no one will use it. Each inference is a prediction, and each experiment designed on the basis of said inference is a test of this prediction. And these predictions are not about what researchers want the nuclear receptors to do; they are about what the nuclear receptors can and will do.

Only if the properties and powers of the complex objects tracked via the classification system are regular, and the tracking apt, will the relevant affordances be granted with their own regularity. This dual regularity makes inferences to these affordances reliable. These features, of reliability and regularity, are products of the objects and systems in which they are embedded. In other words, these are epistemic and ontological constraints on the design and enduring use of the classification systems.³⁸ And these constraints are not subject to the whims of the scientists themselves. The scientists may pick out and study those features of interests to them; but the features they are studying are still constrained to those of the objects and their behavior, (ir)regular or (un)reliable as they are.

In the field of nuclear receptors, for example, researchers are especially interested in knowing three particular things about each nuclear receptor: one, its evolutionary history (what other receptors it's related to); two, its accessibility (whether and how it is

³⁸ One might even call these constraints appeals to explanatory and predictive power. Any proposed classification systems that fails to satisfy said appeals will simply vanish into obscurity.

targetable by natural ligands or designer drugs); and three, its involvement in particular physiological phenomena (its relevance to disease and/or enhancement). I call these things affordances, because knowing about the evolutionary history or ancestry, ligand or drug targetability, and physiological applicability of a particular nuclear receptor allows researchers to infer and apply their research in valuable ways. Knowledge of these things *affords* research. So, nuclear receptors are classified in ways that group together those receptors that are similar to one another in ways relevant to the ability to make inferences to those affordances vital to each of the three perspectives.

In another sub-field of protein science, researchers might or might not be interested in these affordances, as well as other, different ones. Whatever the affordances are in any particular case, it is awareness of the various properties and powers of the complex objects under study that grants inferences to the relevant affordances. So, researchers might group some proteins together on the basis of one kind of shared property, like amino acid sequence. They might group other proteins together on the basis of another kind of shared property, like secondary structure. Still other groupings might be based on shared powers—like the ability to bind to DNA.

But each of these groupings are created because knowing about and tracking different sets of properties and powers grants inferences to relevant affordances.³⁹ With

³⁹ The distinction between properties and powers is required because of one particular feature of proteins as complex objects—evolved contingency. To explain why, I need to digress a bit.

Whether biological kinds are proper natural kinds has been long and intensely debated. There are two common positions to take on biological kinds: as essentialist or cluster kinds. As an example of the former, Brian Ellis has recently advocated an essentialism about biological kinds that he calls ‘New Essentialism’ (see Ellis 2002). To be a member of an essentialist kind requires possessing the set of essential or intrinsic properties which define the kind. Cluster kinds are more forgiving. One paradigmatic example of a cluster kind account is Richard Boyd’s account of HPC or homeostatic property cluster kinds. Members of these kinds generally share but do not essentially share all relevant kind-making properties. Recently, Anjan Chakravartty has offered an account of property sociability that is meant to supplement

this information about the complex objects themselves, scientists can infer from what they know about certain proteins in the group in order to generate hypotheses about other proteins in the group. More groupings means more possible inferences—but the number of groupings is limited to those that are based in properties and powers of the complex objects themselves, via their granting of the relevant affordances. In other words, even diverse classification is constrained by what facts and information are relevant to meeting practical demands in the field.

I call this process by which multiple yet constrained classification systems are generated *selective naturalism*. In other words, this is how complex objects in the context of scientific investigation with distinct perspectives can lead to a complicated set of diverse yet grounded classifications of the complex objects under study.

Section 5: Conclusion

The fact that these classification systems are diverse yet grounded is an important one. In the introduction to this chapter I discussed work by both Kitcher and Dupré. Both of these authors have pluralist philosophies of science. But one of them is known for his “explosive” realism (Dupré), and the other for his “modest” realism (Kitcher).

What I hope to have shown here is how we might allow for and explain a case of modest

accounts of cluster kinds, by explaining the basis for clustering in a new way (see Chakravartty 2007). The idea is that properties are not randomly distributed, but are rather systematically sociable in various ways.

Chakravartty’s suggestion is a good one but it is insufficient when applied to biological objects with evolved contingency. The problem should be familiar by now: some biological kinds will share powers but not properties, or properties but not powers. In these cases property sociability will either not detect a relevant kind, or it will wrongly detect. So, biological kinds that display evolved contingency—like species and proteins—will not be adequately characterized by Chakravartty’s property sociability. Since these are two of the main problem candidates for biological kindhood, we need an account that explains their clustering as well. Adding the distinction that I have made between properties and powers supplements Chakravartty’s proposal in order to deal with evolved contingency.

classificatory pluralism à la Kitcher, without inevitably sliding towards the promiscuity of Dupré.

I also hope that by explaining how the complexity of proteins arises, as well as how this complexity affects their classification, we might begin to explain why the classification of complex biological and chemical kinds in general has been so difficult to understand—so complicated, in other words. These things tend to be complex; proteins are just one example of complex biochemical things. But by picking out what features make proteins complex, and explaining how this complexity complicates attempts to classify them, we can also construct a candidate account of complex biochemical classification. We can then look for these complexity-generating features in other biological and chemical things, and consider the candidate account as an explanation of other cases of complicated classification of complex biochemical objects.

I have been trying to show how protein complexity structures protein classification, and how this complexity—while complicating matters greatly—has also been co-opted as a tool for not just tracking information, but also for generating successful inferences, predictions, experiments. Basically, the idea is that classification of complex objects, while certainly complicated by that complexity, can also be used as a tool in scientific practice—as a tool for discovery. More on this idea in the next and final chapter.

REFERENCES

- Bechtel, W., & R. C. Richardson. 1993. *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton: Princeton University Press.
- Bookout, A., Jeong, Y., Downes, M., Yu, R. T., Evans, R. M., & Mangelsdorf, D. J. 2006. Anatomical profiling of nuclear receptor expression reveals a heirarchical transcriptional network. *Cell* 126:789–799.
- Cartwright, N. 1983. *How the laws of physics lie*. Oxford: Oxford University Press.
- Cartwright, N. 1999. *The dappled world: A study of the boundaries of science*. Cambridge: Cambridge University Press.
- Chakravartty, A. 2007. A metaphysics for scientific realism: Knowing the unobservable. Cambridge: Cambridge University Press.
- Cracraft, J. 1983. Species concepts and speciation analysis. In R. F. Johnston (ed.), *Current ornithology*, 159–87. New York: Plenum Press.
- Dupré, J. 1993. *The disorder of things: Metaphysical foundations of the disunity of science*. Cambridge, MA: Harvard University Press.
- Dupré, J. 2001. In defense of classification. *Studies in the History and Philosophy of Biology and Biomedical Science* 32: 203–219.
- Ellis, B. 2002. *The philosophy of nature: A guide to the new essentialism*. Chesham, UK: Acumen.
- Feyerabend, P. K. 1975. *Against method: Outline of an anarchistic theory of knowledge*. London: New Left Books.
- Kitcher, P. 1981. Explanatory unification. *Philosophy of Science* 48:507–531.
- Kitcher, P. 2001. *Science, truth, and democracy*. Oxford: Oxford University Press.
- Kuhn, T. 1996. *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press. (Original work published 1962)
- Mayr, E. 1957. Species concepts and definitions. In Ernst Mayr (ed.), *The species problem*, 371–88. Washington, D.C.: American Association for the Advancement of Science.
- Mitchell, S. D. 2003. *Biological complexity and integrative pluralism*. Cambridge:

Cambridge University Press.

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

Nuclear Receptors Naming Committee. 1999. A unified nomenclature system for the nuclear receptor superfamily. *Cell* 97: 161–163.

Putnam, H. 1975. “The meaning of ‘meaning’.” In K. Gunderson (ed.), *Language, mind, and knowledge*, 131–193. Minneapolis: University of Minnesota Press.

Shi, Y. 2007. Orphan nuclear receptors in drug discovery. *Drug Discovery Today* 12: 440–445.

Simak, A., & Coombes, R. C. 2002. Endocrine-responsive breast cancer and strategies for combating resistance. *Nature Reviews Cancer* 2: 101–112.

Simon, H. A. 1981. *The science of the artificial*. Cambridge: MIT Press.

Stanford, K., & Kitcher, P. 2000. Refining the causal theory of reference for natural kind terms. *Philosophical Studies* 97: 99–129.

Templeton, A. 1989. The meaning of species and speciation: A genetic perspective. In D. Otte & J. Endler (eds.), *Speciation and its consequences*, 25. Sunderland: Sunauer Associates.

Van Valen, L. 1976. Ecological species, multispecies, and oaks. *Taxon* 25: 233–9.

Wimsatt, W. C. 1974. Complexity and organization. In K. F. Schaffner & R. S. Cohen (eds.), *PSA 1972*, 67–86. Dordrecht: D. Reidel.

Wimsatt, W. C. 1986. Forms of aggregativity. In B. Donagon, A. Perovich, & M. Wedin (eds.), *Human nature and natural knowledge: Essays presented to Marjorie Grene on the occasion of her seventy-fifth birthday*. Boston: D. Reidel: 259–291.

CHAPTER 4: CLASSIFICATION AS A TOOL FOR DISCOVERY

New ideas are exciting. Yet how they develop is mysterious.

Lindley Darden's *Theory Change in Science: Strategies from Mendelian Genetics* (1991)

Abstract: The previous chapter demonstrated how object complexity combined with distinct scientific perspectives produces a constrained classificatory pluralism that is both diverse and grounded. This final chapter shows how having a set of grounded but diverse classification systems helps rather than hinders protein scientists, by suggesting reliable inferences to relevant affordances. In other words, the chapter documents how distinct ways of organizing consistently individuated proteins can be used, alternately and together, to direct and further research. So, this final chapter of the dissertation both documents a novel aspect of scientific practice—the use of classification as a tool for discovery—and explains how this practice works and why—drawing again on the account of complex objects and their pluralist classification.

Section 1: Introduction

In chapter 2, I explored the (intradisciplinary, organizing) classificatory pluralism of proteins. In chapter 3, I sought to explain this classificatory pluralism by providing an account of proteins as complex objects. Briefly, the idea was that protein classification is diverse yet grounded—that it is both pluralist and constrained. Both the pluralism and the constraint are each a product of interaction between the object complexity of proteins and the perspectival variety within protein science—between the complexity of the objects being studied and the variety of perspectives that scientists can take throughout the course of that study.

So, there is a plurality of classification systems and actual classifications on offer in protein science. This plurality generates an interesting question: how does scientific investigation within the field of protein science manage this multiplicity of classification schemes? This chapter provides an answer, and it is not just about coping with some sort of messy pluralism: as it turns out, constrained classificatory pluralism can be used as a tool for discovery.

In this chapter of the dissertation the way in which multiple classifications of the same set of entities can be used to narrow and direct research is made explicit, explained in context and by example, conceptually demonstrated, and finally, examined for further application to both scientific and philosophical investigations. Again, I use nuclear receptor research as an exemplar.

The next section of the chapter, section 2, summarizes the complex state of the field, reiterating the extent of the challenge presented by nuclear receptors—their ubiquity, variety, ambiguity, and disorganization. Section 3 explores the way in which traditional research assumptions and practices are often confounded by these complexities of the superfamily, generating at least four major constraints on research in the field. Despite all this, the scientific study of nuclear receptors has been wildly successful. Section 4 explains how these constraints can be overcome by using multiple classification systems together, as a means of tracking various and distinct similarity relations to make inferences to affordances that guide research. This section also shows how the use of classification as a tool for discovery comes out of and can be seen in current laboratory practice and publications. In section 5 the concept of using classification as a tool for discovery is systematically demonstrated, in order to more

precisely characterize the phenomenon and explore its broader usefulness. The concluding section summarizes the generating mechanism of the now precisely characterized phenomenon and briefly explores some implications.

Section 2: Complications

In the previous chapter I argued that proteins are complex objects. This includes nuclear receptors. So it has already been established that these gene regulatory proteins are macromolecules with lots of interacting parts—and that they display multilevel organization, multipart causal interaction, contextual plasticity, and evolved contingency. The complexity of these proteins is one thing that makes their study quite complicated.

But there are several other factors that further complicate matters. Many of these issues have already been mentioned in passing, elsewhere in the dissertation. So in this section I will simply conduct a brief review.

There are countless nuclear receptors within the entire animal kingdom—but only the animal kingdom. No instances of or precursors to the nuclear receptors have been found in plants, fungi, or bacteria. Known totals within particular species range somewhere between 21 nuclear receptors in *Drosophila melanogaster*, and almost 300 in *Caenorhabditis elegans* (Zhang et. al 2004).

The nuclear hormone receptor superfamily is occasionally called ‘the nuclear receptor superfamily’, ‘the nuclear hormone receptor family’, or even just ‘the nuclear receptor family’. An unspecified reference to ‘the nuclear hormone receptor superfamily’ (or any of the other, related titles) often means the nuclear receptors found only in humans, or mice and humans, or mice, humans, and rats.

For quite some time it was thought that there were 49 nuclear receptors in humans. However, it turns out that one of the sequences, which had been thought to code for a particular receptor (FXR β), is actually a pseudogene¹ in humans. Three additional gene sequences, potentially related to the nuclear receptor gene sequences, have also been identified, each of which contain what look like two DNA binding domains. Since it is unknown whether these sequences actually code for receptors in humans, they are only occasionally included in the superfamily.

Some of the members of the nuclear “hormone” receptor superfamily are receptors for hormones (these are called the ‘endocrine receptors’), but the ligands of others aren’t hormones at all. These nuclear receptors are simply activated by different metabolic ligands. Other nuclear receptors are orphans—meaning they have no known ligands at all. And some of the “nuclear” receptors reside in the cytoplasm, rather than the nucleus.

The nuclear receptors were also named rather haphazardly, as they were discovered. An attempt has been made to standardize the naming (see Nuclear Receptors Naming Committee 1999). But some of the supposedly replaced names have lingered (for example, FXR is often still called BAR). In other cases there is simple ambiguity: in nuclear receptors with names including numerals, Roman numerals, and Greek letters, these are often used interchangeably (as in COUP-TFII, also known as COUP-TF2, or COUP-TF β).² In certain cases this can be genuinely confusing (since, for example, TR2

¹ A pseudogene is a DNA fragment within one genome so similar to a functional gene in another genome that it looks like a fully functional gene, but it doesn’t actually code for proteins or get expressed in cells.

² At least some of this ambiguity results from the fact that various key databases in the field are not Greek-letter and/or Roman-numeral friendly. This has led to many of the receptors whose names include these symbols gaining additional names, whereby the symbols have been switched out for regular numerals and

and TR β are actually two different receptors). In another case, there is a priority dispute preventing standardization: one lab discovered PPAR β , another PPAR δ , and it turns out that these are one and the same receptor. Neither lab will adopt the others' terminology for fear of conceding discovery.

Finally, as was documented in parts of chapters 2 and 3,³ the classification of the nuclear receptors in the superfamily is quite complicated. There is a trio of classification systems on offer; and many actual classifications, each of which belongs to one of the three systems. This classificatory pluralism is, at least, a constrained pluralism.

Now it might seem like this rather complex, complicated, and confused state of the science would impede investigation—that nuclear receptor research would be struggling under the weight of it all. But just the opposite is true. For the last 30 years or so, practically since the discovery of the nuclear hormone receptor superfamily, the field has been in a constant state of production and success, as can be judged on various metrics—including publication rates, acquisition of funding, formation of new laboratories, discovery of molecules and mechanisms, generation of explanations and models, even winning of awards, making of personal fortunes, and even on occasion, investment by industry.

The fact is that researchers in the field have developed tools to overcome many of the constraints. In the next section I will present some of the most significant constraints that have been overcome with the help of a particular tool—the use of classification for discovery—that I will discuss in the section after that.

spelled-out versions of the Greek letters.

³ In both cases, the part in question is section 4.

Section 3: Constraints

My study of the scientific field of nuclear receptors has led me to identify four common constraints on the research, each of which will be discussed here.⁴ These constraints are all products of the ways in which complex or challenging features of nuclear receptors confound traditionally reliable scientific assumptions or techniques. In other words, there are ways in which generally good scientific assumptions get blocked, or common techniques fail to produce instructive results, because of distinctive features of the nuclear hormone receptor superfamily. These impediments to common practice generate constraints on nuclear receptor research.

Sub-Section 3.1: Inaccessibility

Many of the nuclear receptors have been identified, not as nuclear receptors *per se*, but as gene sequences that are sufficiently related to other sequences that do code for known nuclear receptors. In other words, sequence alignment rather than functional similarity has determined superfamily composition to a large extent. By 1990, 15 nuclear receptors with known ligands among the fat-soluble hormones had been identified, but the rest of the burgeoning superfamily were ‘orphans’—meaning, without any known ligand or function as receptors (Mangelsdorf et. al 1995). Some of those have since been ‘adopted’—meaning, ligands have been found—or co-factors⁵ have been identified.

⁴ Of course, at different times in the history of the field there have been different and evolving constraints—some of which are discussed in appendix A. The discussion in this chapter will focus mainly on currently observed constraints and successful responses to them—and the basis for these observations is the sociological study described in appendix B.

⁵ Co-factors are other regulatory proteins that, along with specific ligands, affect the activity of nuclear receptors. They can be either co-activators (recruited in the presence of the ligand-receptor complex to activate transcription) or co-repressors (actively mediating unliganded receptors).

Today, out of the 48 known nuclear receptors in humans, at least ten are still total orphans.

Correspondingly, some of the nuclear receptors in the superfamily have been extensively studied and are quite well understood; but some have been little more than identified as nuclear receptors. The true orphans, for example, tend to be poorly understood. This is in no small part due to the fact that they lack known ligands or co-factors. Without knowing what activates the orphans, and thus how to activate them, these receptors can be incredibly hard to study.

This constraint can be overcome by discovery of a ligand or co-factor for a previously-orphaned receptor, of course, and this often spurs research on that nuclear receptor. For example, research on the orphan receptor NURR1 stalled for many years as efforts to find a ligand were stymied by, it turned out, hydrophobic side chains in the portion of the LBD that normally acts as a pocket for the ligand. Once a novel hydrophobic interaction surface, outside the normal LBD, was discovered, along with potential co-activators to bind to it, NURR1 research flourished (Shi 2007).

This first constraint remains an obstacle for many nuclear receptors, however—as mentioned above, quite a few orphans have not yet been adopted. So a common assumption—receptors have ligands—has not always been borne out in the case of the nuclear receptors. Even for the endocrine and adopted receptors, there are additional complications.

Sub-Section 3.2: Information Overload

This includes the second constraint: nuclear receptors are often expressed in a

variety of cell types, so screening for expression often provides something like a surfeit of information. In other words, a generally helpful technique—generating an expression pattern for a particular receptor in order to determine where that receptor is active and thus gaining some clues as to its effects—can be, in the case of nuclear receptors, rather unhelpful—finding out that a receptor is active all over doesn't provide any special clues as to what it might be doing.

Again, of course, this constraint can be overcome, and has been in some cases. PNR displays a uniquely restricted expression, and the discovery of its almost exclusive presence in the eye led to determination of its role in the development and function of photoreceptors. For at least half of the nuclear receptors found in humans, however, expression has been detected in more than 30 cell types—in each case potentially including, but not limited to eye, brain stem, cerebellum, corpus striatum, olfactory bulb, spinal cord, hypothalamus, pituitary, thyroid, tongue, brown adipose, muscle, heart, white adipose, preputial gland, thymus, epididymis, prostate, lungs, vas deferens, uterus, aorta, seminal vesicles, spleen, bone, skin, stomach, kidney, adrenal, testis, ovary, pancreas, duodenum, jejunum, ileum, colon, gall bladder, and liver (Bookout et. al 2006).

So, in the case of nuclear receptors, screening for expression can offer substantial quantities of information, but often little of it is substantially helpful. Widespread detection does very little to indicate a receptor's mechanism of action or what aspects of physiology it affects. Other common techniques also fail because of widespread expression of nuclear receptors. This leads to the third constraint: simple knockout experiments often fail in the case of nuclear receptors, when widespread expression of a receptor makes knocking it out lethal.

