ARTICLE

# Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences

Naomi Oreskes,* Kristin Shrader-Frechette, Kenneth Belitz

Verification and validation of numerical models of natural systems is impossible. This is because natural systems are never closed and because model results are always nonunique. Models can be confirmed by the demonstration of agreement between observation and prediction, but confirmation is inherently partial. Complete confirmation is logically precluded by the fallacy of affirming the consequent and by incomplete access to natural phenomena. Models can only be evaluated in relative terms, and their predictive value is always open to question. The primary value of models is heuristic.

In recent years, there has been a dramatic increase in the use of numerical simulation models in the earth sciences as a means to evaluate large-scale or complex physical processes. In some cases, the predictions generated by these models are considered as a basis for public policy decisions: Global circulation models are being used to predict the behavior of the Earth's climate in response to increased $CO_2$ concentrations; resource estimation models are being used to predict petroleum reserves in ecologically sensitive areas; and hydrological and geochemical models are being used to predict the behavior of toxic and radioactive contaminants in proposed waste disposal sites. Government regulators and agencies may be required by law to establish the trustworthiness of models used to determine policy or to attest to public safety (*1*, *2*); scientists may wish to test the veracity of models used in their investigations. As a result, the notion has emerged that numerical models can be "verified" or "validated," and techniques have been developed for this purpose (*1*, *3*–*5*). Claims about verification and validation of model results are now routinely found in published literature (*6*).

Are claims of validity and verity of numerical models legitimate (*2*, *7*)? In this article, we examine the philosophical basis of the terms "verification" and "validation" as applied to numerical simulation models in the earth sciences, using examples from hydrology and geochemistry. Because demand for the assessment of accuracy in numerical modeling is most evident at the interface between public policy and scientific usage, we focus on examples relevant to policy (*8*). The principles illustrated, however, are generic.

N. Oreskes is in the Department of Earth Sciences and the Department of History, Dartmouth College, Hanover, NH 03755. K. Shrader-Frechette is in the Department of Philosophy, University of South Florida, Tampa, FL 33620. K. Belitz is in the Department of Earth Sciences, Dartmouth College, Hanover, NH 03755.

*To whom correspondence should be addressed.

## Verification: The Problem of "Truth"

The word verify (from Latin, *verus*, meaning true) means an assertion or establishment of truth (*9*). To say that a model is verified is to say that its truth has been demonstrated, which implies its reliability as a basis for decision-making. However, it is impossible to demonstrate the truth of any proposition, except in a closed system. This conclusion derives directly from the laws of symbolic logic. Given a proposition of the form "p" entails "q," we know that if "p" is true, then "q" is true if and only if the system that this formalism represents is closed.

For example, I say, "If it rains tomorrow, I will stay home and revise this paper." The next day it rains, but you find that I am not home. Your verification has failed. You conclude that my original statement was false. But in fact, it was my intention to stay home and work on my paper. The formulation was a true statement of my intent. Later, you find that I left the house because my mother died, and you realize that my original formulation was not false, but incomplete. It did not allow for the possibility of extenuating circumstances (*10*). Your attempt at verification failed because the system was not closed.

This example is trivial, but even an apparently trivial proposition can be part of a complex open system. Indeed, it is difficult to come up with verbal examples of closed systems because only purely formal logical structures, such as proofs in symbolic logic and mathematics, can be shown to represent closed systems. Purely formal structures are verifiable because they can be proved by symbolic manipulations, and the meaning of these symbols is fixed and not contingent on empirically based input parameters (*11*).

Numerical models may contain closed mathematical components that may be verifiable, just as an algorithm within a computer program may be verifiable (*12*). Mathematical components are subject to verification because they are part of closed systems that include claims that are always true as a function of the meanings assigned to the specific symbols used to express them (*13*). However, the models that use these components are never closed systems. One reason they are never closed is that models require input parameters that are incompletely known. For example, hydrogeological models require distributed parameters such as hydraulic conductivity, porosity, storage coefficient, and dispersivity, which are always characterized by incomplete data sets. Geochemical models require thermodynamic and kinetic data that are incompletely or only approximately known. Incompleteness is also introduced when continuum theory is used to represent natural systems. Continuum mechanics necessarily entails a loss of information at the scale lower than the averaging scale. For example, the Darcian velocity of a porous medium is never identical to the velocity structure at the pore scale. Finer scale structure and process are lost from consideration, a loss that is inherent in the continuum mechanics approach.

Another problem arises from the scaling-up of nonadditive properties. The construction of a numerical simulation model of a ground-water flow system involves the specification of input parameters at some chosen scale. Typically, the scale of the model elements is on the order of meters, tens of meters, or kilometers. In contrast, the scale on which input parameters are measured is typically much smaller, and the relation between those measurements and larger scale model parameters is always uncertain and generally unknown. In some cases, it is possible to obtain input data at the scale chosen by the modeler for the model elements (for example, pump tests), but this is not often done, for practical reasons. Even when such measurements are available, they are never available for all model elements (*14*).

Another reason hydrological and geochemical models are never closed systems is that the observation and measurement of both independent and dependent variables are laden with inferences and assumptions. For example, a common assumption in many geochemical models of water-rock interaction is that observable mineral as-

semblages achieve equilibrium with a modeled fluid phase. Because relevant kinetic data are frequently unavailable, kinetic effects are assumed to be negligible (*15*). But many rocks contain evidence of disequilibrium on some scale, and the degree of disequilibrium and its relation to kinetic controls can rarely, if ever, be quantified. To attempt to do so would necessarily involve further inferences and assumptions. Similarly, the absence of complete thermodynamic data for mineral solid solutions commonly forces modelers to treat minerals as ideal end-members, even when this assumption is known to be erroneous on some level. Measurement of the chemical composition of a mineral phase to estimate the activities of chemical components within it requires instrumentation with built-in assumptions about such factors as interference effects and matrix corrections. What we call data are inference-laden signifiers of natural phenomena to which we have incomplete access (*16*). Many inferences and assumptions can be justified on the basis of experience (and sometimes uncertainties can be estimated), but the degree to which our assumptions hold in any new study can never be established a priori. The embedded assumptions thus render the system open.

The additional assumptions, inferences, and input parameters required to make a model work are known as "auxiliary hypotheses" (*17*). The problem of deductive verification is that if the verification fails, there is often no simple way to know whether the principal hypothesis or some auxiliary hypothesis is at fault. If we compare a result predicted by a model with observational data and the comparison is unfavorable, then we know that something is wrong, and we may or may not be able to determine what it is (*18*). Typically, we continue to work on the model until we achieve a fit (*19*). But if a match between the model result and observational data is obtained, then we have, ironically, a worse dilemma. More than one model construction can produce the same output. This situation is referred to by scientists as nonuniqueness and by philosophers as underdetermination (*20, 21*). Model results are always underdetermined by the available data. Two or more constructions that produce the same results may be said to be empirically equivalent (*22*). If two theories (or model realizations) are empirically equivalent, then there is no way to choose between them other than to invoke extraevidential considerations like symmetry, simplicity, and elegance, or personal, political, or metaphysical preferences (*19, 23–25*).

A subset of the problem of nonuniqueness is that two or more errors in auxiliary hypotheses may cancel each other out. Whether our assumptions are reasonable is not the issue at stake. The issue is that often there is no way to know that this cancellation has occurred. A faulty model may appear to be correct. Hence, verification is only possible in closed systems in which all the components of the system are established independently and are known to be correct. In its application to models of natural systems, the term verification is highly misleading. It suggests a demonstration of proof that is simply not accessible (*26*).

## Validation

In contrast to the term verification, the term validation does not necessarily denote an establishment of truth (although truth is not precluded). Rather, it denotes the establishment of legitimacy, typically given in terms of contracts, arguments, and methods (*27*). A valid contract is one that has not been nullified by action or inaction. A valid argument is one that does not contain obvious errors of logic. By analogy, a model that does not contain known or detectable flaws and is internally consistent can be said to be valid. Therefore, the term valid might be useful for assertions about a generic computer code but is clearly misleading if used to refer to actual model results in any particular realization (*28*). Model results may or may not be valid, depending on the quality and quantity of the input parameters and the accuracy of the auxiliary hypotheses.

Common practice is not consistent with this restricted sense of the term. Konikow and Bredehoeft (*2*) have shown that the term validation is commonly used in at least two different senses, both erroneous. In some cases, validation is used interchangeably with verification to indicate that model predictions are consistent with observational data. Thus, modelers misleadingly imply that validation and verification are synonymous, and that validation establishes the veracity of the model. In other cases, the term validation is used even more misleadingly to suggest that the model is an accurate representation of physical reality. The implication is that validated models tell us how the world really is. For example, the U.S. Department of Energy defines validation as the determination "that the code or model indeed reflects the behavior of the real world" (*29*). Similarly, the International Atomic Energy Agency has defined a validated model as one that provides "a good representation of the actual processes occurring in a real system" (*30*). For all the reasons discussed above, the establishment that a model accurately represents the "actual processes occurring in a real system" is not even a theoretical possibility.

How have scientists attempted to demonstrate that a model reflects the behavior of the real world? In the Performance Assessment Plan for the proposed high-level nuclear waste repository at Yucca Mountain, Nevada, Davis and co-workers (*1*) suggest that "[t]he most common method of validation involves a comparison of the measured response from in situ testing, lab testing, or natural analogs with the results of computational models that embody the model assumptions that are being tested" (*31*). But the agreement between any of these measures and numerical output in no way demonstrates that the model that produced the output is an accurate representation of the real system. Validation in this context signifies consistency within a system or between systems. Such consistency entails nothing about the reliability of the system in representing natural phenomena.

## "Verification" of Numerical Solutions

Some workers would take as a starting point for their definition of terminology the analytical solution to a boundary value or initial value problem. In this context, they may compare a numerical solution with an analytical one to demonstrate that the two match over a particular range of conditions under consideration. This practice is often referred to as verification (*4*, pp. 7–8; *32*).

The comparison of numerical with analytical solutions is a critical step in code development; the failure of a numerical code to reproduce an analytical solution may certainly be cause for concern. However, the congruence between a numerical and an analytical solution entails nothing about the correspondence of either one to material reality. Furthermore, even if a numerical solution can be said to be verified in the realm of the analytical solution, in the extension of the numerical solution beyond the range and realm of the analytical solution (for example, time, space, and parameter distribution), the numerical code would no longer be verified. Indeed, the raison d'être of numerical modeling is to go beyond the range of available analytical solutions. Therefore, in application, numerical models cannot be verified. The practice of comparing numerical and analytical solutions is best referred to as bench-marking. The advantage of this term—with its cultural association with geodetic practice—is that it denotes a reference to an accepted standard whose absolute value can never be known (*33*).

## Calibration of Numerical Models

In the earth sciences, the modeler is commonly faced with the inverse problem: The distribution of the dependent variable (for example, the hydraulic head) is the most well known aspect of the system; the distribution of the independent variable is the least well known. The process of tuning the

model—that is, the manipulation of the independent variables to obtain a match between the observed and simulated distribution or distributions of a dependent variable or variables—is known as calibration.