Sub-Section 3.3: Diffuse Effects

A knockout experiment is one in which a normally functional gene is “knocked out” or removed by an experimenter from an otherwise undisturbed genome. Then, the experimenter attempts to generate an organism with the disrupted genome and study the effect the loss of that gene has on the organism’s physiology. This is supposed to generate a sense of which aspects of physiology are normally influenced by the gene.⁶

In nuclear receptor research, knockout experiments are usually done in mice, by removing the gene for a nuclear receptor of interest. But nuclear receptors often have widespread effects in various tissues, and elimination of even one nuclear receptor can prohibitively affect development. It is difficult to study the disruption of an organism’s physiology without the organism whose physiology has been disrupted.

Even in cases where non-lethal knockouts are produced, results can be difficult to interpret. For example, a series of knockout experiments conducted on the TR α and TR β receptors produced different effects for each experiment. In other words, introducing different mutations within the same genes led to differential disruption of physiology. Out of 17 introduced mutations, 13 different non-lethal knockouts were produced, generating at least 11 distinct expression patterns, each with its own expression profile. As the researchers reporting these results said, “it appears that different mutations introduced in the TR α or TR β loci can have strikingly different consequences, making, at first glance, the interpretation of this data difficult” (Flamant and Samarut, 2003, p. 85). So, these loci can be incredibly complex.

⁶ This is somewhat analogous to cognitive studies that reason from the impact of a brain lesion on mental activity to the usual mental function of that area of the brain. In the case of a knockout experiment, the inference is from the impact of a gene knockout on physiology to the usual physiological contribution of that gene.

Sub-Section 3.4: Strain on Resources

The frequent failure of general, initial techniques to discern the relevance of a particular nuclear receptor to the maintenance of physiology indicates that more targeted, specific approaches are often required. But there is a fourth constraint: nuclear receptors can be found in so many different kinds of tissues, they are active in so many distinct physiological pathways, and their malfunction contributes to a host of pathologies, that it would be massively wasteful, almost to the point of impossibility, to attempt to elucidate the influence of any particular nuclear receptor on physiology by simply going down the list of things that other nuclear receptors have been determined to affect, in each case applying the particular techniques and specific tests from that sphere to the nuclear receptor of interest. Relatedly, taking any experiment that is at all complicated and simply conducting it almost fifty times, on each receptor, to see if any interesting effects are produced, takes a lot of time and material resources.

This constraint, like the others, can be overcome in some cases. In fact, this fourth one is, in some sense at least, an opportunity rather than a constraint. An extraordinarily well-funded and equipped laboratory could adopt what might be called a ‘shotgun approach’, simply applying every known test, screen, experiment, and technique to an understudied receptor, or repeating one experiment for all the receptors, and then looking for any interesting result and pursuing it. The effect of the enzyme OGT on insulin resistance, for example, was discovered using a version of the latter approach (discovery documented in Yang et. al 2008). But labs with resource constraints (i.e., most labs) are not in a similar position, being unable to afford such vast and undirected expenditure. So this constraint on the process and design of research is one with rather

interesting sociological and institutional implications.

Section 4: Classification as a Tool for Discovery

The previous section introduced four constraints on nuclear receptor research, each of which can be individually overcome, but which together act to generate and shape a complex and challenging field for scientific investigation. Researchers in the field have developed not just specific means of overcoming each of these challenges, but also, somewhat accidentally it seems, a general framework for tracking overall complexity and targeting research efforts.

This, finally, is where the set of diverse yet grounded classification systems comes in handy. Each different system stores information about a distinct set of features of the nuclear receptors: phylogenies of the superfamily track evolutionary, functional, and microstructural properties—loosely, their relational powers—and the tracking of these properties and powers is what grants reliable inferences to ancestry. Typologies of the superfamily track superstructural, mechanistic, and (intracellular) locational properties—what I’ve called their chemical powers—and the tracking of these properties and powers is what grants reliable inferences to targetability. Finally, the classification by expression pattern tracks (intercellular) locational and physiological properties—or, the biological powers of the nuclear receptors—and the tracking of these properties and powers is what grants reliable inferences to applicability.

The classification systems were each created and now persist simply because they were and are useful in the understanding and management of the incredible complexity of the nuclear hormone receptor superfamily and its network of nuclear receptor action

within organisms. Interestingly, no one in the field set out to construct a multiplicity of classification systems and actual classifications; no one has attempted to consolidate them; none of them have been used to try and outcompete the others. Rather, the various classifications have each sprung up as relevant and useful features of nuclear receptors have been identified, and as the need to track said features throughout the superfamily has evolved.

In the next sub-section I will explain the origin of each of the three classification systems for the nuclear hormone receptor superfamily. This brief history demonstrates the way in which each classification system develops from a pragmatic combination of ability and need. After that I'll explain how the various classification systems can be applied, separately or together, to target research. Finally, in sub-section 4.3, I'll discuss two specific cases of classification being used as a tool for discovery, which I have detected—one in the laboratory, and the other in print.

Sub-Section 4.1: Origins

The first classification system to develop was the typology, since this is the typing which tracks, among other things, the location of nuclear receptors within the cell. The presence of these receptors inside the nucleus was precisely what brought them to the attention of researchers, and which offered a challenge to Sutherland's Nobel-winning second messenger hypothesis.⁷ Then the challenger became the challenged, with the discovery of "nuclear" receptors outside the nucleus. Of course, these were simply made a part of the (super)family. As the superfamily grew, it became important to keep track

⁷ Please see appendix A for more historical detail.

of which resided inside the nucleus and which resided outside it, if only for practical purposes. Tagging, for example, worked differently on the receptors in each group. And so the first classification system was born, and nuclear receptors were each typed as: type I (inside the cytoplasm); type II (inside the nucleus); or type III (orphan). This classification system aims to reliably track targetability.

The advent of cheaper and easier sequencing techniques led to the generation of the phylogenetic classification system and the various actual phylogenetic classifications. Throughout the development of the superfamily and alongside the relevant technological innovations, amino acid sequence alignments, partial gene sequence alignments, domain-based alignments, and finally whole-sequence alignments have all been conducted. Perhaps the definitive phylogeny dates to 1999 and accompanies the published effort to standardize the nomenclature in the field (in Nuclear Receptors Naming Committee 1999). Regardless, these phylogenies are primarily used to track interspecies and intraspecies evolutionary relations among nuclear receptors—these classifications aim to track ancestry.

The physiologic superfamily is the most recently constructed classification system and dates only to 2006 (initially published in Bookout et. al 2006). This late development is unsurprising given the massive effort required in tracking the expression of almost 50 receptors throughout 39 different cell types. The task became possible only with the various developments of qPCR,⁸ certain robotic equipment, and assorted computing and online resources. But now that the technology is available (at least, to incredibly well-funded labs), a classification of the nuclear receptors based on expression

⁸ Quantitative polymerase chain reaction.

pattern in various cell types has been developed. The point of this classification system is to group receptors by cell, and associated cellular processes, in order to relate particular receptors to certain physiological functions. Normal physiological function is of course associated with health of an organism; whereas abnormal physiological function is associated with disease. By grouping the nuclear receptors physiologically, researchers can make testable inferences as to with which physiological and disease pathways any given receptor is predominately involved. This classification aims to track applicability.

Sub-Section 4.2: Application

Each of these classification systems registers information about shared properties and powers among distinct sets of features of the nuclear receptors. These shared features, or similarity relations, make inferences from one receptor to another reliable. The inferences are to affordances, which are useful for research. This set of relations among classificatory information, properties and powers, reliable inferences, and relevant affordances explains how altogether, and even implicitly, a set of diverse yet grounded classification systems can be used to guide research.

For example, when initiating investigation of a relatively understudied nuclear receptor, possibly an orphan, researchers in the field face the unpleasant combination of having, on the one hand, no particularly promising avenues of research and, on the other hand, a positive excess of possibilities, none more plausibly relevant than any other. This is the combined problem of constraints 2 and 4.⁹ But each of the mechanistic, phylogenetic, and physiological classifications of the superfamily can be used to select a

⁹ From section 3 of this chapter.

few possibly relevant approaches from among the rather extensive list of potential avenues for research.

This is because any particular, relatively unknown nuclear receptor of interest can be located within any of the available classifications of the superfamily. In each classification system, each nuclear receptor is grouped with others, indicating that these receptors have certain features in common. The presence of these similarities suggests the possibility that neighboring receptors share additional, functionally relevant similarities such as of mechanism of action, effect on physiology, possible pathology, and susceptibility to certain techniques, for example.

So a relatively understudied nuclear receptor, grouped in a particular classification system with any other relatively well-studied nuclear receptors, can be investigated for involvement in the same sorts of things that its neighbors are. This narrows down, but does not excessively restrict, the list of potential ideas, effects, applications, and techniques to pursue in the early stages of the investigation of a nuclear receptor. This is classification being used as a tool for discovery.

Using classification in this way is not necessarily done intentionally, nor does it have to be noted explicitly by those who are doing it. Researchers might rely on one classification system for a particular inference, then refer to another to make a different comparison, without intentionally adopting the strategy of or even recognizing that they are using a multiplicity of classification systems to track distinct sets of features and direct research. Adopting a practice is not the same as nor does it necessarily entail theorizing about that practice. This is especially true in this case, where the practice being described—using multiple classification systems together as a tool for discovery—

is actually composed of a set of much more obvious, explicitly adopted practices—using any particular classification system at any given time.

Sub-Section 4.3: Evidence

This account of classification being used as a tool for discovery derives from extensive observation¹⁰ as well as theoretical analysis of those observations along with a broader examination of scientific practice,¹¹ rather than from scientists' descriptions of the practice. This practice is rarely discussed explicitly or extensively even by those who use it—something which is, unfortunately, all too common for many of those aspects of scientific practice that happen “in the head.” (Hence the lead quote from Lindley Darden.) But it is still possible to find traces of the use of classification as a tool for discovery. In this sub-section I'll discuss two places where these traces can be found. The first is in the laboratory; the second is in print.

First: as in documented in appendix B, a laboratory study was conducted as part of the examination of the field of nuclear receptor research. The members of this laboratory included several graduate students, and observation of their progress was particularly helpful for elucidation and understanding of the often-implicit practice of using multiple classification systems as a tool for discovery.

Observation of graduate students, as they progressed through periods of supervision and instruction towards independence and self-sufficiency, provided particular insight into how research in the field progresses. Upon joining the lab, these

¹⁰ Documented in appendix B.

¹¹ Documented in chapters 2–4.

students have to select and design projects that are both doable and worth doing. In other words, they need to pick a project easy enough to accomplish within the time frame of the program (which, in this case, allowed for a maximum of 7 years), but difficult enough to accomplish something novel and worthwhile. As they consider, select, design, and implement their projects, many aspects of practice that are implicit or self-directed in more experienced researchers, like post-docs and PIs, are made explicit in order to train the students.

For example, I witnessed several students succeed in getting their PhD from the lab and go on to postdoctoral positions in the field; I witnessed others fail to do most or all of these things. Those who succeeded were generally the ones who joined ongoing research programs or who took direction with respect to which receptors to work on. Those who failed tackled receptors that no one else in the lab was working on, and did not heed recommendations about which receptors were “unbreakable.”

Direction was usually given in the form of which receptors were “near” others on which classification systems, and what was already known about their neighbors. Receptors in certain groups (such as the PPARs) were particularly apt subjects of study because they were well-positioned in terms of the affordances granted by all three classification systems: they had obvious evolutionary relatability (to well-studied paralogs), targetability (they all had known ligands), and applicability (they were implicated in diabetes and cardiac disease). Both graduate students who worked on PPARs for their dissertations, despite having some disciplinary and behavior issues within the lab, managed to complete their projects and move on. In contrast one of the technically most adept and diligent graduate students in the lab failed to progress after

insisting on working on an orphan receptor, ERR γ , whose expression pattern was indistinct relative to other receptor in that cluster and whose knockouts were lethal.

Second: it is also possible to find textual traces of the influence that all three of the systems for classifying nuclear receptors wield, even together. For example, in a turn-of-the-century review of the superfamily (Chawla, Repa, Evans, & Mangelsdorf 2001), the first three paragraphs are dedicated, one per system, to each of the three ways of classifying nuclear receptors and relating them to one another. The first paragraph is about the family relationship—about how cloning groups the receptors together. These claims stem from phylogenies of the superfamily. The second paragraph is about structural, mechanical, and intracellular locational concerns—about how receptors share domains, motifs, and conformational changes. These claims stem from typologies of the superfamily. And the third paragraph is about “the importance of nuclear receptors in maintaining the normal physiological state” (Chawla et. al 2001, p. 1866)—about how there are two distinct paradigms of nuclear receptor action.¹² This is a claim that foreshadows the construction of a physiological classification system.

Although the review (Chawla et. al 2001) predates the publication of the “ring of physiology” (Bookout et. al 2006), the final two authors (Evans and Mangelsdorf) are the same on both papers, and the terminology of ‘paradigms’ and ‘clusters’ reappears with and is applied to the ring of physiology and its physiologically-laden groupings. This shows that a classification system was already in development—a system from the physiological perspective—at the time of the 2001 paper, although a fully instantiated

¹² According to the review, one paradigm of nuclear receptor action is involvement with metabolism and development; the other is acting as lipid sensors (Chawla et. al 2001).

actual classification had to wait for the 2006 publication. In that publication the authors declare, with respect to their newly proposed classification: “this resource should provide new avenues for more advanced investigation into the biological role of each receptor individually and together as a superfamily” (Bookout et. al 2006, p. 789). That is a direct statement of intent that at least that particular classification might be able to “provide new avenues” of research—i.e., it is a classification system that can be used as a tool for discovery.

Section 5: Demonstrating the Concept

The previous section focused on the development and integration of multiple classification systems with diverse perspectives, affordances, and resources in the science of investigation of nuclear receptors. In short, it was focused on scientific practice. In this section I hope to make the concept of using classification as a tool for discovery perfectly clear even to those with no experience working on nuclear receptors or observing scientific practice in a gene expression laboratory.

I will do this by demonstrating how one might systematically adopt the practice of using multiple classification systems as a tool for discovery. This requires a series of steps. First, take one each of the mechanistic, physiologic, and phylogenetic classifications of the nuclear hormone receptor superfamily. Second, by tracking each grouping along a separate dimension (e.g., mechanistic with shapes, physiologic with color, and phylogenetic with spatial location), it is possible (third step) to generate a single, cohesive depiction of the superfamily that tracks all three classification systems at once. Finally, examine the cohesively classified nuclear hormone receptor superfamily

and the various similarity relations it depicts. Then, target any intriguing relationships with future research.

The following set of three figures demonstrates how to generate just such a depiction, in the three steps described above:

Type I [★]	Type II [⊕]	Unspecified [✕]
AR - ★	FXRalpha - ⊕	CAR - ✕
ERalpha - ★	LXRalpha - ⊕	COUP-TFalpha - ✕
ERbeta - ★	LXRbeta - ⊕	COUP-TFbeta - ✕
GR - ★	PPARalpha - ⊕	COUP-TFgamma - ✕
MR - ★	PPAR(beta/delta) - ⊕	DAX - ✕
PR - ★	PPARgamma - ⊕	ERRalpha - ✕
	PXR - ⊕	ERRbeta - ✕
	RARalpha - ⊕	ERRgamma - ✕
	RARbeta - ⊕	GCNF - ✕
	RARgamma - ⊕	HNF4alpha - ✕
	RXRalpha - ⊕	HNF4gamma - ✕
	RXRbeta - ⊕	LRH1 - ✕
	RXRgamma - ⊕	NGF1B - ✕
	TRalpha - ⊕	NOR1 - ✕
	TRbeta - ⊕	NURR1 - ✕
	VDR - ⊕	PNR - ✕
		Rev-ErbAalpha - ✕
		Rev-ErbAbeta - ✕
		RORalpha - ✕
		RORbeta - ✕
		RORgamma - ✕
		SF1 - ✕
		SHP - ✕
		TLX - ✕
		TR2 - ✕
		TR4 - ✕

Figure 4.1: Step 1 – Indicate the difference types of nuclear receptor with different shapes in a mechanistic nuclear hormone receptor superfamily.

Cluster IA	Cluster IB	Cluster IC
DAX - ✕	AR - ★	COUP-TFalpha - ✕
*FXRbeta - +	COUP-TFbeta - ✕	ERRbeta - ✕
SF1 - ✕	ERalpha - ★	ERRgamma - ✕
	ERbeta - ★	LXRbeta - +
Cluster IIA	RARalpha - +	MR - ★
CAR - ✕	RARgamma - +	NGF1B - ✕
FXRalpha - +	PR - ★	NOR1 - ✕
HNF4alpha - ✕		NURR1 - ✕
HNF4gamma - ✕	Cluster IIB	RARbeta - +
LRH1 - ✕	COUP-TFgamma - ✕	Rev-ErbAalpha - ✕
PXR - +	ERRalpha - ✕	Rev-ErbAbeta - ✕
RORgamma - ✕	GCNF - ✕	RORalpha - ✕
SHP - ✕	PPARalpha - +	RORbeta - ✕
VDR - +	PPAR(beta/delta) - +	RXRbeta - +
	RXRalpha - +	RXRgamma - +
	TR2 - ✕	TLX - ✕
	TRbeta - +	TR4 - ✕
		TRalpha - +
		Cluster IIC
		GR - ★
		LXRalpha - +
		PPARgamma - +
PNR - ✕		

Figure 4.2: Step 2 – Associate a different color with each cluster of nuclear receptors in the physiologic nuclear hormone receptor superfamily, while also importing the type-indicating shapes from the mechanistic superfamily. [Color figure available online.]

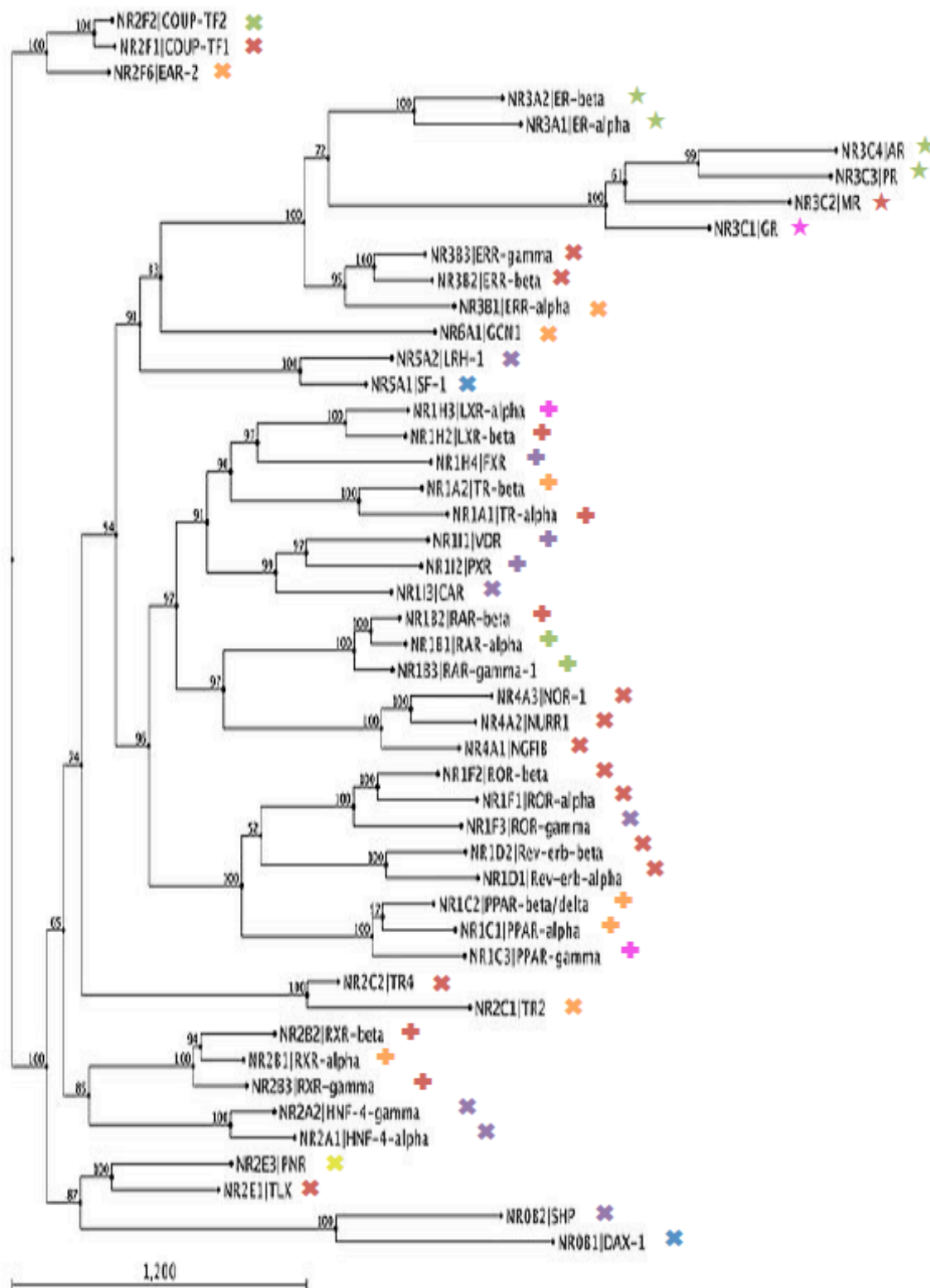


Figure 4.3: Step 3 – Place an appropriately colored and shaped indicator next to each nuclear receptor in a phylogenetic nuclear hormone receptor superfamily. [Color figure available online.]