Some hydrologists have suggested a two-step calibration scheme in which the available dependent data set is divided into two parts. In the first step, the independent parameters of the model are adjusted to reproduce the first part of the data. Then in the second step the model is run and the results are compared with the second part of the data. In this scheme, the first step is labeled "calibration," and the second step is labeled "verification." If the comparison is favorable, then the model is said to be "verified" (*3*, p. 110; *4*, p. 253). The use of the term verification in this context is highly misleading, for all the reasons given above. A match between predicted and obtained output does not verify an open system. Furthermore, models almost invariably need additional tuning during the so-called verification phase (*3*, p. 110). That is, the comparison is typically unfavorable, and further adjustments to the independent parameters have to be made. This limitation indicates that the so-called verification is a failure. The second step is merely a part of the calibration.

Given the fundamental problems of verification, Bas van Fraassen (*22*) has argued that the goal of scientific theories is not truth (because that is unobtainable) but empirical adequacy. Using van Fraassen's terminology, one could say that a calibrated model is empirically adequate. However, the admission that calibrated models invariably need "additional refinements" (*3*, p. 110) suggests that the empirical adequacy of numerical models is forced. The availability of more data requires more adjustments. This necessity has serious consequences for the use of any calibrated model (or group of models) for predictive purposes, such as to justify the long-term safety of a proposed nuclear or toxic waste disposal site. Consider the difference between stating that a model is "verified" and stating that it has "forced empirical adequacy" (*34*).

Finally, even if a model result is consistent with present and past observational data, there is no guarantee that the model will perform at an equal level when used to predict the future. First, there may be small errors in input data that do not impact the fit of the model under the time frame for which historical data are available, but which, when extrapolated over much larger time frames, do generate significant deviations. Second, a match between model results and present observations is no guarantee that future conditions will be similar, because natural systems are dynamic and may change in unanticipated ways (*35*).

## Confirmation

If the predicted distribution of dependent data in a numerical model matches observational data, either in the field or laboratory, then the modeler may be tempted to claim that the model was verified. To do so would be to commit a logical fallacy, the fallacy of "affirming the consequent." Recall our proposition, "If it rains tomorrow, I will stay home and revise this paper." This time, you find that I am home and busily working on my paper. Therefore you conclude that it is raining. Clearly, this is an example of faulty logic. The weather might be glorious, but I decided that this paper was important enough to work on in spite of the beautiful weather. To claim that a proposition (or model) is verified because empirical data match a predicted outcome is to commit the fallacy of affirming the consequent. If a model fails to reproduce observed data, then we know that the model is faulty in some way, but the reverse is never the case (*36*).

This conclusion, which derives strictly from logic, may seem troubling given how difficult it can be to make a model or develop a hypothesis that reproduces observed data. To account for this discrepancy, philosophers have developed a theory of confirmation, founded on the notion of science as a hypothetico-deductive activity. In this view, science requires that empirical observations be framed as deductive consequences of a general theory or scientific law (*37*). If these observations can be shown to be true, then the theory or law is "confirmed" by those observations and remains in contention for truth (*17*). The greater the number and diversity of confirming observations, the more probable it is that the conceptualization embodied in the model is not flawed (*38*). But confirming observations do not demonstrate the veracity of a model or hypothesis, they only support its probability (*39, 40*).

Laboratory tests, in situ tests, and the analysis of natural analogs are all forms of model confirmation. But no matter how many confirming observations we have, any conclusion drawn from them is still an example of the fallacy of affirming the consequent. Therefore, no general empirical proposition about the natural world can ever be certain. No matter how much data we have, there will always be the possibility that more than one theory can explain the available observations (*41*). And there will always remain the prospect that future observations may call the theory into question (*42*). We are left with the conclusion that we can never verify a scientific hypothesis of any kind. The more complex the hypothesis, the more obvious this conclusion becomes. Numerical models are a form of highly complex scientific hypothesis. Confirmation theory requires us to support numerical simulation results with other kinds of scientific observations and to realize that verification is impossible.

## Numerical Models and Public Policy

Testing hypotheses is normal scientific practice, but model evaluation takes on an added dimension when public policy is at stake. Numerical models are increasingly being used in the public arena, in some cases to justify highly controversial decisions. Therefore, the implication of truth is a serious matter (*43*). The terms verification and validation are now being used by scientists in ways that are contradictory and misleading. In the earth sciences—hydrology, geochemistry, meteorology, and oceanography—numerical models always represent complex open systems in which the operative processes are incompletely understood and the required empirical input data are incompletely known. Such models can never be verified. No doubt the same may be said of many biological, economic, and artificial intelligence models.

What typically passes for validation and verification is at best confirmation, with all the limitations that this term suggests. Confirmation is only possible to the extent that we have access to natural phenomena, but complete access is never possible, not in the present and certainly not in the future. If it were, it would obviate the need for modeling. The central problem with the language of validation and verification is that it implies an either-or situation. In practice, few (if any) models are entirely confirmed by observational data, and few are entirely refuted. Typically, some data do agree with predictions and some do not. Confirmation is a matter of degree. It is always inherently partial. Furthermore, both verify and validate are affirmative terms: They encourage the modeler to claim a positive result (*44*). And in many cases, a positive result is presupposed. For example, the first step of validation has been defined by one group of scientists as developing "a strategy for demonstrating [regulatory] compliance" (*1, 45*). Such affirmative language is a roadblock to further scrutiny.