Having completed steps 1–3, the final step is to examine the cohesively classified nuclear hormone receptor superfamily and the various similarity relations it depicts. Then, use this information about shared properties and powers to make inferences—hopefully reliable ones—to target future research towards relevant affordances.

For example: there are some apparent relations among certain nuclear receptors, now obvious because of the above diagram, which might be worth investigating. The uniformly clustered stars indicate a particularly tight link between some portion of shared, aligned sequence among those receptors and their type I mechanism of receptor action. In other words, aligned regions of those receptors are a potentially excellent place¹³ to look for a genetic basis for various behaviors typical of type I receptors—like forming homodimers and binding to inverted repeat response elements.

This clustering of type I receptors, in contrast with the scattering of type II receptors throughout the figure, also suggests a possible difference between the type I and type II categories themselves, perhaps in uniformity of members or basis in sequence.

Additionally, next to the pair of red (cluster IC) orphan Rev-ErbA receptors is the trio of ROR receptors, two of which are also red. This suggests a possible relationship between some portion of aligned, shared sequence among four receptors (Rev-ErbA α , Rev-ErbA β , ROR α , and ROR β) and cluster IC function (central nervous system with circadian and basal metabolism) that might be worth investigating.

Finally, it is also worth noting which kinds of spatial relationships are indicated by the diagram, and which are not. This is because phylogenies have an interesting,

¹³ Not just a general sequence to general structure relationship, but a particular sequence to particular structure relation.

complicated, and not necessarily intuitive structure. Things that are simply next to each other vertically are not necessarily closely related, since it is only connected pairings that matter, and the number of branches required to make a connection. The fewer the branches required, the closer the connection.

This means that, looking near the top of the diagram, it should be clear that $ER\beta$ and $ER\alpha$ are very closely related to each other, since they are paired directly together, connected by only one branch. They are more related to each other than $ER\alpha$ is to AR, since there are 5 branches connecting those two. And the ERs are (unsurprisingly) much more closely related to each other than $ER\beta$ is to $EAR2$, since connecting those receptors requires 10 branches.

This is worth noting because one might think, looking at the stretch of pluses near the center of the diagram (running from $LXR\alpha$ to $RAR\gamma$) with one cross (at CAR) in the middle of it, that this disruption might indicate something of interest in the figure. And it does, but not as straightforwardly as one might expect. This is because CAR, though closely related to the receptors marked with pluses from $LXR\alpha$ to PXR , is not nearly as closely related to the RARs. So the diagram notes a surprising divergence of CAR from some, but not all of the receptors near it, which might be worth investigating.

To summarize: these are only a few immediately observable possibilities that arise from this particular depiction of the superfamily. Of course, use of different iterations of any of the three classification systems will produce different cohesive depictions and thus will suggest different potential relationships among the receptors. Employing either a LBD-only or DBD-only phylogeny of the superfamily along with mechanistic and physiological classifications will suggest ways to relate those select

sequences with receptor type and cluster.

But this is just an example for the sake of demonstration.

Section 6: Conclusion

The principal goals of this chapter were to make explicit the way in which multiple classifications of the same set of entities can be used to narrow and direct research, to explain this practice of using classification as a tool for discovery in context and by example, and then to demonstrate the practice systematically. In this brief conclusion I will summarize the mechanism by which classification as a tool for discovery works, and then consider several implications of this discussion for science and philosophy.

Classification works as a tool for discovery by aiming to track information about which properties and powers of the complex objects being classified regularly generate certain affordances. The classification systems grant inferences to these affordances. For example: in the field of nuclear receptors, researchers are especially interested in knowing three particular things about each nuclear receptor—its ancestry, targetability, and applicability. Inferences to those affordances are granted by tracking certain evolutionary, functional, locational, microstructural, physiological, and mechanical properties and powers in classification systems targeted to each relevant affordance.

Knowledge of these properties and powers affords research and so, nuclear receptors are classified in ways that group together those receptors that are similar to one another in ways relevant to the ability to make inferences to those affordances vital to each of the three perspectives. But only if the properties and powers of the complex

objects tracked via the classification system are regular, and the tracking apt, will the relevant affordances be granted with their own regularity. The classification systems have to work: each classification must generate reliable inferences, or no one will use it. When a classification system does work in this way, it is being used as a tool for discovery.

I will close with two promising implications of this discussion. On the scientific side of things, it is possible that just making the practice explicit could be of aid to researchers, especially during education in the field. Perhaps more clearly and directly presenting this method during instruction and use would aid understanding of how to overcome some of the major constraints in the field. On the philosophical side of things, characterization of this practice enriches philosophical accounts of both scientific classification and discovery by uniting these concepts in a surprising way. The discussion also suggests a novel heuristic that might be of broader use in the philosophy of science, or even in other parts of scientific practice.

REFERENCES

- Bookout, A., Jeong, Y., Downes, M., Yu, R. T., Evans, R. M., & Mangelsdorf, D. J. 2006. Anatomical profiling of nuclear receptor expression reveals a hierarchical transcriptional network. *Cell* 126:789–799.
- Chawla, A., Repa, J. J., Evans, R. M., & Mangelsdorf, D. J. 2001. Nuclear receptors and lipid physiology: Opening the X-Files. *Science* 294: 1866–1870.
- Darden, L. 1991. *Theory change in science: Strategies from Mendelian genetics*. New York: Oxford University Press.
- Flamant, F. & Samarut, J. 2003. Thyroid hormone receptors: Lessons from knockout and knock-in mutant mice. *TRENDS in Endocrinology and Metabolism* 14: 85–90.
- Mangelsdorf, David J., Carl Thummel, Miguel Beato, Peter Herrlich, Günther Schütz, Kazuhiko Umesono, Bruce Blumberg, Philippe Kastner, Manuel Mark, Pierre Chambon, and Ronald M. Evans. 1995. The Nuclear Receptor Superfamily: The Second Decade. *Cell* 83: 835–839.
- Mitchell, S. D. 2003. *Biological complexity and integrative pluralism*. Cambridge: Cambridge University Press.
- Mitchell, S. D. 2009. *Unsimple truths: Science, complexity, and policy*. Chicago: University of Chicago Press.
- Nuclear Receptors Naming Committee. 1999. A unified nomenclature system for the nuclear receptor superfamily. *Cell* 97: 161–163.
- Shi, Y. 2007. Orphan nuclear receptors in drug discovery. *Drug Discovery Today* 12: 440–445.
- Yang, X., Ongusaha, P. P., Miles, P. D., Havstad, J. C., Zhang, F., Venus So, W., Kudlow, J. E., Michell, R. H., Olefsky, J. M., Field, S. J., & Evans, R. M. (2008). Phosphoinositide Signalling Links O-GlcNAc Transferase to Insulin Resistance. *Nature* 451: 964–970.
- Zhang, Z., Burch, P. E., Cooney, A. J., Lanz, R. B., Pereira, F. A, Jiaqian, W., Gibbs, R. A., Weinstock, G., & Wheeler, D. A. (2004). Genomic Analysis of the Nuclear Receptor Family: New Insights into Structure, Regulation, and Evolution from the Rat Genome. *Genome Research* 14: 580–590.

CONCLUSION

It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.

Albert Einstein's *On the Method of Theoretical Physics* (1934)

The aim of science is to seek the simplest explanations of complex facts. We are apt to fall into the error of thinking that the facts are simple because simplicity is the goal of our quest. The guiding motto in the life of every natural philosopher should be, "Seek simplicity and distrust it."

Alfred North Whitehead's *The Concept of Nature* (1920)

One might expect genes, rather than proteins, to be the central topic in a conversation about complex biochemical objects. This would be a mistake, for several reasons. For one, genes have already been extensively studied and discussed in the philosophy of science, with good reason. As the primary units of heredity within the usual units of selection within the fundamental units of evolution (i.e., the genes in the organisms of species), genes are obviously a crucial component of evolutionary theory. Understanding genes is required in order to understand evolution. But there is more to biology, and thus there should be more to the philosophy of biology, than evolution. It is worth understanding, not just how the "endless forms most beautiful and most wonderful" have evolved,¹ but also how these forms function—in other words, how the many different kinds of organism develop and survive to be selected for or against. This implicates proteins as well as, and perhaps even more than, genes.

For another reason—one much more important to this project—genes are not actually very complex, at least not in the sense of complex objects discussed here. Genes do display several of the features that generate entity complexity, though to a limited extent: they have multilevel organization produced by multipart causal interaction

¹ From Darwin's *Origin* (1859/2001, p. 490).

(among individual chemical bases, which generate the double helix of DNA, for example); and they are absolutely the contingent products of evolution (every gene that has evolved could have evolved differently, at least by a little bit²). But genes are not actually that plastic, despite the slight flexibility that demonstrates their contingency.

This is because genes do, basically, only one tightly constrained and contextually embedded thing, and any change disruptive of that function is debilitating. In the right circumstances, with the necessary help, genes make proteins.³ Though genes have many other *properties*, that is their chief *power*. Proteins have both many properties and many powers. There are approximately 20,000 different genes in the human genome (Pennisi 2012), far fewer than expected, and all of them make proteins. Estimates of the number of different human proteins range from one hundred thousand to one million (Hinz 2010), far more than expected according to the initial “one gene – one protein” mantra. Of course that bit of dogma had to be abandoned long ago: alternative splicing, posttranslational modification, and complex pathway interactions are a few of the ways in which the expression of a single gene might result in a variety of different proteins.

Why so many different proteins? Because they do just about everything—they are the actors within cells and the bulk of their biological content. There are millions of individual proteins in each of the trillions of cells in a human body. Half of the dry weight of an *E. coli* cell is protein, whereas a fifth is RNA, and only about 3% is DNA

² Genes encode proteins by indicating primary amino acid sequence with trios of nucleobases. There are four different kinds of nucleobase in DNA—adenine, cytosine, guanine, and thymine—and thus there are 4³ or 64 possible arrangements of three of the four bases in sequence. But there are only 20 amino acids. All but two of them can be indicated by more than one trio, and there are no proteins made only of methionine and tryptophan. As a result, every known gene that codes for a protein could have had a slightly different nucleobase sequence than it actually does, while still producing the same amino acid sequence.

³ I have left out RNA because I am not talking, here at least, about how DNA makes RNA makes proteins. Rather, I am talking about how genes make proteins—which is slightly different but just as true.

(Voet & Voet 2004). I mentioned above that proteins are at least as important as genes for the development and maintenance of organisms. I would go so far as to say that proteins are something like the atomic particles of biology.

Imagine a multicellular organism as its own universe. It starts with a Big Bang of meiosis, and many characteristics of the resulting cosmos are fixed in the first three minutes of chromosomal recombination. As the universe expands countless galaxies of cells are formed. Imagine each galaxy has only one central star or supernova—a stellar nucleus. Almost all of the elements found scattered throughout the universe are generated by nucleosynthesis within these stellar bodies. But knowing where and even how these atomic particles are made is not sufficient for knowing what they do once they travel beyond the stars in which they were born. In the bulk of the physiological universe, outside of the nuclear supernovae of cellular galaxies, proteins are the atoms that are traveling and emitting, clashing and bonding, building and degrading. Understanding the organismal cosmos requires understanding more than just the *creation* of its atomic particles—the countless proteins—but also their *action*.

Of course, understanding the disparate action of countless scattered particles is hard—it's an incredibly complicated task. But this is a dissertation about the complications of grappling with complexity, as manifest in many objects of study throughout the life sciences. Sadly, complexity has not always been a friend to scientific understanding. This is because simple things are often much easier to understand than complex things—they are less complicated. And complex things are often very difficult to understand—they are more complicated. But we cannot always understand complex things by pretending that they are simple.

Pretending that complex things are simple is often a mistake. There is, however, a long and storied tradition—among both scientists and philosophers of science—of trying to understand complex things by *making* them simpler.⁴ This is often a fruitful strategy. It is also different than *pretending* complex things are simple. Pretending that something complex is simple is a way of tricking yourself and others into thinking that you’ve done one thing when really you’ve done another—like imagining you’ve understood a complex thing by pretending that it is simple.

But when you make something that is complex simple in order to understand it—without pretending that the complex thing *is* simple—you can often gain complete understanding of the simple thing that you have constructed, as well as partial understanding of the complex thing from which you derived the simple thing. But you must keep in mind that the thing you have (potentially) gained complete understanding of is the simple construct, rather than the complex original.

Gaining complete understanding of complex things themselves, without making them simpler, is much harder. In many cases it is probably impossible. However, we can at least strive for fuller understanding of some complex things by avoiding simplification when we can, and marking it when we cannot. And we should not let ourselves be lulled into thinking that the only things worth understanding are the simple ones, or that the only real kind of understanding is the complete kind that comes so much more readily to simple rather than complex things.

To succumb to that temptation is to indulge in, at best, a misleading kind of wishful thinking: that of letting ourselves think the world is one way rather than another

⁴ By abstracting or idealizing, for example.

because, well, it would just be so much easier if it were so. But it is not so. The world is full of complex things. Making them simpler can at times lend itself to partial understanding. Embracing their complexity can often do a lot more.

In this dissertation I have endeavored to grapple with the complexity of proteins. I hoped that this engagement would help to dissolve some of the classificatory complications evident in this and other biochemical sciences. To finally summarize the results of my approach: in the first chapter I explored the membership conditions of several kind terms from chemistry—including elements, the compound water, and proteins. I concluded that protein individuation, at least, is monist. In chapter 2 I looked at the ways in which monistically individuated proteins are related to one another, and concluded that protein organization, in contrast with individuation, is pluralist. In chapter 3 I offered an explanation for the constrained, intradisciplinary, and organizing classificatory pluralism of proteins. In short, the complexity of the objects involved, combined with the distinct scientific perspectives that can be adopted in the course of their study, generate a diverse yet grounded set of protein classification systems. Lastly, in chapter 4, I showed that this kind of classification is useful as a tool for discovery.

REFERENCES

- Darwin, C. 2001. *On the origin of species*. Cambridge, MA: Harvard University Press. (Originally published in 1859)
- Einstein, A. 1934. On the method of theoretical physics. *Philosophy of Science* 1(2): 163–169.
- Hinz, U., & The UniProt Consortium. 2010. From protein sequences to 3D-structures and beyond: The example of the UniProt Knowledgebase. *Cellular and Molecular Life Sciences* 67: 1049–1064.
- Pennisi, E. 2012. ENCODE project writes eulogy for junk DNA. *Science* 337: 1159–1161.
- Voet, D. & Voet, J. G. 2004. *Biochemistry*, vol. 1 (3rd ed.). Hoboken, New Jersey: Wiley.
- Whitehead, A. N. 1920. *The concept of nature: Tarner lectures delivered in Trinity College, November, 1919*. Cambridge: Cambridge University Press.

APPENDIX A

A special distinctive name seemed appropriate and ‘protein’ was the name he [Berzelius] suggested in his reply [to Mulder] a month later. The name is derived from the Greek word ‘πρωτειος’ which means ‘standing in front’, ‘in the lead’. ‘We’re number one’ is how one might translate it in the modern vernacular.

Tanford & Reynolds’ *Nature’s Robots: A History of Proteins* (2001)

Each cell has a special task that it performs only when instructed by another cell. How do cells receive these instructions? How do they interpret the instructions so as to perform the right task? The answer ... is a family of proteins called nuclear receptors.

Michael Brown’s Presentation of the *Albert Lasker Basic Medical Research Award* (2004)

Abstract: Much of this dissertation relies on a study of the nuclear hormone receptor superfamily for its case material. This appendix traces the history of the superfamily, documenting its origin in several different fields: protein science, molecular biology, and endocrinology. As the nuclear hormone receptor superfamily began to take shape as an object of scientific study in the mid-1980s, a synthesis occurred between parts of both endocrinology and molecular biology, due to the dual nature of these gene regulatory proteins as both transcription factors and nuclear hormone receptors. This composite research program endures today.

Section 1: Introduction

The history of nuclear receptors is an interesting one. These gene regulatory proteins have at least a tripartite historical origin: as proteins, as transcription factors, and as nuclear hormone receptors. In this appendix, I trace the origin of the nuclear hormone receptor superfamily through its roots in protein science, molecular biology (especially the study of gene expression and regulation), and endocrinology (the study of hormone signaling). Then I show how parts of two distinct fields, molecular biology and

endocrinology, were united in order to investigate these gene regulatory proteins as both transcription factors and nuclear hormone receptors.

In specific organizational terms: after this brief introductory section and prior to the concluding one, there are four substantive sections of this appendix to the dissertation. Sections 2 (“Protein Science”) and 3 (“Molecular Biology”) are generally pieced together from standard histories of the relevant fields, supplemented by details from the occasional philosophical or lexicographical source. Section 4 (“Endocrinology”) is a history of that particular sub-field which I have put together using a combination of secondary sources and my own primary research. Finally, section 5 (“The Nuclear Hormone Receptor Superfamily”) is based chiefly on primary sources, as well as on my own observational study (which is further documented in appendix B). All sources are noted throughout.

Section 2: Protein Science

Although the first proteins were identified more than two hundred years ago, only halting progress was made in the study of these macromolecules until rather recently. So the history of proteins, despite spanning several centuries, is a rather skewed one, with little to report from the early days and almost too much to disclose in terms of recent developments.¹

But to start all the way back at the beginning: the uncharacterized and unknown contents of living cells were once simply referred to as ‘protoplasm’ by (then-called)

¹ A plausible, though partial, explanation of the shape of this distribution is the complexity of proteins discussed in chapter 3.

natural philosophers studying organismal products and processes.² In 1789, Antoine Fourcroy distinguished albumin, fibrin, and gelatin, all from animals. Related substances from plants had also been identified, and these were “often collectively called ‘albumins’, in recognition of the prototype derived since time immemorial from egg white; the German equivalent was ‘Eiweisskörper’” (Tanford & Reynolds 2001, p. 11).

Organic chemists did not, in Fourcroy’s time, have very many tools for molecular analysis at their disposal. Early in the nineteenth century it became possible to determine the elemental composition of molecules, to give compositional formulas, and to calculate formula weights. This led to the publication of many papers that simply listed the numbers of atoms of various elements in one particular substance. With any of the albumins, the total numbers would include many hundreds or even thousands of carbon, hydrogen, nitrogen, and oxygen atoms (possibly some phosphorous and sulphur too). So the albumins had very high molecular weights, especially given the standards of the time.

Without an understanding of chemical bonding, or even the tetravalent model of carbon (not proposed by August Kekulé until the second half of the nineteenth century), little progress understanding the nature and function of the albumins was made. However, they continued to be thought relevant and important to the study of living things because of the great variety and versatility of the albumins found in, and only in, organisms of all kinds.

The term ‘protein’ was coined by Jöns Jakob Berzelius in 1838,³ just two years

² For a much more comprehensive account of the history of proteins, please see Tanford & Reynolds’ *Nature’s Robots* (2001).

³ The first usage of the term in publication was by Gerardus Johannes Mulder, also in 1838. But a close study of correspondence between Mulder and Berzelius revealed that the term was suggested to Mulder by Berzelius just before Mulder’s usage of the term in print. This misattribution of the term’s origin to Mulder

before William Whewell coined the term ‘scientist’.⁴ A term was required because there was speculation, chiefly by Gerardus Johannes Mulder, that there might be a single substance, consisting only of carbon, hydrogen, nitrogen, and oxygen, at the core of all the albumins. This was debunked shortly after the idea was proposed, but the name ‘protein’ stuck. Eventually ‘albumin’ and even the German term, ‘Eiweisskörper’, were replaced in usage by ‘proteids’ and ‘protein’.

In 1902, Emil Fischer and Franz Hofmeister both (though independently⁵) suggested that what is now known as the ‘peptide bond’ is the fundamental means by which amino acids are joined together to form proteins. Fischer himself coined the terms ‘peptide’ and ‘polypeptide’. He never realized his dream of creating, in his laboratory, the first synthetic enzyme, but he had co-discovered and named the link between one amino acid and another. This link, repeated many times, chains amino acids together in sequence, in what is today called the ‘primary structure’ of proteins.

But between Fourcroy and Fisher there is more than a hundred years. It took that long, after the discovery of these substances, for even their most basic structure to be elucidated. Proteins, though discovered rather early in the history of biochemistry, had to wait for the discovery of amino acids, theories of chemical bonding, Kekule’s tetravalent model of carbon, and finally Fischer and Hofmeister’s suggestion, for an account of their primary structure to emerge.

over Berzelius was first revealed by Harold Hartley, as he describes in a letter to *Nature* (1951).

⁴ According to the OED’s entry for “scientist, n.” this occurs in Whewell’s *Philosophy of the Inductive Sciences: Founded upon Their History* (1840). Interestingly, the OED entry also describes a prior attempt to introduce the term, by members of the British Society for the Advancement of Science, documented in their *Quarterly Review* (1834). But it was apparently found, at the time, “not generally palatable.”

⁵ Interestingly, they both (unknowingly and unintentionally) presented this finding at the same conference, on the same day. This was at the 74th meeting of the Society of German Scientists and Physicians on September 22nd 1902, in Karlsbad. Hofmeister called the substances of interest ‘Eiweissmoleküls’ and Fischer called them ‘Proteinstoffe’, however. See Fruton’s “Contrasts in Scientific Style” (1985).

And it has taken another hundred years, throughout the 20th century and into the 21st, for scientists to develop sophisticated methods for identifying the secondary, tertiary, and quaternary structure of many of these (large, incredibly complex) macromolecules. The use of X-ray crystallography to solve the structures of relatively simple proteins didn't begin until the late 1950s. And by then, the development and use of tools for protein sequencing⁶ had already begun to add to the pile of proteins whose structure needed to be solved. The development of tools for gene sequencing⁷ in the late seventies only added to the ever-increasing heap. The size and complexity of many of these sequenced molecules soon overwhelmed scientists' ability to solve their structures using contemporary techniques.

In the 21st century tools for modeling protein structure have finally begun to catch up with tools for sequencing. Coupling mass spectrometry with ever-increasing computational power has facilitated the solving of many large protein structures. But the number of known (aka sequenced) proteins is immense, and still far exceeds the number of solved (in terms of structure) proteins. There is currently specific sequential data on tens of millions of proteins. It is anybody's guess how many proteins exist in total (among humans and other organisms). Some estimate 100 million, others more than 100 billion.⁸

It is the unique combination of sequential variability and structural complexity that produces such a variety, versatility, and vast number of proteins.⁹ Historically accounting for the construction, discovery, and stabilization of even a fraction of this

⁶ Primarily, Edman degradation. See Edman (1950).

⁷ Primarily, Sanger sequencing. See Sanger, Nicklen, & Coulson (1977).

⁸ Please see chapter 1, section 3 for precise statistics about the number of proteins.

⁹ Please see chapter 2, section 2 for an account of the extent and variety of possible protein configurations.

immense group of molecules is a difficult task. In what follows I will restrict my attention to one group, the nuclear receptors, which are gene regulatory proteins and thus also have origins in the field of molecular biology.

Section 3: Molecular Biology

Scientific investigation of molecular biology began with a focus on the identification of genetic material and its mechanism of replication.¹⁰ The origin of the field is usually dated only a few decades prior to the publication in the journal *Nature* of the structure of DNA (Watson & Crick 1953). But molecular biology also has roots in the field of biochemistry, and research in that field goes quite a way back.

Biochemistry has traditionally been concerned with the chemical reactions that provide the requisite energy and other components necessary to the maintenance and function of living organisms. But the fields of biochemistry and molecular biology underwent something of a synthesis when it turned out that the molecular basis for metabolism and other key physiological processes was the same genetic material responsible for reproduction and development. In other words, DNA is complicit in a host of these crucial processes, and is thus of central importance to both biochemistry and molecular biology. Actually, the principal focus is now on the genetic material itself and all its activities: gene expression, gene regulation, gene mutation, and gene replication. Sometimes this collection of subjects is referred to as the relatively new field of molecular genetics.

¹⁰ This section provides just a brief sketch of the field. For a much more detailed historical account, please see, for example, Olby's *The Path to the Double Helix* (1974).

Regardless, research in the subfield of gene expression focuses on how the expression of genes is regulated and functions to produce various gene products, like RNA and protein, which then go on to run and power the machinery of cellular life. A more comprehensive and accurate name for the field might be gene expression and regulation, since these two processes are inextricably linked and thus often studied together. It is because of these complex relations between gene expression and regulation that the field of gene expression does not employ a unidirectional model of the relationship between the genotype and the phenotype. Work in the field is not simply a study of how the genotype produces the phenotype; nor is it a study of to what extent nature versus nurture generates the organism. In fact, research in the field of gene expression reveals precisely how the genotype ("nature") is inextricably bound up with both its immediate and distant environments ("nurture"), demonstrating just how inaccurate and unhelpful it often is to try to conceptually isolate these influences.

The tangled and complex relations between gene expression and gene regulation arise from the fact that it is the expression of genes that generates the very proteins that regulate gene expression. The presence and concentration of gene regulatory proteins in the vicinity of genes affects which genes are expressed; likewise, which genes are expressed affects the presence and concentration of gene regulatory proteins. Additionally, the system is open, rather than closed. Intracellular transfer of exogenous—externally generated—as well as endogenous—internally generated—signaling, binding, and regulatory molecules means that gene expression is affected by internal and external processes of regulation; likewise, gene regulation is affected by internal and external processes of expression. Each of these internal and external processes of expression and

regulation are in turn affected by other internal and external processes of regulation and expression, and so on, and so forth.

Of course, scientists throughout the various labs within the general field of gene expression do not simply investigate whichever of the many molecules implicated in these processes of expression and regulation they feel like. There are subfields within the field of gene expression and most laboratories focus on these and the sets of related molecules that comprise the objects of investigation within them.

The primary subject of this appendix is, of course, one particular subfield (the study of nuclear receptors) and the objects of investigation within it (the group of gene regulatory proteins known as the nuclear hormone receptor superfamily). To briefly describe the current understanding of the science, nuclear receptors (sometimes called 'genetic activation switches') are proteins that, when in complexes with ligands and occasionally cofactors, bind to special portions of DNA (often called 'response elements') and regulate the expression of various sets of genes, thereby controlling certain aspects of the physiology of organisms, including portions of development, homeostasis, and metabolism.

Because nuclear receptors are “genetic activation switches”—i.e., they act as transcription factors—they are part of the study of gene expression and regulation. But because nuclear receptors are also receptors for hormones, among other ligands, they are correspondingly an important part of the study of hormone signaling. It is to the history of this study that I will now turn.

Section 4: Endocrinology

The scientific study of hormones began just before the turn of the twentieth century, although it turns out that people of various cultures have been utilizing hormones from animal tissues to affect human physiology for millennia.¹¹ The term 'hormone', and its definition as an endogenous chemical messenger used to regulate physiological activity, was established just after the turn of the century.

Hormones rapidly became the focus of intense scientific investigation because of their perceived effect on the development and maintenance of organisms.¹² But the study of hormones was initially complicated by the fact that it was incredibly difficult to obtain any of the substances under investigation in sufficient amount. Hormones are present only in very small quantities in organisms and the molecules often have a short half-life. These traits make sense given the role of hormones as chemical messengers rather than as substrates, catalyzers, or products of chemical reactions, but they also make hormones quite difficult to isolate and extract.

In the 1910s, a few precious milligrams of thyroid hormone were extracted and purified from massive quantities of thyroid gland collected from stockyards in Chicago. When it proved possible to crystallize the substance that had been extracted, the popular and relatively new technique of using x-ray crystallography to determine molecular structure was applied. The structure was purportedly solved near the end of 1914 and presented at a conference early in 1915 (Kendall 1915). However, the initially reported

¹¹ See Fausto-Sterling (2000).

¹² For a more detailed history of the development of this field, please see Oodshoorn's *On the Making of Sex Hormones* (1990) or Fausto-Sterling's *Sexing the Body* (2000).

structure turned out to be incorrect.¹³ The correct structure of thyroid hormone, or thyroxine, was not solved until 1926, and thyroxine was first synthesized in 1927 (Harrington and Barger 1927).¹⁴ Knowledge of the structure and its eventual synthesis permitted a degree and kind of scientific investigation that had not been possible when an excess of thirty tons of offal was required to obtain just a tiny bit of the substance of interest.

Investigation of the structures and effects of this and other endogenous chemical messengers, such as the steroid hormones, progressed and was eventually linked to the study of various substances that turned out to be exogenous precursors to other chemical messengers, such as vitamins A and D. In other words, two seemingly unrelated kinds of things, endogenous hormones and exogenous vitamins, were brought together in the study of the regulation of physiological activity via chemical signaling. But theoretical accounts of precisely how such insignificant amounts of any of these small molecules managed to regulate the many complex and variable aspects of physiology remained, for decades, largely speculative. The predominant expectation was biochemical: it was postulated that hormones acted as enzymatic cofactors, regulating the rate of catalyzed reactions via allosteric modification.

This expectation can be explained with the help of a concise account of the science of enzymes. These molecules are defined as chemical catalysts: enzymes increase

¹³ Various explanations for this mistake have been offered: for example, that the scientist in question, E. C. Kendall, was rushing at the end of 1914 in order to be able to present his findings at the meeting in early 1915; that he was unskilled at x-ray crystallography because he was a physiologist and not a chemist; and that he failed to appropriately isolate the hormone from the gland material in the first place.

¹⁴ This work, correcting the structure of thyroxine, was done by a biochemist named C. R. Harrington. Kendall redeemed himself with his work on the hormones of the adrenal gland, for which he eventually received a Nobel Prize, in 1950.

the rate of a chemical reaction without being consumed by the reaction themselves. Uncatalyzed reactions occur when substrates transform unassisted into products, and vice versa. The rate at which a transformation of this sort occurs is limited by the activation energy that is required to initiate the transformation. Enzymes increase the rate of reactions by lowering this activation energy. This is possible because enzymes are molecules with an affinity for binding to substrates and inducing a slight conformational shift in the substrate, bringing it into a transition state with a much lower activation energy for the transformation from substrate to product. Individual enzymes target specific substrates and catalyze particular reactions: the position on an enzyme where a substrate binds, called the active site, has to be particularly suited to the relevant substrate in order to appropriately attract and alter it.

But enzymes themselves can be altered as well. The presence or absence of another small molecule, called a cofactor, can induce a conformational shift in the enzyme, rendering the active site for substrate binding either effective or inert. This is called allosteric modification: it is a way of regulating the effect of enzymes on reactions by activating and deactivating their catalytic power. And this is how, it was thought, hormones must work: they were cofactors that regulated the action of various enzymes that catalyzed reactions crucial to physiology via allosteric modification. At least, this was the idea until, in the fifties, concepts from molecular biology as well as biochemistry were used to develop a theory of extracellular signaling linking hormones with receptors.

The scientific use of the term 'receptor' began around the turn of the twentieth century, but its initial meaning was variable and uncertain. In immunology, for example, the term was used to refer to an endogenous molecule that detects and binds to foreign

and invasive substances; now, these are called antibodies. There were (and still are) at least two significantly distinct uses of 'receptor' in physiology: either as an organ, cell, or nerve ending that responds to sensory stimuli (a primarily neurological use of the term) or as an intracellular resident molecule that binds to an extracellular signaling or otherwise interactive molecule (a more biological use of the term). It is the latter rather than the former sense of the term that is relevant to endocrinology.

Interestingly, this sense of 'receptor' originated in pharmacology, as part of an early effort to model how drugs might affect physiology. Adopting a variation of H. E. Fischer's early lock-and-key model of enzyme action, J. N. Langley proposed that certain exogenous substances can act as a "key" in the "lock" of particular endogenous 'receptive substances'. The term 'receptive substance' was first used in this context in 1905 (Langley 1905), and an equation quantitatively expressing the receptor idea in terms of a biochemical reaction was first proposed by Langley's student A. V. Hill in 1909 (Hill 1909). Although both Langley and Hill's papers were published in the *Journal of Physiology*, and both discussed the study of the poisons nicotine and curare in order to understand "receptive substances," Hill's proposed equation was overlooked. A little less than a decade later, I. Langmuir constructed what turned out to be an equivalent formulation (see, for example, Langmuir 1918), and Langmuir's proposed equation was supported by extensive experimental data in 1926 by A. J. Clark and J. H. Gaddum (Clark 1926a, 1926b; Gaddum 1926). However, establishing the existence of a concentration-effect curve did not help to determine the elusive identity of the relevant receptive substances. It was not until the forties that it was conclusively established that enzymes were proteins, and as late as the sixties, it was still not certain that receptors,

like enzymes, were proteins too.¹⁵

Section 5: The Nuclear Hormone Receptor Superfamily

So, receptors were eventually identified as proteins. And hormones were identified as endogenous chemical messengers, or signaling molecules. But it took quite a while for scientists to develop an account of how hormones and receptors work together such that chemical signals are received and used to regulate the physiology of organisms. It was not until the fifties that Sutherland began the research that eventually led to the development of his second messenger hypothesis, and the foundation of receptor theory.

Sutherland's work eventually demonstrated how a hormone could interact with a receptor on the surface of a target cell, triggering another molecule in the cytoplasm to initiate a series of cellular events that determine the physiological response to the hormonal message (see, e.g., Sutherland & Robison 1966). His theoretical account was named the second messenger hypothesis because it includes the hormone—the first, extracellular messenger—as well as another signaling molecule—the second, intracellular messenger—linked together by a receptor, embedded in the boundary of the cell. In other words, Sutherland proposed a model of hormone signaling in which the hormone binds to the external surface of a trans-membrane receptor molecule, releasing another signaling molecule, previously bound to the internal surface of the receptor, into the cytoplasm.

¹⁵ For a more detailed history of the development of the receptor concept, see Kenakin (2004) or Rang (2006).

This model is still used to characterize the signaling action of various hormones, and in 1971 Sutherland was awarded the Nobel Prize in Chemistry for “his discoveries concerning the mechanism of the action of hormones.” But even as he gave his Nobel lecture the honoree already had to add a caveat: “of course the various relations in the endocrine system cannot all be understood in terms of a simple concept ... the primary action of the steroid hormones appears to follow an entirely different pattern” (Sutherland 1972, 405). This qualification had to be made because of Jensen’s work on the estrogen receptor.

It has been reported that Jensen began his work on the estrogen receptor as early as 1958 (see, e.g., Moore 2012). Certainly, by 1960, he was already publishing the results from a series of experiments tagging and tracking estrogen with radioactive labels. Jensen’s experiments showed the hormone accumulating in the nucleus of cells, rather than bound to the surface of cells or even in the cytoplasm where most cellular interactions occur (see, e.g., Jensen and Jacobson 1962).

This is when the mechanism by which some hormones act as chemical signals to regulate physiology was first experimentally implicated as straightforwardly genetic. Understanding this part of hormone signaling became a problem whose solution might involve genes far more directly than had previously been suspected. And this part of hormone signaling became a candidate solution for what had been a long-standing and fundamental puzzle in molecular biology.

It was T. H. Morgan who first postulated that, since all the cells of an organism contain the same genetic material, but different cell types develop and function very differently, there must be some kind of system of differential expression of different

genes in different cell types, or differential gene expression. But what could this system be? That was Morgan's puzzle, and the answer to it remained unknown so for quite some time. Now of course the standard answer is that a network of transcription factors are what regulate differential gene expression—and nuclear receptors are an important part of that network.

But Jensen's data was not immediately linked to Morgan's proposed system of differential gene expression. For one, the results of Jensen's experiments astounded the field. It took a while for his data to be accepted and then incorporated into the corpus of material suitable for transmission and further investigation. For another, Morgan's insight had occurred long before it was even known what precisely genes were made of, and it was not until decades later, after the identity of the genetic material had been determined, that investigation of even the possibility of a system of differential gene expression became feasible. Finally, there were practical as well as ideological impediments to the study of both differential gene expression and hormone signaling.

In the seventies, the development of some key techniques (initially from virology and then from bacteriology) encouraged a renewal of interest in Morgan's puzzle. After a long incubation period, the search for something that could differentially control gene expression—what was dubbed a 'transcription factor'—rapidly became hurried and fierce. And as has already been mentioned, both hormones and the receptors they bind to are present in only very small quantities in certain tissues, making them incredibly difficult to find and work with. Because of the challenge produced by their scarcity, isolation of the first hormones as well as detection of the first hormone receptors are all considered highly significant scientific accomplishments. But the advent of cloning

technology in the late seventies and early eighties made it possible to amplify the amount of these substances available for research. And once cloning had allowed scientists to finally generate sufficient amounts of the molecules of interest, other techniques like sequencing could be applied.

By the early eighties, the combination of all three of these factors—realization of the import of Jensen’s work, the resurgence of interest in differential gene expression, and the development of techniques enabling the investigate the role of steroid hormones as candidate transcription factors—had transformed steroid hormone research into an incredibly competitive and potentially lucrative scientific endeavor. Whoever could identify the protein-coding genetic sequence of a steroid hormone receptor could get credit for both that and for the discovery of a transcription factor. Perhaps even more importantly, with knowledge of the protein-coding genetic sequence of a receptor, scientists could use cloning to generate myriad copies of the molecule. With both the hormone and its receptor identified and readily available, they could begin to investigate and manipulate the role of the hormone-receptor complex as a transcription factor, regulating the expression of genes.

It is hard to overstate the importance of the development of these practical techniques for the ideological advances that followed them. Perhaps the simplest way of indicating just how crucial they were is to refer again to the complexity of proteins,¹⁶ and to the delayed development of receptor theory that resulted from this complexity. Even with the development of x-ray crystallography, specifying the structure of thyroid hormone was difficult enough that the supposed “solution” mistakenly published in 1915

¹⁶ Detailed in chapter 3.

was not corrected until more than ten years later. And this is a molecule composed of only a few dozen atoms. The thyroid hormone receptors have primary sequences that are hundreds of amino acids long, and molecular weights over 50,000 Da. These are huge, unbelievably complicated molecules. They, and others like them, could not even begin to be understood until their encoding sequences were identified, and they could not be manufactured until scientists found out how to trick bacteria into making the receptors for them.

So, the development of cloning, sequencing, and other biochemical techniques revolutionized the field. The first hormone to be cloned was human growth hormone, in 1979 (Martial et al. 1979), and by 1985 the genes encoding two different hormone receptors, for glucocorticoid (Hollenberg et al. 1985) and estrogen (Walter et al. 1985), had been sequenced and cloned as well. Most important for this discussion is that the ability to clone and sequence steroid hormones and their receptors generated an unprecedented opportunity to characterize the as-yet-unknown mechanism of steroid hormone signaling.

As Sutherland's remark from 1971 reveals, Jensen's work tracking estrogen into the nucleus of cells implied that, though the signaling mechanism of some hormones might have been accounted for by the second messenger hypothesis, there were still parts of the story left untold. As Sutherland put it, "the primary action of the steroid hormones appears to follow an entirely different pattern." This statement also reflects the early assumption that steroid hormones in particular formed a special group whose mechanism of action needed to be worked out and added to Sutherland's second messenger hypothesis in order to characterize all cases of hormone signaling.

But in 1986, only a year after the initial sequencing of the glucocorticoid and estrogen receptors, sequence alignment studies detected a surprising alignment of sequences from the *erb-A* protooncogene family with the sequences for both glucocorticoid (Evans, Weinberger, Hollenberg, & Rosenfeld, 1986) and estrogen (Green et al. 1986). By the end of year it had been determined that what had been known as the *c-erb-A* gene sequence actually encoded the receptor for thyroid hormone (Weinberger, Giguere, Hollenberg, Rosenfeld, & Evans, 1986). However, thyroid hormone is not a steroid (all steroids are derivatives of cholesterol; thyroid hormone is a tyrosine-based hormone). Alignment of the sequences for all three of these receptors—glucocorticoid, estrogen, and thyroid—suggested that use of a signaling mechanism in which the receptor invaded the nucleus might not be restricted to just steroid hormone receptors, but might include other hormone receptors, such as the one for thyroid, as well.