A neutral language is needed for the evaluation of model performance. A model can certainly perform well with respect to observational data, in which case one can speak of the precision and accuracy of the fit. Judgmental terms such as excellent, good, fair, and poor are useful because they invite, rather than discourage, contextual definition. Legitimately, all we can talk about is the relative performance of a model with respect to observational data, other models of the same site, and our own expectations based on theoretical precon-

ceptions and experience of modeling other sites. None of these things can be discussed in absolute terms.

## Then What Good Are Models?

Models can corroborate a hypothesis by offering evidence to strengthen what may be already partly established through other means. Models can elucidate discrepancies in other models. Models can be also be used for sensitivity analysis—for exploring "what if" questions—thereby illuminating which aspects of the system are most in need of further study, and where more empirical data are most needed. Thus, the primary value of models is heuristic: Models are representations, useful for guiding further study but not susceptible to proof.

The idea of model as representation has led the philosopher Nancy Cartwright to the claim that models are "a work of fiction" (*46*). In her words, "some properties ascribed to objects in the model will be genuine properties of the objects modeled, but others will be merely properties of convenience." Her account, which is no doubt deliberately provocative, will strike many scientists as absurd, perhaps even offensive. While not necessarily accepting her viewpoint, we might ponder this aspect of it: A model, like a novel, may resonate with nature, but it is not a "real" thing. Like a novel, a model may be convincing—it may "ring true" if it is consistent with our experience of the natural world. But just as we may wonder how much the characters in a novel are drawn from real life and how much is artifice, we might ask the same of a model: How much is based on observation and measurement of accessible phenomena, how much is based on informed judgment, and how much is convenience? Fundamentally, the reason for modeling is a lack of full access, either in time or space, to the phenomena of interest. In areas where public policy and public safety are at stake, the burden is on the modeler to demonstrate the degree of correspondence between the model and the material world it seeks to represent and to delineate the limits of that correspondence.

Finally, we must admit that a model may confirm our biases and support incorrect intuitions. Therefore, models are most useful when they are used to challenge existing formulations, rather than to validate or verify them. Any scientist who is asked to use a model to verify or validate a predetermined result should be suspicious.