The general term for any signaling molecule capable of activating a receptor is ‘ligand’. So, alignment of the sequences of the glucocorticoid, estrogen, and thyroid receptors in 1986 extended the group of known ligands for these receptors from just steroid hormones to also include thyroid hormone. To complicate matters even further, in 1987 additional sequences were aligned, these sequences were shown to encode novel receptors, and the ligands identified for these receptors were vitamin D (McDonnell et al. 1987) and retinoic acid, the derivative of retinol or vitamin A (Giguere, Yang, Segui, & Evans 1987; Petkovich, Brand, Krust, & Chambon 1987). This discovery—that some vitamins and their derivatives might function, hormone-like, as signaling molecules—was another startling revelation to come from this area of investigation. And one way of describing the evolving situation within this field, during the rapid developments of the

mid-eighties, is as one of surprising confrontation with a spreading pool of potential ligands for an ever-expanding group of receptors.

Early on in this process there was uncertainty about just what to call the group of molecules being discovered. In the presentation abstract of one of the initial talks on the subject of a potential "family of eukaryotic transcriptional regulatory factors" (Evans et. al 1986, p. 63), the terms 'paradigm', 'class', 'mechanism', and 'family' were all used. The term 'superfamily' was first used publicly in another talk later that year, entitled "Human Steroid-Receptors and erbA Protooncogene Products: Members of a New Superfamily of Enhancer Binding-Proteins" (Weinberger et al.1986). That did not settle the matter, however. Early in 1988, Giguere, Yang, Segui, and Evans published an article in *Nature*, titled "Identification of a New Class of Steroid-Hormone Receptors," in which the authors claimed that "the isolation of novel steroid hormone receptor cDNAs is a step towards identifying a new hormone response system" (p. 94). In this discussion the terms 'class', 'group', and 'system' were used. But only a few months later, what turned out to be the definitive review was published in *Science* (Evans, 1988), and the term 'superfamily' became firmly established.¹⁷

Section 6: Conclusion

The formation of the nuclear hormone receptor superfamily marked the beginning of what would become an extremely productive and enduring research program. Just to get a sense of the impact of this area of study, consider the fact that a search of Google

¹⁷ Despite this, the nuclear hormone receptor superfamily is also occasionally called 'the nuclear receptor superfamily', 'the nuclear hormone receptor family', or even just 'the nuclear receptor family'. More on the term 'superfamily' and its application to this group of proteins can be found in chapter 2 of this dissertation.

Scholar shows that Evans' (1988) review in *Science* has been cited 7,073 times.

According to the same forum, Watson and Crick's (1953) *Nature* paper announcing the structure of DNA has 9,153 citations.¹⁸

Throughout the subsequent decade after Evans' review, additional cloning and alignment studies as well as development of various screening techniques (such as low stringency hybridization screening) soon uncovered novel receptors, some with undetermined ligands (described in, e.g., Manglesdorf et al. 1995). Eventually, families within the superfamily of nuclear receptors were also established: the thyroid hormone receptor-like family, the retinoid X receptor-like family, the estrogen receptor-like family, the nerve growth factor IB-like family, the steroidogenic factor-like family, and the germ cell nuclear factor-like family. Within each of these families there are further groupings into subfamilies: for example, within the thyroid hormone receptor-like family, there is the thyroid hormone receptor subfamily, the retinoic acid receptor subfamily, the peroxisome proliferator-activated receptor subfamily, the Rev-ErbA subfamily, the retinoic acid receptor-related orphan receptor subfamily, the liver X receptor-like subfamily, and the Vitamin D receptor-like subfamily. Within each subfamily there are anywhere between one and four receptors.

Most of the receptors have several names, since each lab that works with them tends to pick a name and attempt to get it established in the literature, because this influences perceptions of ownership and discovery. Various attempts have been made to organize and standardize the nomenclature (for example, see Nuclear Receptors Nomenclature Committee 1999), but these have only partially succeeded. Regardless,

¹⁸ Both searches conducted on December 25, 2013.

today there are 48 commonly recognized members of the nuclear receptor superfamily in humans. There is one additional found in mice for a total of 49 in that species, and another one missing in rats for a total of 47 there. Among less related species the numbers fluctuate quite a bit more: there are 21 nuclear receptors found in *Drosophila melanogaster*, for example, but more than 270 in *Caenorhabditis elegans* (Zhang et al., 2004). Nuclear receptors have been found in all but only the animals, and no evolutionary precursors have been found.

Among the 48 nuclear receptors found in humans, however, at least 10 are still missing the ligand, or receptor activation component, of their ligand-receptor-gene triad. In a cute extension of the family metaphor, these nuclear receptors are called 'orphan' receptors. Of course, this means that the ligands for this group of nucleus-invading receptors are not restricted to steroids, or even to hormones at all. And so, though it is often still called 'the nuclear hormone receptor superfamily', the fundamental concept currently uniting the superfamily is that of a nuclear receptor, whether a receptor for a hormone or not.

REFERENCES

- Clark, A. J. 1926a. The reaction between acetyl choline and muscle cells. *Journal of Physiology* 61: 530–546.
- Clark, A. J. 1926b. The antagonism of acetyl choline by atropine. *Journal of Physiology* 61: 547–556.
- Edman, P. 1950. Method for determination of the amino acid sequence in peptides. *Acta Chemica Scandinavica* 4: 283–293.
- Evans, R. M. 1988. The steroid and thyroid hormone receptor superfamily. *Science*, 240, 889–895.
- Evans, R. M., Weinberger, C., Hollenberg, S., & Rosenfeld, M. G. 1986. Structure and expression of human glucocorticoid receptor cDNA: A trans-acting factor related to the c-erbA protooncogene family. *Journal of Molecular and Cellular Biology*, 5, 63.
- Fausto-Sterling, A. 2000. *Sexing the body: Gender politics and the construction of sexual identity*. New York: Basic Books.
- Fruton, J. S. 1985. Contrasts in scientific style. Emil Fischer and Franz Hofmeister: Their research groups and their theory of protein structure. *Proceedings of the American Philosophical Society* 129: 313–370.
- Gaddum, J. H. 1926. The action of adrenaline and ergotamine on the uterus of the rabbit. *Journal of Physiology* 61: 141–150.
- Giguere, V., Yang, N., Segui, P., & Evans, R. M. 1988. Identification of a new class of steroid-hormone receptors. *Nature* 331: 91–94.
- Green, S., Walter, P., Kumar, V., Krust, A., Bornert, J.-A., Argos, P., & Chambon, P. 1986. Human oestrogen receptor cDNA: Sequence, expression, and homology to v-erb-A. *Nature* 320: 134–139.
- Harrington, C. R., & Barger, G. 1927. Chemistry of thyroxine: Constitution and synthesis of thyroxine. *Biochemical Journal* 21: 169–183.
- Hartley, H. 1951. Letter to *Nature* 168: 244.
- Hill, A. V. 1909. The mode of action of nicotine and curari, determined by the form of the contraction curve and the method of temperature coefficients. *Journal of Physiology* 39: 361–373.

- Hollenberg, S. M., Weinberger, C., Ong, E. S., Cerelli, G., Oro, A., Lebo, R., Thompson, E. B., Rosenfeld, M. G., & Evans, R. M. 1985. Primary structure and expression of a functional human glucocorticoid receptor of cDNA. *Nature* 318: 635–641.
- Jensen, E. V. & Jacobson, H. I. 1962. Basic guides to the mechanism of estrogen action. *Recent Progress in Hormone Research* 18: 387–414.
- Kenakin, T. 2004. Principles: Receptor theory in pharmacology. *TRENDS in Pharmacological Science* 25: 186–192.
- Kendall, E. C. 1915. The isolation in crystalline form of the compound containing iodine, which occurs in the thyroid. *Journal of the American Medical Association* LXIV: 2042–2043.
- Langley, J. N. 1905. On the reaction of cells and of nerve-endings to certain poisons, chiefly as regards the reaction of striated muscle to nicotine and to curare. *Journal of Physiology* 33:374–413.
- Langmuir, I. 1918. The adsorption of gases on plane surfaces of glass, mica and platinum. *Journal of the American Chemical Society* 40: 1361–1403.
- Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., Blumberg, B., Kastner, P., Mark, M., Chambon, P., & Evans, R. M. 1995. The nuclear receptor superfamily: The second decade. *Cell* 83: 835–839.
- Martial, J. A., Hallewell, R. A., Baxter, J. D., & Goodman, H. M. Human growth hormone: Complementary DNA cloning and expression. *Science* 205: 602–607.
- McDonnell, D. P., Mangelsdorf, D. J., Pike, J. W., Haussler, M. R., & O'Malley, B. W. Molecular cloning of complementary DNA encoding the avian receptor for vitamin D. *Science* 235: 1214–1217.
- Moore, D. D. 2012. A conversation with Elwood Jensen. *Annual Review of Physiology* 74: 1–11.
- Nuclear Receptors Naming Committee. 1999. A unified nomenclature system for the nuclear receptor superfamily. *Cell* 97: 161–163.
- Olby, R. 1974. *The path to the double helix: The discovery of DNA*. New York: Dover.
- Oudshoorn, N. 1990. On the making of sex hormones: Research materials and the production of knowledge. *Social Studies of Science* 20: 5–33.
- Oxford English Dictionary Online (December 2012). "scientist, n.". Oxford University Press. 9 January 2013:

<http://www.oed.com/view/Entry/172698?redirectedFrom=scientist&>

- Petkovich, M., Brand, N. J., Krust, A., & Chambon, P. 1987. A human retinoic acid receptor which belongs to the family of nuclear receptors. *Nature* 330: 444–450.
- Rang, H. P. 2006. The receptor concept: Pharmacology's big idea. *British Journal of Pharmacology* 147: S9–S16.
- Sanger, F., Nicklen, S., & Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of the Sciences* 74: 5463–5467.
- Sutherland, E. W. 1972. Studies on the mechanism of hormone action. *Science* 177: 401–408.
- Sutherland, E. W. & Robison, G. A. 1966. Role of cyclic-3' 5'-AMP in responses to catecholamines and other hormones. *Pharmacological Reviews* 18: 145.
- Tanford, C. & Reynolds, J. 2001. *Nature's robots: A history of proteins*. Oxford: Oxford University Press.
- Walter, P., Green, S., Greene, G., Krust, A., Bornert, J.-M., Jeltsch, J.-M., Staub, A., Jensen, E. V., Scrace, G., Waterfield, M., & Chambon, P. 1985. Cloning of the human estrogen receptor cDNA. *Proceedings of the National Academy of the Sciences* 82: 7889–7893.
- Watson, J. D. & Crick, F. H. C. 1953. Molecular structure of nucleic acids. *Nature* 171: 737–738.
- Weinberger, C., Giguere, V., Hollenberg, S., Rosenfeld, & Evans, R. M. 1986. Human steroid-receptors and erbA protooncogene products: Members of a new superfamily of enhancer binding-proteins. *Cold Spring Harbor Symposia on Quantitative Biology* 51: 759–772.
- Zhang, Z., Burch, P. E., Cooney, A. J., Lanz, R. B., Pereira, F. A., Jiaqian, W., Gibbs, R. A., Weinstock, G., & Wheeler, D. A. 2004. Genomic analysis of the nuclear receptor family: New insights into structure, regulation, and evolution from the rat genome. *Genome Research* 14: 580–590.

APPENDIX B

I am now convinced that this kind of direct examination of scientists at work should be extended and should be encouraged by scientists themselves in our own best interest, and in the best interest of society. Science, in general, generates too much hope and too much fear, and the history of the relationship of scientists and nonscientists is fraught with passions, sudden bursts of enthusiasm, and equally sudden fits of panic. If the public could be helped to understand how scientific knowledge is generated and could understand that it is comprehensible and no more extraordinary than any other field of endeavor, they would not expect more of scientists than they are capable of delivering, nor would they fear scientists as much as they do. This would clarify not only the social position of scientists in society, but also the public understanding of the substance of science, of scientific pursuits and of the creation of scientific knowledge.

Jonas Salk's Introduction to *Laboratory Life* (1979)

Abstract: This appendix documents a laboratory study conducted à la Bruno Latour and Steve Woolgar's *Laboratory Life* (1979). Although many science studies scholars have analyzed Latour and Woolgar's work, now a classic in the field, prior to this study the experiment of *Laboratory Life* had not yet been replicated. Given the authors' stated intent—to be reflexive in applying the methods of science to their own work while studying said methods—it seemed a shame not to subject their work to this particularly scientific ideal of reproducibility. And so I conducted the following study—thirty years later, and once again at the Salk Institute. As much as possible, I repeated Latour and Woolgar's sociological methods and reapplied their conceptualizations. I also did my best to mimic the authors' tone and style. Interestingly, a surprising amount of Latour and Woolgar's work turned out to be remarkably salient today.

Section 1: Introduction

In October of 1975, Bruno Latour began nearly two years of observation within Roger Guillemin's neuroendocrinology laboratory at the Salk Institute for Biological Studies in La Jolla, California. In June of 1979, *Laboratory Life: The Social*

Construction of Scientific Facts, was published. The authors, sociologists Bruno Latour and Steve Woolgar,¹ presented an analysis of everyday scientific practice based on the observations Latour made during his time at the laboratory.

Reviews were favorable, although criticism increased the farther the reviewer's area of specialization was from the authors'. Among sociologists, there was general consensus that the book was a great achievement: Roger Krohn, writing in *Contemporary Sociology*, declared that he "highly recommended" it (1981, p.434), and Nicholas Mullins said in *Science, Technology, and Human Values* that "this is an important book. It should be read by everyone doing social studies of science, science policy or sociology of science" (1980, p.55). Historians were also enthusiastic, although a bit more reserved in their praise: Alan Irwin wrote in the *British Journal for the History of Science* that "this book provides a reasonable introduction to recent developments in a rapidly growing area of sociology of science" (1982, p.209), and Donna Haraway, who reviewed the book for *Isis*, concluded that "Latour and Woolgar's achievement is an exquisitely detailed story, rather than another pronouncement of belief in social construction of facts. ...For those who hope for more from a material analysis of scientific construction, the value of the project is not yet realized" (1980, p.489).

Perhaps unsurprisingly, philosophers resisted the constructivism advanced in Latour and Woolgar's piece: Verloren van Themaat wrote in the *Journal for General Philosophy of Science* that "in spite of the merits of this book as a sociological study of a

¹ Although Jonas Salk, in his introduction to *Laboratory Life*, calls Latour "a young French philosopher" and Woolgar "an English sociologist" (1979/1986, p.12), both are now more commonly identified as sociologists. In fact, in the postscript to *Laboratory Life*, Latour describes himself as initially an epistemologist who, in the course of his investigation at the Salk Institute, became a sociologist (1979/1986, p.274).

laboratory, its plea for the replacement of a realistic by a constructivist epistemology is not convincing” (1982, p.170). And David Bearman, reviewing the book for *Science*, wrote that “unfortunately the work as a whole will be difficult for outsiders to the sociology or philosophy of science to penetrate. ...As a result, a major contribution to the specialty literature that deserves and could reward a broader reading audience will, I’m afraid, fail to attract or excite as it could” (1979, p.825).

Fortunately for the authors of *Laboratory Life*, Bearman’s fears went unrealized. A second edition was printed in 1986, and by the time *Isis* reviewed it again, H. M. Collins could declare that “*Laboratory Life* was the first and, deservedly, the most influential book-length account of day-to-day work in a single laboratory setting” (1988, p.148). Collins argued that, although the book undoubtedly had flaws, it also contained some intriguing and novel insights that could contribute to the conceptualization of scientific practice. As he put it: “*Laboratory Life* succeeds and will continue to succeed, and to win friends and allies, because it contains good, persuasive ideas” (p.149).

In the twenty or so years since its republication, *Laboratory Life* has continued to succeed. It has been cited more than 6,000 times² and is routinely assigned reading on syllabi for sociology, communication, history, and philosophy of science courses, and especially, science and technology studies courses. It is a premier exemplar of an ethnography; it is the paradigmatic laboratory study.

Latour and Woolgar called their approach "an anthropology of science" (1979/1986, p.27), but now it is often referred to as 'ethnographic methodology'. The text of *Laboratory Life* begins with an excerpt from Latour’s field notes: a dispassionate

² According to Google Scholar in June 2011.

description of what is going on in the lab at some place and time on some random day.

Each entry in the notes begins with the time, and then there is a short description of what some particular lab member is doing. These notes are followed by a summary of life in the laboratory: a description of what scientific activity in this lab consists of, on a daily basis. Later in the chapter, the authors describe their approach:

Scientists in our laboratory constitute a tribe whose daily manipulation and production of objects is in danger of being misunderstood, if accorded the high status with which its outputs are sometimes greeted by the outside world. There are, as far as we know, no a priori reasons for supposing that scientists' practice is any more rational than that of outsiders. We shall therefore attempt to make the activities of the laboratory seem as strange as possible in order not to take too much for granted. (Latour & Woolgar, 1979/1986, pp.29–30)

In short, Latour and Woolgar chose not to take the scientists' reports of their own activities for granted. Instead, the authors relied on observation and their own conceptual framework in order to characterize what they observed.

But there have been fewer ethnographic studies of scientific practice than one might expect, given the enduring reputation and influence of *Laboratory Life*. Karin Knorr-Cetina's (1981) *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science* and Michael Lynch's (1985) *Art and Artifact: A Study of Shop Work and Shop Talk in a Research Laboratory* are key exemplars. The case study recounted in this appendix attempts to continue in this clearly foundational, but somewhat neglected, science studies tradition of conducting laboratory studies.

In summary, this case study has four main goals: to participate in the ethnographic tradition; to repeat Latour and Woolgar's experiment; to understand and evaluate the ideas of *Laboratory Life*; and to characterize more recent scientific practice. These goals

are realized in a series of three steps. The first step is to describe some recent scientific practice in a particular laboratory—Principal Investigator X’s gene expression laboratory at the Salk Institute (where Roger Guillemin’s lab, which Latour studied, was as well).³ The second step is analyzing this practice using key ideas from Latour and Woolgar’s *Laboratory Life*, such as daily activity, literary inscription, origin stories, statement types, cycles of credit, fact construction, and the creation of order out of disorder. The third and final step is evaluating how various differences—between the time, place, and subject of Latour’s study and the more current observational experience—will require further conceptual investigation in order to generate a thorough and accurate characterization of this recent scientific practice.

The appendix is split into sections: this introduction; two middle parts (which correspond with the first two steps described in the previous paragraph); and a conclusion (which takes the third and final step). After the introduction, the next step in the process is to offer a thorough description of the laboratory being observed. Because the intent is to compare more recent experience with Latour’s, an attempt has been made to replicate some of his methods of observation and description. So the next section of the appendix presents a detailed look at the laboratory that is the subject of this analysis. A judicious attempt has been made to present the account with an element of Latourian ethnography: without assuming familiarity on the part of either the observer or the reader. This style of description results in an extensive sense of the time, place, and practice in the laboratory being analyzed, and this context provides a helpful familiarity during the later stages of

³ At the time of original publication, Latour and Woolgar refrained from identifying the particular lab in which they conducted their ethnography, and all references to individuals in said lab were blinded. I replicate that practice here.

conceptualization.

The process of conceptualization begins in the third section of the paper, when elements of the characterization of scientific practice provided by Latour and Woolgar in *Laboratory Life* are applied to the more recent observations. The fourth and final section of the appendix explores some aspects of the recently observed scientific practice that are not accounted for within the conceptual scheme of *Laboratory Life*, and reviews the aims of the study presented in this introduction, assessing the matter of their accomplishment as well as the overall value of both this particular and this general kind of study. The hope is that, by providing a detailed description of the lab and the activity within it, a variety of interesting and current scientific practices will be brought to light. Some of them might already be comprehensible or expected, within Latour and Woolgar's characterization of scientific activity. But those that are not will be candidates for further examination.