### REFERENCES AND NOTES


1. P. A. Davis, N. E. Olague, M. T. Goodrich, "Approaches for the validation of models used for performance assessment of high-level nuclear waste repositories," SAND90-0575/NUREG CR-5537 (Sandia National Laboratories, Albuquerque, NM, 1991). These workers cite the case of Ohio versus EPA, in which a federal appeals court found the state responsible for testing computer models used to set emission limits on electric power plants [*U.S. Law Week* **54**, 2494 (1986)]. The court found the government liable because it had made no effort to determine the reliability of the model used. Given that the legal necessity of model testing has been established, what claims are justified on the basis of such tests?
2. L. F. Konikow and J. D. Bredehoeft, *Adv. Water Resour.* **15**, 75 (1992).
3. H. F. Wang and M. P. Anderson, *Introduction to Groundwater Modeling: Finite Difference and Finite Element Methods* (Freeman, San Francisco, 1982).
4. M. P. Anderson and W. M. Woessner, *Applied Groundwater Modeling: Simulation of Flow and Advective Transport* (Academic Press, New York, 1992).
5. The volume of *Adv. Water Resour.* **15** (1992) is entirely dedicated to the discussion of validation and verification of computer models.
6. In our experience, such claims are particularly abundant in cases in which an obvious public policy interest is at stake, such as in work surrounding the proposed high-level nuclear repository at Yucca Mountain, Nevada. Examples include N. Hayden, "Benchmarking: NNMSI flow and transport codes: Cove 1 Results" (Sandia National Laboratories, Albuquerque, NM, 1985); K. Stephens *et al.*, "Methodologies for assessing long-term performance of high-level radioactive waste packages," NUREG CR-4477, ATR-85 (5810-01) 1 ND (U.S. Nuclear Regulatory Commission, Washington, DC, 1986); T. Brikowski *et al.*, "Yucca Mountain program summary of research, site monitoring, and technical review activities," (State of Nevada, Agency for Projects–Nuclear Waste Project Office, Carson City, Nevada, 1988); L. Costin and S. Bauer, "Thermal and mechanical codes first bench-mark exercise, Part I: Thermal analysis," SAND88-1221 UC 814, (Sandia National Laboratory, Albuquerque, NM, 1990); R. Barnard and H. Dockery, " 'Nominal Configuration,' Hydrogeologic parameters and calculation results," vol. 1 of "Technical summary of the performance assessment calculational exercises for 1990 (PACE-90)," SAND90-2727 (Sandia National Laboratories, Albuquerque, NM, 1991).
7. For recent critiques of verification and validation in hydrology, see (*2*); J. D. Bredehoeft and L. F. Konikow, *Groundwater* **31**, 178 (1993). For a similar critique in geochemistry, see D. K. Nordstrom, *Eos* **74**, 326 (1993); *Proceedings of the Fifth CEC Natural Analogue Working Group Meeting and Alligator Rivers Analogue Project Final Workshop*, Toledo, Spain, 5 to 9 October 1992; H. von Maravic and J. Smellie, Eds. (EUR 15176 EN, Commission of the European Community, Brussels, 1994).
8. Two recent editorials dealing with the interface of modeling and public policy at Yucca Mountain are C. R. Malone, *Environ. Sci. Technol.* **23**, 1452 (1989); and I. J. Winograd, *ibid.* **24**, 1291 (1990).
9. For example, the *Random House Unabridged Dictionary* gives the first definition of verify as "to prove the truth of" (New York, 1973). Dictionary definitions of verify, validate, and confirm reveal the circularity present in common use, thus highlighting the imperative for consistent scientific usage.
10. This is the same as saying that there was an implicit ceteris paribus clause.
11. Godel questioned the possibility of verification even in closed systems [see E. Nagel and J. R. Newman, *Godel's Proof* (New York Univ. Press, New York, 1958)].
12. On the problem of verification in computer programming, see J. H. Fetzer, *Commun. ACM* **31**, 1048 (1988); *Not. Am. Math. Soc.* **36**, 1352 (1989); *Minds Mach.* **1**, 197 (1991).
13. This is equivalent to A. J. Ayer's classic definition of an analytic statement as one that is "true solely in virtue of the meaning of its constituent symbols, and cannot therefore be confirmed or refuted by any fact of experience" [A. J. Ayer, *Language, Truth and Logic* (Dover, New York, 1946; reprinted, 1952), p. 16]. Analytic statements are verifiable because "they do not make any assertion about the empirical world, but simply record our determination to use symbols in a certain fashion" (*ibid.*, p. 31). Also see A. J. Ayer, Ed., *Logical Positivism* (Free Press, New York, 1959).
14. If it were technically and economically possible to undertake exhaustive sampling on the scale of model elements, then we would run the risk of modifying the continuum properties we are trying to measure. The insertion of closely spaced drill holes into a porous medium may change the hydraulic properties of that medium. Furthermore, the dependent variables of the system—hydraulic head, solute concentration, and mineral assemblages—cannot be obtained at the model element scale. To know these parameters perfectly would be to mine out the region being modeled. This point is also made by C. F. Tsang [*Groundwater* **29**, 825 (1991)].
15. Recently, geochemists have made considerable progress on the kinetics of mineral reactions, but the point remains the same: In the absence of adequate data, many modelers assume that kinetics can be ignored. Similarly, in the absence of complete thermodynamic data, modelers necessarily extend available data beyond the range of laboratory information. To call this bad modeling is to miss the point: Data are never complete, inferences are always required, and we can never be certain which inferences are good and which ones are not as good.
16. An obvious example from atmospheric modeling is the notion of the mean global temperature. How do we measure the average temperature of the Earth? Our most basic data can be very deeply layered.
17. C. G. Hempel and P. Oppenheim, *Philos. Sci.* **15**, 135 (1948); C. G. Hempel, *Aspects of Scientific Explanation* (Free Press, New York, 1965); *Philosophy of Natural Science* (Prentice-Hall, Englewood Cliffs, NJ, 1966).
18. For this reason, C. F. Tsang (*14*) proposes that model evaluation should always be a step-by-step procedure.
19. This perspective refutes a simple Popperian account of falsification, wherein we are expected to throw out any model whose predictions fail to match empirical data. As many philosophers have emphasized, especially Imre Lakatos and Thomas Kuhn, scientists routinely modify their models to fit recalcitrant data. The question is, at what point do scientists decide that further modifications are no longer acceptable? Philosophers are still debating this question [T. S. Kuhn, *The Structure of Scientific Revolution* (Univ. of Chicago Press, Chicago, ed. 2, 1970); *The Essential Tension: Selected Studies in Scientific Tradition and Change* (Univ. of Chicago Press, Chicago, 1977); I. Lakatos, in *Criticism and the Growth of Knowledge*, I. Lakatos and A. Musgrave, Eds. (Cambridge Univ. Press, Cambridge, 1970), pp. 91–196; K. R. Popper, *The Logic of Scientific Discovery* (Basic Books, New York, 1959); *Conjectures and Refutations: The Growth of Scientific Knowledge* (Basic Books, New York, 1963)].
20. Nonuniqueness may arise on a variety of levels: Konikow and Bredehoeft (*2*) have emphasized the heterogeneity of the natural world; C. Bethke [*Geochim. Cosmochimi. Acta* **56**, 4315 (1992)] has emphasized the possibility of multiple roots to governing equations. Also see L. N. Plummer, D. L. Parkhurst, D. C. Thorstenson, *ibid.* **47**, 665 (1983); L. N. Plummer, "Practical Applications of Groundwater Geochemistry," First Canadian-American Conference on Hydrogeology (National Water Well Association, Worthington, OH, 1984), pp. 149–177; L. N. Plummer, E. C. Prestemon, D. L. Parkhurst, *U.S. Geol. Surv. Water Resour. Invest. Rep. 91-4078* (1991), p. 1.
21. Also referred to as the Duhem-Quine thesis, after Pierre Duhem, who emphasized the nonuniqueness of scientific explanation, and W. V. O. Quine,

who emphasized the wholistic nature of scientific theory. Both perspectives refute any simple account of the relation between theory and observation. The classic essays on underdetermination have been reprinted and critiqued in S. Harding, Ed., *Can Theories Be Refuted? Essays on the Duhem-Quine Thesis* (Reidel, Dordrecht, the Netherlands, 1976).