To conclude this introduction: in his “Two Cultures or One?” Robert Westman (1994) reexamines Kuhn’s *The Copernican Revolution* (1957) and considers how the historical account provided there contributed as a case study to the characterization of social constructivism offered in *The Structure of Scientific Revolutions* (1962/1996). Westman compares Kuhn’s understanding of the Copernican Revolution with his own, at times “providing alternative interpretations that subsequent historical scholarship has brought to light” (1994, p.81). This juxtaposition allows the author to offer some crucial insight into the way that the different conceptions of the specific historical event contributed to the accuracy and completeness of Kuhn’s characterization of scientific revolution in general.

Similarly, this appendix juxtaposes Latour's observations of laboratory experience with a more recent set of observations made in another laboratory, in order to evaluate the accuracy and completeness of Latour and Woolgar's characterization of scientific practice, and finally to offer a revised and updated characterization of laboratory life as practiced today. There are differences between the two sets of experiences in the lab, just as there are differences between Kuhn and Westman's accounts of the Copernican Revolution. Unsurprisingly, there will also be differences between the two characterizations of scientific practice, just as the portrayal of scientific revolutions in Kuhn's *Structure* differs from what Westman concludes based on his comparison in "Two Cultures or One?"⁴

But first, some similarities with which to begin. The initial task of this exploration is to describe some recent scientific practice in a particular laboratory: Principal Investigator X's gene expression lab at the Salk Institute (hereafter known as GEL-X). Gene expression is, like neuroendocrinology, a scientific field within molecular biology. And, like Roger Guillemin's lab where Latour observed, Principal Investigator X's lab is also at the Salk. So, it is worth describing the context in which both the more recent and Latour's past observations of scientific activity were conducted.

Section 2: GEL-X

The Salk Institute for Biological Studies was founded by Jonas Salk, designed by Louis Kahn, funded by the March of Dimes, and built in La Jolla, California, with

⁴ Westman credits C. P. Snow for the "two cultures" meme. Interestingly, Salk also mentions the phrase in his introduction to *Laboratory Life*: "I have doubts about this way of thinking and, in my own work, find many details which do not fit this picture, but I am always stimulated by attempts to show that the two 'cultures' are, in fact, only one" (1979/1986, p.13).

scientific research commencing in its first laboratory in 1963. The Salk has two main buildings perched on a cliff above Black's Beach and the Pacific Ocean. They run parallel to each other east to west and are referred to as the North Building and the South Building. There is a stone courtyard (made of travertine marble) between them. Standing at the east end looking west, there is a spectacular view of a green canyon, the water, and above this the sky, all framed by the tiered buildings of the Institute. East of the courtyard, behind the observer, is a newer building that houses the administration, conference halls, and some lab space. It is called the East Building.

Below is a figure composed of labeled pictures of the Salk campus:

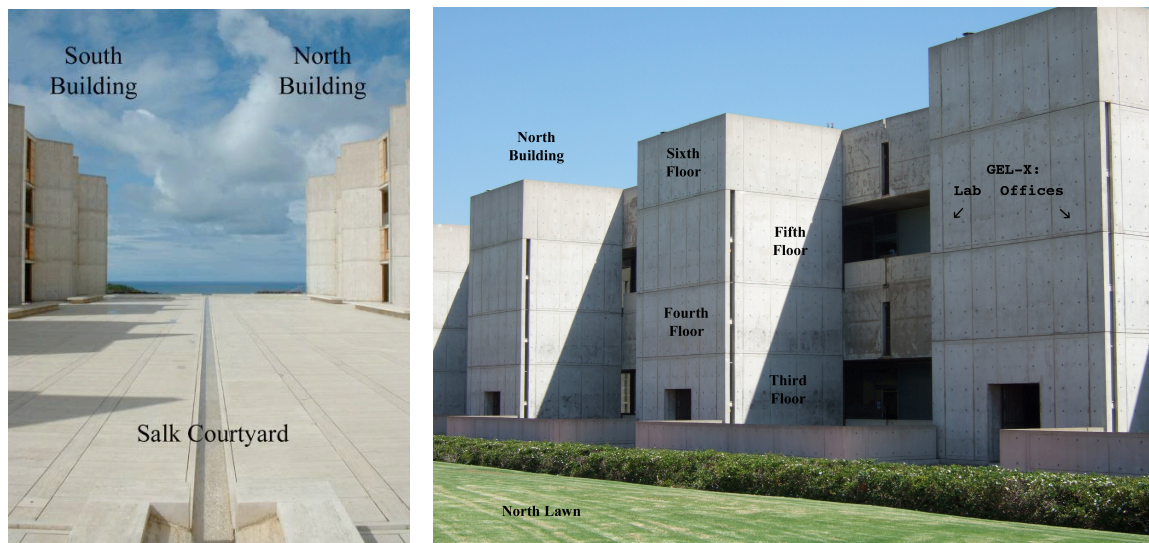


Figure B.1: On the left, the Salk Courtyard, facing the Pacific Ocean; on the right, a view of the North Building from the lawn (pictures by author).

Most of the Salk Institute's laboratories occupy the first, third, and fifth floors of the North and South Buildings. The second, fourth, and sixth floors are pipe space, crammed with fridges, freezers, shakers, incubators, and tanks for cold storage. The third floor is actually the one that is level with the lawn around the Institute. The second and first floors lie below the ground. At the first floor the two buildings are connected by several

tunnels. There are also courtyards outside the laboratories to let in light from above. Walking along the east tunnel there are cupboards, fridges, freezers, and kiosks labeled with names like Invitrogen, New England Biolabs, Ambion, and Fisher. Labs can send members down to a kiosk to enter a code and immediately retrieve required supplies. But the bulk of supplies are ordered via computer and delivered directly to the lab.

Principal Investigator X's gene expression lab (GEL-X) is on the fifth floor of the North Building. The 'GEL' in GEL-X stands for Gene Expression Laboratory, and the letter which is here 'X' would normally indicate the first letter of the last name of the lab's principal investigator (PI). There are, in total, four gene expression labs at the Salk (GEL-E, GEL-B, GEL-P, and GEL-W), each doing work in the general field of gene expression but led by different principal investigators (whose last names begin with E, B, P, and W, respectively). All the GEL labs are located somewhat together on the fifth floor of the North Building.

There are wood-paneled offices for the PIs of the GEL labs located along the westernmost wall of the building, separated from the lab space by an outdoor hallway. Facing eastward and entering the main door to the laboratory, there is first a computer room and a kitchen, and then a large space filled with benches and equipment. Desks (with spots for up to 22 postdoctoral researchers and graduate students) stretch along the southern wall for the entire length of the laboratory. The lab benches (nine in total, fitting approximately four researchers each, one pair on each side of the bench) span the width of the lab space, perpendicular to the desks. The countertops are cluttered and the shelves crammed with vials, tubes, jars, and other supplies. The opposite wall, facing the desks, also holds packed shelves and chest freezers.

The wall is broken by gaps that lead into another lab space. These passageways are also crammed with more equipment and workstations for specific procedures. The first has polymerase chain reaction (PCR) machines on one side and the radioactive station on the other. Often the stations are so packed that there is no actual space open on the countertop. The next gap is a cold room. It is a walk-in refrigerator with cupboards for materials and shakers for more procedures. After this is the chemical weigh station. There are big containers of crystals and powders. There are mixers and a pH detector. Many of the solutions made in the lab are made here. Past this is the tissue culture room. This is a room with several sterile hoods, some incubators and cold cases. After this there is another cold room, and then the back wall of the main lab space.

A row of centrifuges stretches to the left, and a wall of lab materials is on the right. There are carts with empty cages in front of the wall. These are brought up from what is colloquially called the mouse house. This is where most of the animals in the Institute are kept, in large rooms full of cages, and supervised by members of the Animal Research and Development department (ARD). Once brought up to the lab, animals cannot be returned for fear they will infect the rest of the population with germs from outside the mouse house. When the cages are empty they are taken from the lab and placed in a giant rack on the bottom floor of the institute. Carcasses are also taken downstairs in plastic bags and left in freezers for disposal. Workers from ARD take the cages, clean and sterilize them in a giant underground autoclave, and then reuse them. The carcasses are destroyed.

Beyond the main lab space are more benches, part of a neighboring lab. Then there is another room with the robots for quantitative real-time polymerase chain reaction

(Q-PCR), as well as a bench and a hood. Past this is another room with old equipment, supplies, and a few hoods for viral work. This is about where the lab's territory ends and another lab begins, although the lab has additional storage and work rooms scattered around the Institute. Altogether the main lab space of GEL-X occupies about one half of the fifth floor, in the southwest quadrant of the North Building.

In *Laboratory Life*, Latour and Woolgar include a detailed diagram of Guillemin's laboratory, which depicts the lab space as a sort of cellular entity with certain inputs and outputs. This is one of the most creative figures from the text, and well worth reproducing here:

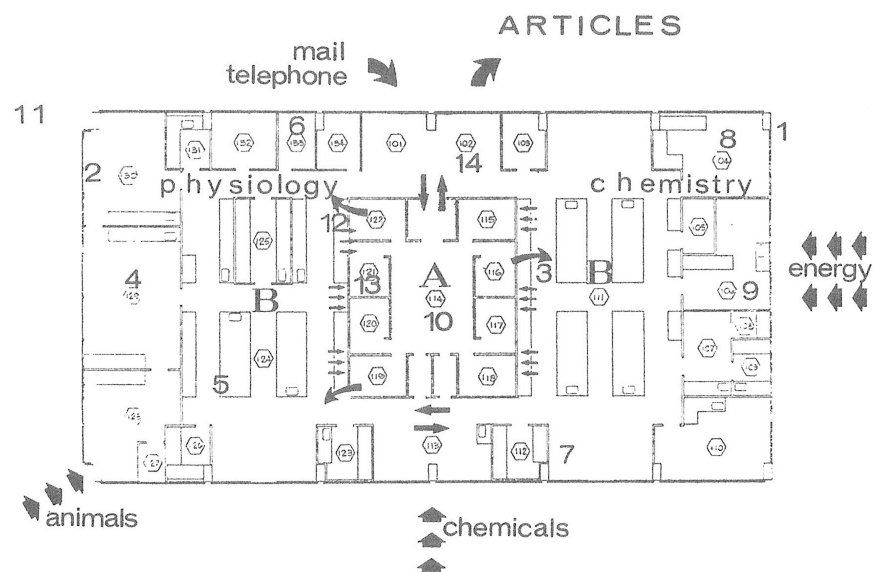


Figure B.2: Diagram of the interior of Guillemin's laboratory (from Latour & Woolgar 1979/1966, p.46).

Latour and Woolgar note the various inputs as (proceeding counterclockwise from the top): mail, telephone, animals, chemicals, and energy. The sole output is articles. A blueprint can also be provided for GEL-X, with inputs and outputs identified:

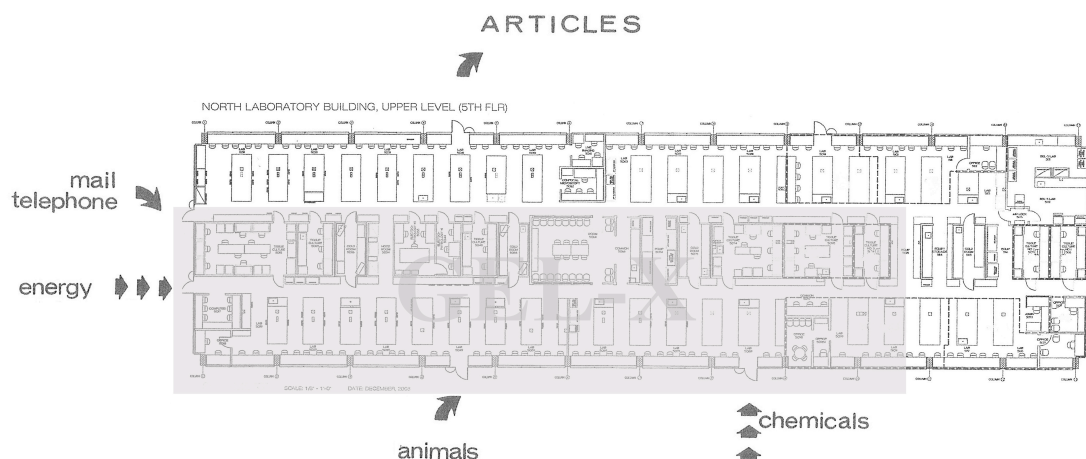


Figure B.3: Diagram of the interior of the GEL-X laboratory (from blueprints for construction).

Some of the inputs and outputs identified in *Laboratory Life* are preserved here, but some changes have also been made.

One change to the diagram is the fact that 'people' and 'money' have been added as inputs. Neither of these integral components were included in the original diagram, though Latour and Woolgar did acknowledge each in the text of *Laboratory Life*. For example, when discussing the diagram the authors discuss as another input "the technicians and doctors who comprised the work force" (1979/1966, p.52). And in the Observer's Story (discussed in the next section), there is mention of funds: "cheques of taxpayers' money arrive periodically, courtesy of the N.I.H., to pay bills and salaries" (p.16). But neither are explicitly included in the diagram—perhaps the cheques were considered part of the mail input? Regardless, these omissions have been rectified in the newer diagram. And in keeping with the previous description of GEL-X, the input 'animals' has been specified to 'animals in cages', and the input 'chemicals' has been expanded to 'biological & chemical compounds' because both kinds of compounds are

critical inputs. Finally, the input 'energy' has been replaced by 'electricity', since other forms of (traditionally called) energy rarely enter the lab.⁵ In total, the inputs include (again, proceeding counterclockwise): mail & telephone, people, money, animals in cages, biological & chemical compounds, and electricity.

Perhaps the most obvious difference, however, is that new diagram includes many more outputs than Latour and Woolgar's. This is simply because far more than just articles exit the modern laboratory. First, the statements which Latour and Woolgar identified as the primary focus of scientific practice are not only made in literary form, as articles; they are also made in spoken form, as conference presentations. Second, the only inputs which are not also outputs are money and electricity. Members of the lab send as well as receive mail; they make as well as receive telephone calls. These people join the lab but they also leave it, often to form their own gene expression laboratory at another academic institution. Both the animals and the cages that enter the lab exit it as well, although the animals have inevitably switched from alive to dead. Chemical compounds and biological samples (most often plasmids) are shipped to other academic labs for further research. Finally, GEL-X produces quite a bit of waste, and this byproduct of the daily activity of science should not be overlooked. There are even different kinds of waste, each with their own signage and methods of disposal: (normal) trash, food waste (from the kitchen), recyclables, biomedical waste (such as used petri dishes), sharps (like needles and broken glass), hazardous waste (such as syringes with viral solutions), and radioactive waste (like pipette tips used to extract an isotope). To

⁵ For instance, food for either animals or persons is not allowed in the lab. Animals are brought up to the lab only to die—once removed from it, they cannot be returned to the mouse house. And people are technically not permitted to eat at their benches or desks—all food must remain in the kitchen. Of course, both of these rules are occasionally violated.

summarize, as the figure shows, total outputs include: articles & presentations, mail & telephone, people, carcasses & cages, biological & chemical compounds, and waste.

The diagram also serves to emphasize how big GEL-X really is. As can be seen in the figure above, GEL-X takes up most of the 5th floor of the North Building, but there are still four other labs on the floor. Three of them run in sequence along the north side and the fourth fits into the leftover pocket at the east end. Technically, that isn't even the entirety of the laboratory: GEL-X is more disparate than is suggested by the above figure. In other words, the lab constitution also includes some satellite locations. There are several office areas in different locations throughout the building. This includes the office of the principal investigator. One set of office suites exists below and west of the actual lab space (on the fourth rather than fifth floor of the building). There are also segregated areas in which certain kinds of scientific activity crucial to the lab are performed. Sequencing is now done in its own area in the Institute, and some animal procedures are performed in the areas where they are also housed. This fragmentation represents a slight variation in lab composition from the time of Latour and Woolgar's study.

Another structural difference can be found among the changes in personnel organization since Latour and Woolgar's time. In *Laboratory Life*, the authors describe "an almost perfect administrative pyramid" (Latour & Woolgar, 1979/1986, pg. 216–7) comprised of technicians, supertechs, and professional researchers for Guillemin's lab. Although the members of Principal Investigator X's lab can still be grouped into technicians, supertechs, and professional researchers, the corresponding structure is only loosely pyramidal. Now researchers share technicians, and crucial aspects of their

employment, duties, and supervision are the purview of the lab manager rather than the researchers themselves. So there is a marked split between two groups: transient but investigative members of the lab, like the postdocs and graduate students, and the more permanent and less academic employees of the lab, such as the administrators and technicians. A comparison of these hierarchies can be seen below:

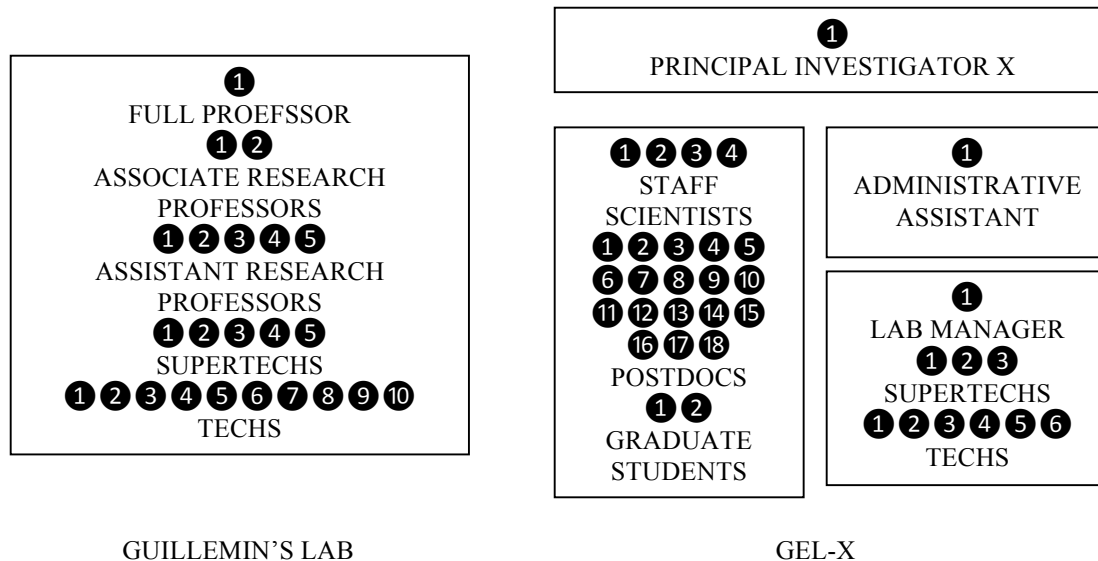


Figure B.4: A comparison of laboratory hierarchies.

On the left, the personnel hierarchy of Guillemin's lab, as described by Latour and Woolgar in *Laboratory Life*. There was one full professor, two associate research professors, five assistant research professors, five supertechs, and ten technicians for a total of twenty-three lab members. On the right, the personnel hierarchy of GEL-X, which is more of an interactive network than a pyramid. Coincidentally, GEL-X was founded in 1979, the same year that *Laboratory Life* was published, and at that time had only two members: Principal Investigator X and one technician. But thirty years later, in the summer of 2009, there was one principal investigator (still Principal Investigator X), four staff scientists, one administrative assistant, one lab manager (the original

technician), eighteen postdocs, two graduate students, three supertechs, and six technicians, for a total of thirty-six lab members.

In the modern structure, there is a sort of division of the workforce into two separate hierarchies: the academics (staff scientists, postdocs, and graduate students), chiefly responsible for the creative aspects of the lab's scientific activity, and the technicians (management, supertechs, and techs), primarily responsible for the execution and management of the lab's scientific activity. Obviously, there is some overlap: postdocs conduct some of their own experiments, in addition to their techs, and sometimes technicians are acknowledged as authors in publications because their expertise is viewed as significant enough to have contributed to the intellectual achievement represented by a manuscript. But in general, the one group dictates the scientific agenda (by reading the relevant literature, designing experiments, attending data club and scientific conferences, and writing and discussing grants and manuscripts), while the other group assists with the scientific agenda (by conducting the dictated experiments) as well as maintains the environment necessary for scientific practice (by ordering supplies, organizing and cleaning the laboratory, maintaining equipment, making buffers, taking care of animals, sending and receiving samples). In GEL-X, Principal Investigator X guides the scientific agenda (by reviewing project ideas and design, overseeing grant applications and manuscript submissions, and networking with related labs), while the lab manager, although subordinate to the principal investigator, is the one who really runs the laboratory on a daily basis, dealing with the employees, equipment, protocols, supplies, and budget. Finally, there is the administrative assistant who handles the principal investigator's schedule and correspondence, and doesn't really

fit into either hierarchy, but is still instrumental in lab function. Postdocs and technicians must contact the administrative assistant in order to get time with the principal investigator.