22. B. van Fraassen, *The Scientific Image* (Oxford Univ. Press, New York, 1980).
23. H. E. Longino [*Science as Social Knowledge* (Princeton Univ. Press, Princeton, NJ, 1990)] examines the role of personal and political preference in generating sex bias in scientific reasoning. Her point is that extraevidential considerations are not restricted to bad science but are characteristic of all science, thus making differentiation between "legitimate" and "illegitimate" preferences difficult.
24. For a counterargument, see C. Glymour, in *The Philosophy of Science*, R. Boyd, P. Gasper, J. D. Trout, Eds. (Massachusetts Institute of Technology Press, Cambridge, MA, 1991), pp. 485–500.
25. Ockham's razor is perhaps the most widely accepted example of an extraevidential consideration. Many scientists accept and apply the principle in their work, even though it is an entirely metaphysical assumption. There is scant empirical evidence that the world is actually simple or that simple accounts are more likely than complex ones to be true. Our commitment to simplicity is largely an inheritance of 17th-century theology.
26. In the early 1920s, a group of philosophers and scientists known as the Vienna Circle attempted to create a logically verifiable structure for science. Led by the philosopher Rudolf Carnap, the "logical positivists" wished to create a theoretically neutral observation language that would form a basis for purely logical structures, free of auxiliary assumptions, for all of science. Such logical constructions would be verifiable [R. Carnap, reprinted in *Logical Positivism*, A. J. Ayer, Ed. (Free Press, New York, 1959), pp. 62–81. Also see A. J. Ayer (1946), in (*13*). For a historical perspective on Carnap and logical positivism, see I. Hacking, *Representing and Intervening* (Cambridge Univ. Press, New York, 1983); R. Creath, Ed., *Dear Carnap, Dear Quine: The Quine-Carnap Correspondence and Related Work* (Univ. of California Press, Berkeley, 1990); and R. Boyd, in (*24*), pp. 3–35. The influence of the Vienna Circle on philosophy of science was profound; W. V. O. Quine has called Carnap "the dominant figure in philosophy from the 1930s onward" (in Creath, above, pp. 463–466). But in spite of Carnap's stature and influence, the philosophical program of "verificationism" collapsed resoundingly in the 1950s [P. Galison, *Sci. Context* **2**, 197 (1988); J. Rouse, *Stud. Hist. Philos. Sci.* **22**, 141 (1991)]. It was officially pronounced dead in the *Encyclopedia of Philosophy* in 1967 [K. R. Popper, *Unended Quest: An Intellectual Autobiography* (Collins, Glasgow, 1976), p. 87]. There now appears to be nothing in the philosophy of science that is as uniformly rejected as the possibility of a logically verifiable method for the natural sciences. The reason is clear: Natural systems are never closed.
27. For example, *Webster's Seventh New Collegiate Dictionary* (Merriam, Springfield, MA, 1963) gives the following definition of validation: to make legally valid, to grant official sanction to, to confirm the validity of (for example, an election). Random House similarly cites elections, passports, and documents [*Random House Dictionary of the English Language* (Random House, New York, 1973)].
28. For example, a widely used and extensively debugged package such as MODFLOW [M. G. McDonald and A. W. Harbaugh, *U.S. Geol. Surv. Tech., Water Resour. Invest.*, book 6 (1988), chap. A1, p. 1] or WATEQ [A. H. Truesdell and B. J. Jones, *Nat. Tech. Inf. Serv. PB2-20464* (1973), p. 1] might be valid, but when applied to any particular natural situation would no longer necessarily be valid. C. F. Tsang has argued that models should be validated with respect to a specific process, a particular site, or a given range of applicability. Unfortunately, even with such a degree of specificity, the elements of the model (the conceptualization, the site-specific empirical input parameters, the estimated temperature range) are still underdetermined. Furthermore, he notes that establishing "the range of application" of a model cannot be done independently of the desired performance criteria. "There is the possibility that a performance criterion could be defined in such a way that the quantity of interest can never be predicted with sufficient accuracy because of intrinsic uncertainties in the data . . . Thus, one has to modify the performance criterion to something more plausible yet still acceptable for the problem at hand" (C. F. Tsang, in (*14*), p. 827]. But this conclusion begs the question, Who decides what is plausible and what is acceptable?
29. "Environmental Assessment: Yucca Mountain Site, Nevada Research and Development Area, Nevada," vol. 2 of *U.S. Department of Energy DOE/RW-0073* (Office of Civilian Radioactive Waste Management, Washington, DC, 1986). This definition conflates the generic numerical simulation code with the site-specific model. A site-specific model might accurately represent a physical system, but there is no way to demonstrate that it does. A code is simply a template until the parameters of the system are put in, and therefore could not, even in principle, accurately represent a physical system.
30. "Radioactive waste management glossary," IAEA-TECDOC-264 (International Atomic Energy Agency, Vienna, 1982). A recent summary of European work in this area in the context of radioactive waste management is given by P. Bogorinski *et al.*, *Radiochim. Acta* **44/45**, 367 (1988).
31. In defining model "validation," these workers use the descriptor "adequate" rather than "good," presumably because they recognize the difficulty of defining what constitutes a "good" representation. They propose that a model need only be "adequate" for a "given purpose," in this case compliance with federal regulations. But this definition begs the question of whether the regulations are adequate. Furthermore, because these workers recognize that models cannot be validated but refuse to relinquish the term validation, they end up with an almost incomprehensible statement of their goals: "[M]odels can never be validated, therefore validation is a process of building confidence in models and not providing 'validated' models" [P. A. Davis *et al.*, in (*1*), p. 8].
32. For example, in the guidelines of the U.S. Nuclear Regulatory Commission Radioactive Waste Management Program (NUREG-0865) (U.S. Nuclear Regulatory Commission, Washington, DC, 1990), "verification" of a code is described as "the provision of an assurance that a code correctly performs the operations it specifies. A common method of verification is the comparison of a code's results with solutions obtained analytically." However, a certain confusion in the literature over terminology is made evident by comparison of Anderson and Woessner (*4*) with Wang and Anderson (*3*). Previously, Anderson had referred to this process as validation, and more recently, and more misleadingly, as verification.
33. Admittedly, computer programmers engage routinely in what they call program "verification." However, the use of the term "verification" to describe this activity has led to extremely contentious debate [see Fetzer (1988), in (*12*) and letters in response in *Commun. ACM* **32** (1989)]. One striking feature of "verification" in computer science is that it appears to be motivated, at least in part, by the same pressure as in the earth science community: a demand for assurance of the safety and reliability of computer programs that protect public safety, in this case, those controlling missile guidance systems (*ibid.*, p. 376). For an interesting historical paper on the problem of establishing certainty in the manufacture of weapons systems, see G. Bugos, *Soc. Stud. Sci.* **23**, 265 (1993).
34. A good example of van Fraassen's concept is the view expressed by G. de Marsily, P. Combes, and P. Goblet [*Adv. Water Resour.* **15**, 367 (1992)], who claim that they "do not want certainty, [but] will be satisfied with engineering confidence. [W]e are only [trying] to do our level best." This is a commendably honest approach but one that will invite a very different public reaction than claims about "verified" models.
35. Using post-audits of "validated" models, Konikow and co-workers have shown that even models that produce a good history match of past data often do terribly when extended into the future [L. F. Konikow and J. D. Bredehoeft, *Water Resour. Res.* **10**, 546 (1974); L. F. Konikow, *Groundwater* **24**, 173 (1986); ——— and M. Person, *Water Resour. Res.* **21**, 1611 (1985); L. F. Konikow and L. A. Swain, in *28th International Geological Congress Selected Papers on Hydrogeology*, V. H. Hiese, Ed. (Hanover, West Germany, 1990), pp. 433–449). Typically, this occurs either because the conceptualization of the system built into the numerical model was incorrect or because modelers failed to anticipate significant changes that subsequently occurred in the system (for example, changes in climatic driving forces). Post-audit studies by these and other workers have been reviewed by M. P. Anderson and W. W. Woessner [*Adv. Water Resour.* **15**, 167 (1992)]. Of five studies reviewed, not one model accurately predicted the future. In several cases, models were calibrated on the basis of short-duration data sets that inadequately described the range of natural conditions possible in the system. This issue of temporal variation becomes particularly important for modeling the long-term disposal of nuclear wastes. Changes in the geological conditions of the repository site, which could lead to changes in the dynamics and structure of the system, are not only possible but, given enough time, almost certain.
36. Various philosophers including A. J. Ayer, W. V. O. Quine, I. Lakatos, and T. S. Kuhn have questioned whether we can in fact prove a hypothesis false. Ayer emphasized that refutations, no less than confirmations, presuppose certain conditions [Ayer, 1946 (*13*, especially p. 38)]. Quine, Lakatos, and Kuhn emphasized the wholistic nature of hypotheses and the flexible options for modifications to "save the phenomena" (*19, 21*). However, none of these moves really undermines Popper's argument that it is still possible in principle to prove a theory false, but not possible even in principle to prove a theory true [Popper (*19*)].
37. Note that this is just one view. Many philosophers have disputed the hypothetico-deductive model.
38. The notion of diversity in confirmation helps to explain why it is important to test a model in a wide variety of circumstances—including those that may appear quite different from the expected circumstances at the modeled site—despite apparent arguments to the contrary. For example, Davis and co-workers (*1*) have argued that testing the performance of a model in areas not relevant to regulatory compliance is a waste of resources and can lead to the needless rejection of models that are adequate to the task at hand. While this may sometimes be the case, confirmation theory suggests that successful testing of a model in a variety of domains provides important support for the conceptualization embodied in the model. Failed tests help to establish the limits of model adequacy, and may cast legitimate doubt on the model conceptualization of the physical or chemical processes involved.
39. In his classic account of the principle of verification, A. J. Ayer [(1946), in (*13*)] opened the door to undermining his own position by recognizing that empirical statements could never be proved certain but only probable. He called this condition "weak verification," an obvious oxymoron. In hindsight it is easy to see that "weak verification" is probabilistic confirmation [Ayers (1946), in (*13*), pp. 99–100 and 135–136]. Popper preferred the term "corroboration" to emphasize that all confir-