The above paragraphs detail the location, materials, layout, and personnel of GEL-X. The intent of this discussion was to provide a rich description of the context in which everyday scientific practice occurs in this particular laboratory. Despite the sharing of location both at the Salk Institute and within the field of molecular biology, various differences between GEL-X and Guillemin's lab are already apparent. This is not surprising given the passage of time and change in the focus of the research. Some of the similarities and differences can be conceptualized with the help of the characterization of scientific practice offered by Latour and Woolgar in *Laboratory Life*. The next section examines the conceptual framework offered by Latour and Woolgar, applying some of the authors' ideas to aspects of the more recently observed scientific activity, in order to better understand both the observations of scientific practice and the characterization of it.

Section 3: Applying the Ideas of *Laboratory Life* to GEL-X

Latour and Woolgar's conceptual framework includes a host of interesting ideas: ethnography, daily activity, literary inscription, hybridization, origin stories, statement types, cycles of credit, fact construction, the creation of order out of disorder, and reflexivity, to name a few. This section applies these older concepts to the newer observation of scientific practice.

Sub-Section 3.1: The Daily Activity of Science

As has already been mentioned in the introductory section of this appendix, the authors of *Laboratory Life* adopt what they call ‘an anthropological perspective’ to their study of science. Unsurprisingly, the ethnographic method begins with making observations. So *Laboratory Life* begins with an excerpt from Latour’s field notes—a detailed description of what he observed occurring in everyday scientific practice. The notes provide a sense of what it might be like to be a fly on the wall of Guillemin’s lab in the seventies:

6 mins. 20 secs. Bill comes from the chemistry section and gives Spencer a thin vial: “Here are your two hundred micrograms, remember to put this code number on the book,” and he points to the label. He leaves the room.

Long silence. The library is empty. Some write in their offices, some work by windows in the brightly lit bench space. The staccato noise of typewriting can be heard from the lobby.

9 mins. Julius comes in eating an apple and perusing a copy of *Nature*.

9 mins. 10 secs. Julie comes in from the chemistry section, sits down on the table, unfolds the computer sheets she was carrying, and begins to fill in a sheet of paper. Spencer emerges from his office, looks over her shoulder and says: “hmm, looks nice.” He then disappears into John’s office with a few pages of draft.

9 mins. 20 secs. A secretary comes in from the lobby and places a newly typed draft on John’s desk. She and John briefly exchange remarks about new deadlines.

9 mins. 30 secs. Immediately following her, Rose, the inventory assistant, arrives to tell John that a device he wants to buy will cost three hundred dollars. They talk in John’s office and laugh. She leaves.

(Latour & Woolgar, 1979/1986, pp.15–16)

Much of this snapshot of everyday events in a laboratory might be replicated today. An

excerpt from field notes gathered on an unremarkable morning in GEL-X follows:

9:09am: Two postdocs enter while conversing. One says: “So, you’re working on macrophages too?” The other responds inaudibly. The first says something back and then asks “How many can you fit on a ten [inaudible]meter dish?” They pass out of earshot. The telephone rings for the first time today, but only once.

9:28am: Several more lab members enter the lab. Some go to their benches and desks. One enters the far cold room. Another does something in the kitchen. Water is running. Glasses clink. A grad student pops her gum while reading. A technician looks in a drawer while holding a cell phone to her ear. She is wearing a lab coat.

9:30am: The phone rings repeatedly in the lab manager’s office.

9:39am: By now there are 11 post docs in the lab. Nine sit at their desks. Two are standing in front of their benches.

9:50am: Another post doc enters the lab, and puts some personal equipment down on the desk next to the observer. They exchange greetings. The grad student says “Hey [post doc’s name], do you think that today I could get a little tiny sample of [certain compound]?” The post doc replies, “No problem.” He unpacks a laptop and then heads to the kitchen. His computer emits a sound that indicates start-up. The grad student gets up from her desk and joins two post docs in the kitchen. Two technicians talk at their benches.

There are a few apparent differences: obviously, technology has changed and laptops have replaced typewriters. A piece of equipment is more likely to cost \$30,000 rather than \$300. Finally, the scientists are just as likely to be women as men, and the technicians and assistants are just as likely to be men as women. There is no longer a high probability that most roles will correlate strongly with gender.⁶ Regardless of these differences, there are also apparent similarities. The scientists still read *Nature*, eat in and around the lab, work on papers at their desks, conduct experiments at the bench,

⁶ It should also be apparent that, in the field notes from GEL-X, names have been replaced with role identifiers. This was done in order to avoid contributing to any impressions of gender or ethnicity. In contrast, Latour's field notes do contain names. These were changed in order to preserve anonymity, but they still give some impressions of gender and ethnicity that might better be either avoided or, if mentioned, then discussed explicitly.

exchange data and materials with each other, and converse informally about many aspects of scientific practice, as the notes show.

The excerpt from Latour's field notes at the beginning of *Laboratory Life* is immediately followed by a synthesis of the entire mass of notes gathered throughout two years of observation in Guillemin's lab. The authors call this summary of everyday scientific practice the 'Observer's Story'. Because it provides many points of comparison with today's laboratory, it is reproduced in its entirety below:

Every morning, workers walk into the laboratory carrying their lunches in brown paper bags. Technicians immediately begin preparing assays, setting up surgical tables and weighing chemicals. They harvest data from counters which have been working overnight. Secretaries sit at typewriters and begin recorrecting manuscripts which are inevitably late for their publication deadlines. The staff, some of whom have arrived earlier, enter the office area one by one and briefly exchange information of what is to be done during the day. After a while they leave for their benches. Caretakers and other workers deliver shipments of animals, fresh chemicals and piles of mail. The total work effort is said to be guided by an invisible field, or more particularly, by a puzzle, the nature of which has already been decided upon and which may be solved today. Both the buildings in which these people work and their careers are safeguarded by the Institute. Thus, cheques of taxpayers' money arrive periodically, by courtesy of the NIH, to pay bills and salaries. Future lectures and meetings are at the forefront of people's minds. Every ten minutes or so, there is a telephone call for one of the staff from a colleague, an editor, or some official. There are conversations, discussions, and arguments at the benches: "Why don't you try that?" Diagrams are scribbled on blackboards. Large numbers of computers spill out masses of print-out. Lengthy data sheets accumulate on desks next to copies of articles scribbled on by colleagues.

By the end of the day, mail has been dispatched together with manuscripts, preprints, and samples of rare and expensive substances packed in dry ice. Technicians leave. The atmosphere becomes more relaxed and nobody runs anymore. There are jokes in the lobby. One thousand dollars has been spent today. A few slides, like Chinese idiograms, have been added to the stockpile; one character has been deciphered, a miniscule, invisible increment. Minute hints have dawned. One or two statements have seen their credibility increase (or decrease) a

few points, rather like the daily Dow Jones Industrial Average. Perhaps most of today's experiments were bungled, or are leading their proponents up a blind alley. Perhaps a few ideas have become knotted together more tightly.

A Philippino cleaner wipes the floor and empties the trash cans. It has been a normal working day. Now the place is empty, except for the lone figure of an observer. He silently ponders what he has seen with a mild sense of bewilderment...

(Latour & Woolgar, 1979/1986, pp.16–17)

Once again, many of the things that are described in this narrative are still true today: there is a frenzy of movement—by technicians, students, researchers, managers, assistants, writers, and janitors, as well as by money, machines, materials, networks, and organisms of all kinds—all required by the daily activity of science. Many experiments begun during the day continue overnight: sometimes they can be left to themselves, in buffer and on a shaker, chilling or heating, in a tissue culture hood or PCR machine; but in other cases they have to be supervised through the night, readings taken every two hours or animals tested at various stages of progression of a chemical or mechanical treatment. In one sense, the scientific activity of the laboratory never ceases. Even if no one is in the lab, there are animal techs in the mouse house, taking care of the subjects of the experiments. And many of the postdocs take articles home to read, or edit their own manuscripts when they're not in the lab. The fact that everyone has their own personal, portable computer has produced some changes in the workplace: there are no longer secretaries who type up manuscripts, since everyone does that for themselves on their laptop, and it's harder to keep work at the lab, since everyone brings it home with them on their personal computer.

The daily activity of science in the academic laboratory is still funded by what is,

in some sense of the word, charity. These scientists do not produce products that are sold on the shelves of any nearby stores. Any profit derived from their work is produced far away in time and space. So, academic scientists require generous investment from those who believe in the value of science, in some sense of value beyond short-time monetary gain. Usually, the majority of funding comes from taxpayer dollars. The National Institute of Health (NIH) still controls most of the funding available for academic labs in the biomedical sciences in the United States.⁷ Scientists apply for grants, which are evaluated by panels of experts in related fields, and funds are distributed to approved projects.⁸ GEL-X, however, gets only a portion of its funding from grants. It has the

⁷ In 2007, the NIH was responsible for approximately 27% of spending in support of *all* biomedical research in the United States, by far the largest contribution by any single agency. The pooled contribution of “industry” (including pharmaceutical, biotech, and medical device firms) contributed another 58%, with the remaining 15% coming from other federal, state, and local governmental agencies as well as foundations, charities, and other non-profit funds. In other words, the NIH controls more than a quarter of all the funding for biomedical research in the United States, whether for profit or not.

Narrowing the focus to *academic* biomedical research—considering only the financial support of biomedical research at colleges and universities in the United States—reveals that the federal government’s share of the expenditures jumps to 65%, with the next nearest contributor being the institutions themselves, at 18%. For every \$2 spent in a biomedical research lab at an American college or university, at least one of those dollars is from the NIH.

With regards to the agency’s total funds: in April of 2010, the current director of the NIH, Francis Collins, appeared before a Congressional subcommittee to request a budget of 32.1 billion dollars for the NIH in the 2011 fiscal year. And in terms of taxpayer dollars: that piece of the federal pie amounts to about \$100 per capita per year. In contrast, national defense spending adds up to about \$1600 per capita per year—16 times the annual per capita rate spent in biomed.

See Loscalzo (2006) for a detailed account of many of the facts mentioned in this note.

⁸ Project approval depends, of course, on many factors. But one aspect of the application process worth noting is something called ‘areas of research emphasis’. Chief among the many functions of the director of the NIH is the stipulation of these categories. These are five areas in which, for whatever reason, the director thinks that biomedical research could be significantly benefitted, in whatever way, by special attention—and this means more funding opportunities in that area.

Past areas of research emphasis have included: Minority Health, Women’s Health, AIDS Research, Disease Prevention, and Behavioral and Social Sciences Research (those are the five areas of research emphasis Harold E. Varmus, the director of NIH during the Clinton administration); or Eliminating Health Disparities, Exploiting Genomic Discoveries, Reinvigorating Clinical Research, Neuroscience Research, and Biomedical Computing (those are the five areas of research emphasis of Elias A. Zerhouni, the director of NIH during the latest Bush administration). At the start of his term as director, Collins named his five areas of research emphasis: High-Throughput Technologies, Translational Medicine, Benefiting Health Care Reform, Focusing More on Global Health, and Reinvigorating and Empowering the Biomedical Research Community. See Collins (2010) for further discussion of the current themes.

good fortune of being a Howard Hughes Medical Institute (HHMI) lab, which means that GEL-X gets a large percentage of its budget secured for every year without having to rely on grant applications. The HHMI is one of the largest philanthropic organizations in America: a nonprofit medical research organization that allocates hundreds of millions of dollars a year (approximately \$730 million in 2009) to those scientists named as HHMI Investigators (approximately 350 currently). Principal Investigator X was named an HHMI Investigator in 1985, and thus the GEL-X lab receives a majority of its funds from HHMI every year, to pay for salaries, supplies, equipment, and overhead.

The Observer's Story also includes mention of a Philippino cleaner. This is worth discussing briefly, since it is one of the only references to ethnicity in the entirety of Latour and Woolgar's text. Interestingly, it turns out to be not at all anomalous for Latour to have observed a Philippino working as a cleaner in the late seventies. This is because a majority of the menial laborers working at the Salk Institute around that time—as janitors and technicians, in landscaping and animal care—were Philippino. That particular ethnic group had a strong presence at the institution in that era, at least in the entry-level positions that did not require degrees from American colleges and universities. Principal Investigator X's first technician was a person of Philippino ethnicity, hired shortly after the principal investigator first came to the Salk in 1979. This technician had initially interviewed at the institute on the recommendation of another member of the Philippino community who was already employed by the Salk. Thirty years later, the former technician still works for GEL-X, but now this person is the lab

manager, and makes just over \$90,000 per year.⁹ This is more money than just about anyone else in the lab, other than Principal Investigator X. A good portion of the techs in GEL-X are also still Philippino: the manager is responsible for hiring the techs, and the kind of social connection that prompted her initial interview still exists, so there is often an increased probability that an opening in the lab will be filled by someone from her tightly-knit community. But throughout the rest of the Salk a sort of ethnic uniformity among the menial laborers is no longer present. It is no longer possible to predict with any likelihood what ethnicity a janitor might be.

Sub-Section 3.2: Literary Inscription and Origin Stories

There is a crucial point yet to be made about the Observer's Story: it includes the first hint at what will become a predominant theme in *Laboratory Life*, the fluctuating currency of statements. After analyzing the observations made by Latour in the laboratory, the authors end up conceptualizing the product of the observed scientific activity as literary inscription. In other words, Latour and Woolgar describe what scientists chiefly do as attempting to produce and support statements. The Observer's Story recounts of daily practice that "one or two statements have seen their credibility increase (or decrease) a few points, rather like the daily Dow Jones Industrial Average." But this report, of scientific activity as concerned mainly with literary inscription, represents a divergence from how the scientists themselves commonly describe their activity, as puzzle solving. This point is also made in the Observer's Story, when Latour and Woolgar write that "the total work effort is said to be guided by an invisible field, or

⁹ Data obtained in 2009.

more particularly, by a puzzle, the nature of which has already been decided upon and which may be solved today.”

There are two interesting things to note about this disagreement. First, scientists still commonly describe scientific activity as predominantly occupied with puzzle solving. When asked why they chose to go into science, scientists in GEL-X often explain their decision by referring to puzzle-solving, making claims such as “I really like to solve problems” and “I love the feeling of figuring something out, like a puzzle.” And second, here Latour and Woolgar are demonstrating one of the results of their adoption of the anthropological perspective. They are not accepting the self-report of the scientists as a description of what the scientists themselves are doing. The scientists say that they are solving puzzles, guided by an invisible field, but Latour and Woolgar claim instead that they are engaged in literary inscription, making statements and getting others to believe that these statements represent facts.

The particular question of how to best describe scientific activity (as a problem or puzzle solving activity as opposed to literary inscription) will be returned to, after mention of another way in which taking the anthropological perspective generates some disagreement for the authors with the scientists. When the scientists were asked what they worked on by non-members of the lab, Latour observed that the scientists often claimed to work “in neuroendocrinology.” To the scientists, this field represented a space which circumscribed the problems they tried to solve, or the things which any statements they produced were about. Latour and Woolgar discuss the scientists’ description of their field of neuroendocrinology, but with some degree of skepticism:

Our observer noticed that when asked by a total stranger, members of the

laboratory replied that they worked (or were) “in neuroendocrinology.” They went on to explain that neuroendocrinology was the result of a hybridization which had taken place in the 1940s between neurology, described as the science of the nervous system, and endocrinology, the science of the hormonal system. It occurred to our observer that such location “in a field” facilitated the correspondence between a particular group, network, or laboratory and a complex mixture of beliefs, habits, systematized knowledge, exemplary achievements, experimental practices, oral traditions, and craft skills. Although referred to as the “culture” in anthropology, this latter set of attributes is commonly subsumed under the term paradigm when applied to people calling themselves scientists. Neuroendocrinology seemed to have all the attributes of a mythology: it had its precursors, its mythical founders, and its revolutions. (1979/1986, p.54)

The fact that the authors of *Laboratory Life* are describing as a mythological origin story what to the scientists represents a historical account of the creation of their field indicates another point of significant disagreement between them.

So, scientists from Guillemin’s lab routinely provide the hybridization account as an explanation of the genesis and object of their research, and yet Latour and Woolgar discredit the authenticity of this activity with their manner of describing it. They describe the scientists of the lab as “a strange tribe who spend the greatest part of their day coding, marking, altering, correcting, reading and writing” (Latour & Woolgar, 1979/1986, p.49). Every tribe has a particular culture, where culture is understood “to refer to the set of arguments and beliefs to which there is a constant appeal in daily life and which is the object of all passions, fears, and respect” (Latour & Woolgar, 1979/1986, p.55). Latour and Woolgar imply that the scientist’s particular story of the hybridization of neurology and endocrinology into their field of neuroendocrinology is mostly fabricated when they write that “the mythology through which a culture represents itself is not necessarily entirely false” (1979/1966, p.55).

But beyond their general reasons for taking an anthropological perspective, particular reason for the authors' disbelief in this instance is not made explicit. In fact, Latour and Woolgar admit that an examination of the literature reveals a surge of papers related to the field at the time indicated by the origin story. The most specific criticism they make when describing the historical/mythological account is that:

As in many mythological versions of the scientific past, the struggle is now formulated in terms of a fight between abstract entities such as models and ideas. Consequently, present research appears based on one particular conceptual event, the explanation of which only merits scant elaboration by scientists. (Latour & Woolgar, 1979/1986, p.54)

This objection, to the stripping of concrete entities and a rich context from a scientific account, is a faint precursor of what will turn out to be the central point of contention between the scientists' own reports of their activity and Latour and Woolgar's distinct characterization of scientific activity. In this case the account of interest is an origin story. But a review of the text of *Laboratory Life* demonstrates that the authors see this behavior—a sort of stripping away of concrete context facilitating a reorientation around abstract entities—throughout the practice of science. And Latour and Woolgar will object to this practice throughout their text. This dispute explains why they characterize scientific activity via literary inscription, as well as why they describe fact elucidation as a process of construction rather than discovery.

Extremely interesting for this comparative discussion, gene expression is a field that—like neuroendocrinology—has an origin story of its own genesis via the hybridization of two fields. When asked about the origin of their area of research, members of GEL-X often describe an eventual synthesis between two split traditions, one in clinical medicine (specifically dealing with hormone physiology) and the other in

molecular biology (referring to early attempts to investigate genetics through virology). Those being indoctrinated into the traditions and culture of the lab may accept this account as it is given. But what *Laboratory Life* helps to show is the way in which these tribal rituals can obscure the many complexities of field creation. By eliminating the details of the persons, places, and materials involved in the synthesis, it can seem as if it is simply a void in the space of ideas that brings a field into existence. But there are many such spaces, and not all of them are filled. Those that are filled come into existence not just because of the possibility, but also because of the actuality of the circumstances of the people, places, and materials that coalesce.

Recognizing this complexity greatly enhances the account that one can provide for the origin of the area of research known as gene expression. For example, some of the founders of the field were scientists with experience both in medicine and in virology. The principal investigator of GEL-X was a pre-med student in undergrad whose father was a doctor. Principal Investigator X went to graduate school in microbiology, and worked with bacteria and viruses during graduate study. The presence of scientists with this sort of familiarity with both fields, at a certain time, interacting at certain conferences, and investigating certain questions of peculiar interest, helps to better explain how such a synthesis between the two fields could have occurred, beyond the abstract account of simply putting two groups of ideas and methods together. It is not just the ideas that have to fit together; there also have to be people familiar with both, capable of applying them to each other, and problems that might be solved with the application of just such a novel approach, and these problems have to have the kind of traction that secures funding and interest. By identifying each of these factors one might

be able to begin to construct a rich, complex, more complete account of the origin of any field, not just gene expression.