mation is inherently weak (Popper, 1959 (*19*)]. For a recent perspective on probabilistic confirmation, see A. Franklin and C. Howson, *Stud. Hist. Philos. Sci.* **19**, 419 (1988); and C. Howson and P. Urbach, *Scientific Reasoning: The Bayesian Approach* (Open Court, La Salle, IL, 1989).
40. Carnap therefore argued that all inductive logic was a logic of probability [R. Carnap, in *The Problem of Inductive Logic*, I. Lakatos, Ed. (North Holland, Amsterdam, 1968), pp. 258–267]. Confirming observations give us warrant for a certain degree of belief.
41. An example is the evidence of faunal homologies in Africa and South America, before the acceptance of plate tectonic theory. These data, which were used as an early argument in favor of continental drift, were considered to be equally well explained by the hypothesis of land bridges [N. Oreskes, *Hist. Stud. Phys. Sci.* **18**, 311 (1988)].
42. An obvious example of this is Ptolemaic astronomy, which was extremely well confirmed for centuries and then overturned completely by the Copernican revolution. See T. S. Kuhn, *The Copernican Revolution* (Harvard Univ. Press, Cambridge, MA, 1957). Indeed, every scientific revolution involves the overturning of well-confirmed theory. See I. B. Cohen, *Revolution in Science* (Belknap Press, Cambridge, MA, 1985).
43. Konikow and Bredehoeft (*2*), on the basis of their extensive experience with both scientists and government officials, emphasize that the language of verified and validated models is typically interpreted to mean that the models under discussion are, in essence, true. It is also clear that this is the intent of many authors who claim to base results on "validated" models.
44. We have never seen a paper in which the authors wrote, "the empirical data invalidate this model."
45. Another example is found in the environmental assessment overview for Yucca Mountain (*29*, p. 4). The task of site selection, as defined in this report, consisted of "evaluat[ing] the potentially acceptable sites against the disqualifying conditions. . . ." The authors concluded that the Yucca Mountain site was "not disqualified." That is, the null hypothesis is that the site is safe; the burden of proof is on those who would argue otherwise.
46. N. Cartwright, *How the Laws of Physics Lie* (Clarendon Press, Oxford, 1983), p. 153.
47. This article was prepared for a session on hydrological and geochemical modeling in honor of David Crerar at the American Geophysical Union, May 1993. We thank the organizers, A. Maest and D. K. Nordstrom, for inviting us to prepare this article; J. H. Fetzer, L. Konikow, M. Mitchell, K. Nordstrom, L. Sonder, C. Drake, and two reviewers for helpful comments on the manuscript; and our research assistant, D. Kaiser. We dedicate this paper in appreciation of the work of David Crerar.

**RESEARCH ARTICLE**

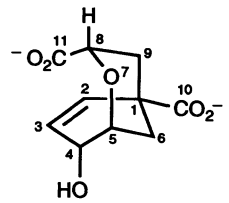
# Routes to Catalysis: Structure of a Catalytic Antibody and Comparison with Its Natural Counterpart

Matthew R. Haynes, Enrico A. Stura, Donald Hilvert, Ian A. Wilson

**The three-dimensional structure of a catalytic antibody (1F7) with chorismate mutase activity has been determined to 3.0 Å resolution as a complex with a transition state analog. The structural data suggest that the antibody stabilizes the same conformationally restricted pericyclic transition state as occurs in the uncatalyzed reaction. Overall shape and charge complementarity between the combining site and the transition state analog dictate preferential binding of the correct substrate enantiomer in a conformation appropriate for reaction. Comparison with the structure of a chorismate mutase enzyme indicates an overall similarity between the catalytic mechanism employed by the two proteins. Differences in the number of specific interactions available for restricting the rotational degrees of freedom in the transition state, and the lack of multiple electrostatic interactions that might stabilize charge separation in this highly polarized metastable species, are likely to account for the observed $10^4$ times lower activity of the antibody relative to that of the natural enzymes that catalyze this reaction. The structure of the 1F7 Fab′-hapten complex provides confirmation that the properties of an antibody catalyst faithfully reflect the design of the transition state analog.**

The mammalian immune system has been successfully exploited by chemists to create antibody molecules with tailored catalytic activities and specificities. Haptens designed to mimic the key stereoelectronic features of transition states can induce antibodies capable of catalyzing various chemical transformations, ranging from simple hydrolyses to reactions that lack physiological counterparts or are normally disfavored (*1*). The ability to create novel active sites in this way (*2*) permits systematic exploration of the basic principles of biological catalysis and, through comparison with naturally occurring enzymes, evaluation of alternative catalytic pathways for particular reactions. In the absence of structural information, it is difficult to determine precisely the extent to which the transition state analog dictates the catalytic characteristics of the induced antibody. Thus, detailed knowledge of the mode of transition state analog binding by antibodies should facilitate the further development, through rational redesign, of both transition state analogs and first-generation catalytic antibodies.

The unimolecular conversion of (−)-chorismate into prephenate (Fig. 1) was one of the first nonhydrolytic reactions to be catalyzed by an antibody (*3*, *4*). This concerted transformation, formally a Claisen rearrangement, has been intensively studied as a rare example of a biologically relevant pericyclic reaction (*5*–*11*). In microorganisms and higher plants prephenate production is the committed step in the biosynthesis of tyrosine and phenylalanine, and the enzyme chorismate mutase accelerates this reaction by more than 2 million. Although the precise factors that contribute to the efficiency of the enzyme are still poorly understood, it is known that the uncatalyzed reaction occurs through an asymmetric chairlike transition state **2** in which carbon-oxygen bond cleavage precedes carbon-carbon bond formation (*7*, *8*). In aqueous solution the flexible chorismate molecule preferentially adopts the extended pseudodiequatorial conformation **1a** and must be converted to the higher energy pseudo-diaxial conformer **1b** on the way to the transition state (*9*). Binding sites that are complementary to the compact transition state species (and the corresponding substrate conformer) would therefore be expected to increase substantially the probability of reaction. The favorable entropy of activation ($\Delta\Delta S^{\ddagger}$ = 13 cal $K^{-1}$ $mol^{-1}$) of the enzyme-catalyzed process compared to the spontaneous thermal rearrangement is consistent with this idea (*6*), as is the observation of strong enzyme inhibition by the conformationally restricted *endo*-oxabicyclic dicarboxylic acid **4** which approximates the structure of **2** (*12*). Stabilization of any charge separation in the transition state through electrostatic or hydrogen bonding interactions might also contribute to the potency of the enzyme (*13*).




**4**

M. R. Haynes, E. A. Stura, and I. A. Wilson are in the Department of Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037. D. Hilvert is with the Departments of Chemistry and Molecular Biology, The Scripps Research Institute, La Jolla, CA 92037.