Sub-Section 3.3: Statement Types and Cycles of Credit

Simply based on a selection of ideas from the first couple of chapters, it should be apparent that *Laboratory Life* has already made a contribution to the development of a better understanding of the modern laboratory. But in chapter 3 the authors move on to their controversial account of fact construction. The following quote helps to explain this transition:

Whereas other tribes believe in gods or complicated mythologies, the members of this tribe insist that their activity is in no way to be associated with beliefs, a culture, or a mythology. Instead, they claim to be concerned only with “hard facts.” (Latour & Woolgar, 1979/1986, p.70)

This brings the discussion back to the aforementioned disagreement over Latour and Woolgar’s characterization of scientific practice as predominantly concerned with literary inscription. Although “the production of papers is acknowledged by participants as the main objective of their activity” (Latour & Woolgar, 1979/1986, p.71), the scientists still object to the characterization.

The authors of *Laboratory Life* explain:

Indeed, our observer incurred the considerable anger of members of the laboratory, who resented their representation as participants in some literary activity. In the first place, this failed to distinguish them from any other writers. Secondly, they felt that the important point was that they were writing *about* something... (Latour & Woolgar, 1979/1986, p.53)

In other words, disagreement arises mainly because Latour and Woolgar’s characterization of scientific practice as literary inscription fails to mention that the

process of inscription is about something in particular: facts. The kind of facts of interest to a particular laboratory are those demarcated by identification with a particular field, such as neuroendocrinology or gene expression. In both cases—of the description of scientific activity as literary inscription and the characterization of accounts of field formation as origin stories—dispute between the scientists and the authors of *Laboratory Life* arises over the role of facts in scientific practice.

Thus the focus of *Laboratory Life* naturally shifts to an examination of the statements of fact found in the manuscripts and articles produced by the fundamental scientific activity of literary inscription. Latour and Woolgar classify the statements they find according to a type system based on the degree of facticity with which these statements are imbued. Type 5 statements refer to facts so taken-for-granted that they are rarely even mentioned. Type 4 statements refer to similarly uncontroversial facts, but they are at least stated explicitly. Type 3 statements are qualified by indicating their claim to fact-hood rather than simply being stated as fact. Type 2 statements are openly acknowledged claims as opposed to established facts. Type 1 statements are mere hypotheses. Using this categorization schema, Latour and Woolgar are able to conceptualize laboratories as places whereby scientists engage in the process of literary inscription in an effort to transform their statements into facts. As they put it:

A laboratory is constantly performing operations on statements; adding modalities, citing, enhancing, diminishing, borrowing, and proposing new operations. Each of these operations can result in a statement which is either different or merely qualified. Each statement, in turn, provides the focus for similar operations in other laboratories. (Latour & Woolgar, 1979/1986, p.86–87)

This, then, is how scientific facts come into existence. Laboratories produce literary

inscriptions containing statements, and these statements fluctuate in type according to the degree of facticity which they acquire. As statements progress up the scale—into types 3, then 4, and eventually 5—facts are born.

Latour and Woolgar are particularly interested in one part of this story: that fact that “the function of literary inscription is the successful persuasion of readers, but the readers are only fully convinced when the sources of persuasion seem to have disappeared” (Latour & Woolgar, 1979/1986, p.76). According to the authors, diminished reference to sources of support for a statement seems to indicate the increased acceptance of the statement as fact. Once again, a process of concrete and contextual stripping seems to be occurring as a fundamental aspect of scientific practice. In an effort to describe the process by which the sources of persuasion are removed from a statement (or, how a statement comes by be made and then eventually accepted as fact), Latour and Woolgar present a detailed case study from Guillemin’s lab. The authors call their case study “the construction of a fact: the case of TRF(H)” (Latour & Woolgar, 1979/1986, p.105), and it tells the story of a major accomplishment for the laboratory: the elucidation and acceptance of the structure of a particular molecule, thyrotropin releasing factor/hormone.

Of course, GEL-X has its own literary inscriptions. The laboratory has produced many manuscripts filled with statements in various stages of facticity, and the scientists of the laboratory perform their own operations, moving the statements along the scale of types, hoping to establish certain claims as fact. Probably the central accomplishment of Principal Investigator X’s laboratory has been the elucidation of the nuclear receptor superfamily. Here is a brief scientific description of this entity, from a review article:

The nuclear receptor superfamily is comprised of over 150 different proteins that have evolved to mediate a complex array of extracellular signals into transcriptional responses. Many, but not all, of these proteins directly bind to signaling molecules, which, because of their small lipophilic character, can easily enter the cell. Thus, unlike membrane-bound receptors, the nuclear receptors are intracellular and function to control the activity of target genes directly. In aggregate, these target genes comprise a genetic network whose coordinate activity defines the physiological response. (Mangelsdorf & Evans, 1995, p.841)

Latour and Woolgar's characterization of scientific practice suggests various ways that might enrich this account. The simply stated paragraph quoted above contains many statements, each of which has its own distinct history. For example, the identification of the various members of the family was a venture involving several labs, including GEL-X, and which took place over many years. And each of the statements in the above paragraph may belong to any of the five statement types.

Or they might belong to several, when there is active disagreement. For example, the final sentence quoted is actually quite a controversial one. Fifteen years after the publication of this review, Principal Investigator X and the rest of GEL-X are still attempting to persuade other members of the gene expression community that the network of genes targeted by the nuclear receptor superfamily is what defines the physiological response. Outside the gene expression community, very different notions of what comprises the physiological response are advocated and applied. So, members of GEL-X would probably classify the statement that "these target genes comprise a genetic network whose coordinate activity defines the physiological response" as type 4 on Latour and Woolgar's scale. Members of rival labs also working on nuclear receptors with an alternate classification scheme might classify it as type 2. Others within the gene expression community, aware of a certain balance between consensus and agreement,

would probably label the statement as type 3. And outside of gene expression, scientists with alternate conceptions of what defines the physiological response (perhaps an organ-based rather than a molecular notion) might consider the statement type 1, or simply think of it as false.

What is important for this discussion is to note the way that Latour and Woolgar's analysis of scientific practice helps to enrich understanding of what is presented in the scientific product as baldly stated fact. There is a scientific process by which laboratories produce literary inscriptions—transforming materials, organisms, equipment, and technical expertise into data which is then translated into statements contained in manuscripts—but there is also a correspondingly scientific process by which these statements are operated on—sliding them up and down the scale of types 1 through 5. Both processes act in a way that strips the products of the context of their production, but the latter also serves to cloak uncertainty and dissent within science. This discussion occupies a somewhat privileged epistemic position, a product of extensive observation of and experience in GEL-X. This familiarity grants access to the increased complexity surrounding the claims stated so simply in the articles produced by the laboratory. But this is not information available to everyone. To those without access, the statements must all seem to have the ring of certainty and the sheen of facticity.

The fact that the scientific process encourages a sort of layering of information, concealing dissent and uncertainty behind the presentation of statements as pure fact, is a matter of some interest. At the very least, such practice must encourage misinterpretation of some statements, as more or less certain or established than they might actually be. What might explain this intriguing procedure? Latour and Woolgar examine some

possible answers to this query during their investigation of “the microprocesses whereby facts are socially constructed” (1979/1986, p.152). In this discussion they offer some insight into why a more complete account of an actual, concrete process of fact construction might be stripped of its context and presented via a simpler account of a more abstract process:

By transforming the second account into the first, the teller transforms a localized, heterogeneous, and material set of circumstances (in which social factors are clearly visible) into the sudden occurrence of a personal and abstract idea which bears no trace of its social construction. (Latour & Woolgar, 1979/1986, p.170)

So the stripping process removes unwanted traces of social construction. There has been extensive discussion of the widespread tendency to emphasize objectivity in science, and to eliminate social factors. In *Laboratory Life*, the authors mention this during their discussion of the commonly assumed divide between social and scientific (or technical or intellectual) factors influencing scientific practice.¹⁰ Latour and Woolgar do not accept the validity of this distinction, but they do acknowledge the common perception that the distinction holds, particularly among scientists themselves.

Given that the externality of social factors is a widely adopted norm for scientific activity, it is not difficult to understand why scientists might engage in processes that would eliminate mention of social factors and a richer context from their accounts, despite the somewhat misleading nature of this activity. The stripping process results in a sort of purification, increasing the credibility of the scientific product, by making it more in line with the standard norms of scientific practice. Interestingly, Latour and Woolgar deny that the majority of scientists are strongly motivated by these norms. For example:

¹⁰ Another very common way of discussing this distinction is as between external and internal factors.

“the explanatory power of norms falls well short of our objective of understanding both science and the scientists who make it” and “appeals to norms were extremely rare among our respondents” (Latour & Woolgar, 1979/1986, p.190)—although it is not clear why this lack of appeal should count as evidence for Latour and Woolgar, since they claim not to be taking the scientists' own proclamations "at face value." Regardless, even without most scientists being motivated by the norms themselves, they can still be motivated by the enhanced credibility that the adoption of predominant norms will produce.

As Latour and Woolgar demonstrate throughout their discussion of cycles of credit, credibility has great import for scientific activity. They explain:

Scientists' behavior is remarkably similar to that of an investor of capital. An accumulation of credibility is prerequisite to investment. The greater this stockpile, the more able the investor to reap substantial returns and thus add further to his growing capital. (Latour & Woolgar, 1979/1986, p.197)

Scientists' accumulation, preservation, and employment of credibility, which acts as capital in the scientific 'market', are essential to their successful practice of science.

Laboratory Life documents the extensive degree to which scientists in Guillemin's lab are concerned with credibility and credit. This is also a feature of members of GEL-X.

Latour and Woolgar write that:

Scientific activity in our laboratory comprised a field of contention in which facts were produced, claims dissolved, artefacts deconstructed, proofs and arguments disproved, careers ruined, and prestige cut down. (1979/1986, p.212)

And this is also a strikingly apt description of current scientific activity in the lab of Principal Investigator X. A single publication in *Nature*, *Science*, or *Cell* is often

sufficient for a graduate student or postdoc from GEL-X to get hired as the principal investigator of their own lab. As this extensive discussion of scientific practice has demonstrated, publications such as these are produced via the generation of literary inscriptions filled with statements possessing varying degrees of facticity and interest. The extent to which other members of the scientific community (including those acting as reviewers for journals) can be convinced as to the facticity and interest of those statements will strongly influence the successful development of each scientist's career. And credibility is instrumental in any attempt to convince or persuade anyone of anything. It is therefore not at all surprising that credibility would be crucial within a field where individuals need to be convincing in order to succeed. Such individuals are likely to be highly motivated by the importance that development of their credibility will have on the success of their career.

Additionally, the importance of credibility persists throughout a scientist's career. Efforts to obtain credit and prestige continue even after acquiring a laboratory of one's own. One reason for this is that funding still has to be secured every year. Having lab space is no guarantee of funds with which to populate it. Even scientists with some secured funding from their home institution or via an award such as that of HHMI Investigator depend on their credibility when applying for additional grants, among other things.

During the first hour of being interviewed for this study, while giving a sort of evaluative history of the lab and its primary field of research, Principal Investigator X mentioned the Nobel Prize fourteen times (almost once every four minutes). Telling the narrative in this way—a sort of “history by award”—is understandable in light of the

immense recognition and corresponding benefit that winning the Nobel and acquiring such prestige virtually guarantees for one's lab. For instance, it affects personal and professional acquisition of resources (it is incredibly rare for a Nobel winner's grant applications to NIH or NSF to get rejected), as well as leads to increased credibility and hence more easily accepted publications and characterizations of statements as facts. Recognition of the fact that science is a place where credibility is the most valuable currency helps to make sense of many of the initially puzzling practices identified throughout this discussion. It even explains why certain seemingly detrimental practices common to scientific activity (such as the stripping of concrete context in favor of more abstract characterizations) might be not just generated but also encouraged by the norms and structure of science.

Sub-Section 3.4: The (Social) Construction of Scientific Facts

The presentation of scientific facts as (socially) constructed is perhaps the most contentious aspect of *Laboratory Life*.¹¹ The idea of construction persists throughout Latour and Woolgar's account of scientific practice, constitutes a point of significant contention between the authors as science studies scholars and the lab members as practicing scientists, and is still of major interest in science studies.

The idea is also one of the authors' most ambiguous. On the one hand, Latour and Woolgar clearly view scientific practice as constructing the objects of study, making

¹¹ The term 'social' is included in parenthesis prior to the term 'construction' because of Latour and Woolgar's inconstancy regarding its inclusion. The initial subtitle of *Laboratory Life* was "The Social Construction of Scientific Facts" (1979) but in the second edition this was altered to "The Construction of Scientific Facts" (1986). According to the postscript of the later version, the term was dropped because: "by demonstrating its pervasive applicability, the social study of science has rendered 'social' devoid of any meaning" (1979/1986, p.281).

contextually dependent claims, and producing subjective, inventive accounts.

Throughout the text, the authors make claims like:

It is not simply that phenomena *depend on* certain material instrumentation; rather, the phenomena *are thoroughly constituted by* the material setting of the laboratory. (Latour & Woolgar, 1979/1986, p.64)

And:

The material setting both makes possible the phenomena and is required to be easily forgotten. Without the material environment of the laboratory none of the objects could be said to exist, and yet the material environment very rarely receives mention. (Latour & Woolgar, 1979/1986, p.69)

Finally:

By contrast, we do not conceive of scientists using various strategies as pulling back the curtain on pre-given, but hitherto concealed, truths. Rather, objects (in this case, substances) are constituted through the artful creativity of scientists. (Latour & Woolgar, 1979/1986, p.129)

On the other hand, when discussing the construction of a fundamental fact in Guillemin's lab, Latour and Woolgar state that:

In one important sense, then, TRF did not exist prior to the imposition of limitations, because such limitations preceded the first experiments and defined what could be accepted in advance. (Latour & Woolgar, 1979/1986, p.121)

Here the authors claim that in one sense, TRF did not exist prior to the investigation. But this implies that in another sense, it did. In fact, even the materiality mentioned in the previous set of quotes is a kind of precursor to the construction as well as part of the process of constructing, a dual object of creator and created.

In other words: as is repeatedly indicated by Latour and Woolgar's discussion, the failure of scientists to acknowledge social and historical context throughout their practice presents an extremely incomplete portrait of scientific activity. So it seems necessary to incorporate at least some elements of creativity and construction into the characterization

of scientific activity. But despite these inventive elements, it is also apparent that reality pushes back. It is not the case that anything goes: there are restrictions on what can be created and constructed, limits that are defined by the phenomena. How these phenomena are detected, categorized, manipulated, and produced are all aspects subject to creation and construction, but there is still the material reality of the stuff that the creations are created from, and that the constructions are constructed of.

The text of *Laboratory Life* is never entirely clear about how to reconcile these two, rather contrary aspects—which might be called constructivist and realist tendencies. Perhaps unsurprisingly, given this latent ambiguity, this is an area where there has been substantial development and evolution of the authors' positions as well as those of others in science studies. Many science studies scholars have made contributions to the constructivist debate—perhaps an especially significant one is Hacking's *The Social Construction of What?* (1999). But with regard to current scientific practice, it is worth noting that the scientists of GEL-X tend to protest just as vociferously today as those in Gullemin's lab in the seventies protested the constructivist account of their practices.

Sub-Section 3.5: The Creation of Order out of Disorder

To conclude *Laboratory Life*, Latour and Woolgar summarize their characterization of scientific practice with the help of one final concept: the creation of order out of disorder. As they explain it:

In sum, then, our discussion is informed by the conviction that a body of practices widely regarded by outsiders as well organized, logical, and coherent, in fact consists of a disordered array of observations with which scientists struggle to produce order. (Latour & Woolgar, 1979/1986, p.36)

In keeping with the reflexive tradition that the authors of the text themselves adopt, this retrospective analysis of *Laboratory Life* attempts to both understand Latour and Woolgar's characterization of science as the creation of order out of disorder, and to apply the concept of the creation of order out of disorder to this newer characterization of sciences and its more recent set of observational data.

In some cases Latour and Woolgar's ideas have straightforwardly applied, and this has contributed to the creation of order out of disorder in the understanding and characterization of current scientific practice. But other aspects of the discussion, applying certain concepts from *Laboratory Life* to illuminate current laboratory experience in GEL-X, are not yet completely resolved. In some cases the ideas introduced by the text serve to create further queries for both the laboratory experience and the conceptual characterization of the daily activity of science. Instances of primary concern include the characterization of field hybridization as origin stories, the stripping of context that conceals fact construction, and the ambiguous nature of construction itself. These issues, evoked by the study of one of today's laboratories, and left unresolved by the characterization of the daily activity of science offered by Latour and Woolgar, are excellent candidates for further investigation, which might contribute to an updated and fuller conception of scientific activity.

Section 4: Conclusion

Discussion from sections two and three of this appendix indicates that some important aspects of laboratory life at the Salk have remained relatively stable since the time of Latour's observation period. Several examples can be easily provided.

One: many of the buildings are the same, although new ones have been added to the grounds as the number of personnel and labs at the Institute has increased. Two: most of the labs still get the majority of their funding from the NIH. A substantial chunk of the funds awarded by successful grants is apportioned to the Salk as overhead. That is how the Institute itself is mainly funded, in addition to donations from public charities such as the March of Dimes and private benefactors. Three: despite the fact that there are now many more people working at the Salk than there were in the seventies, the lab personnel are still divided into lab directors/principal investigators, postdocs, students, and technicians. Ethnic and gender diversity has increased within most of those categories, but not all. Out of the fifty-six professors who currently run their own lab at the Institute, only ten are female. Four: most of the scientific research conducted at the Salk is biological, although the divisions within that field have fluctuated quite a bit during the last thirty years. New departments have appeared and old ones have dissolved. Five: although the technology used to conduct the research has advanced remarkably, in some ways things have really stayed the same. There are new machines, compounds, protocols, and even organisms introduced all of the time. But what has stayed constant is the fact that experimenters must constantly update their equipment and skills in order to stay current.

So, some things have stayed the same. But other things have changed, and even some of the similarities indicate the presence of other interesting differences. To deal with this divergence, as well as some of the unresolved issues of *Laboratory Life*, some of Latour and Woolgar's concepts must be extended or amended, and other entirely foreign concepts must be introduced. Perhaps in another thirty years, the ethnographic

tradition will continue, and new light will be cast on these and other issues alike.

REFERENCES

- Bearman, D. 1979. Review of *Laboratory life: The social construction of scientific facts*. *Science* 206: 824–825.
- Collins, F. 2010. Opportunities for research and NIH. *Science* 327: 36–37.
- Collins, H. M. 1988. Review of *Laboratory life: The construction of scientific facts*. *Isis* 79(1): 148–149.
- Hacking, I. 1999. *The social construction of what?* Boston, Massachusetts: Harvard University Press.
- Haraway, D. 1980. Review of *Laboratory life: The social construction of scientific facts*. *Isis* 71(3): 489–490.
- Irwin, A. 1982. Review of *Laboratory life: The social construction of scientific facts*. *The British Journal for the History of Science* 15(2): 208–209.
- Knorr-Cetina, K. D. 1981. *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Oxford: Pergamon Press.
- Krohn, R. 1981. Review of *Laboratory life: The social construction of scientific facts*. *Contemporary Sociology* 10(3): 433–434.
- Kuhn, T. 1957. *The copernican revolution: Planetary astronomy in the development of western thought*. Cambridge, MA: Harvard University Press.
- Kuhn, T. 1996. *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press. (Original work published 1962)
- Latour, B. & Woolgar, S. 1986. *Laboratory life: The construction of scientific facts* (2nd ed.). Princeton: Princeton University Press. (Original work published 1979)
- Loscalzo, J. 2006. The NIH budget and the future of biomedical research. *New England Journal of Medicine* 354: 1665–1667.
- Lynch, M. 1985. *Art and artifact: A study of shop work and shop talk in a research laboratory*. London: Routledge & Kegan Paul.
- Mangelsdorf, D. & Evans, R. M. 1995. The RXR heterodimers and orphan receptors. *Cell* 83: 841–850.
- Mullins, N. C. 1980. Review of *Laboratory life: The social construction of scientific facts*. *science, technology, and human values* 5(30): 55.

- Salk, J. 1986. Introduction. In B. Latour & S. Woolgar (eds.), *Laboratory life: The construction of scientific facts* (2nd ed., pp.11–14). Princeton: Princeton University Press. (Original work published 1979)
- van Themaat, V. 1982. Review of *Laboratory life: The social construction of scientific facts*. *Journal for General Philosophy of Science* XIII(12): 166–170.
- Westman, R. S. 1994. Two cultures or one? A second look at Kuhn's *The Copernican Revolution*. *Isis* 85: 79–115.