

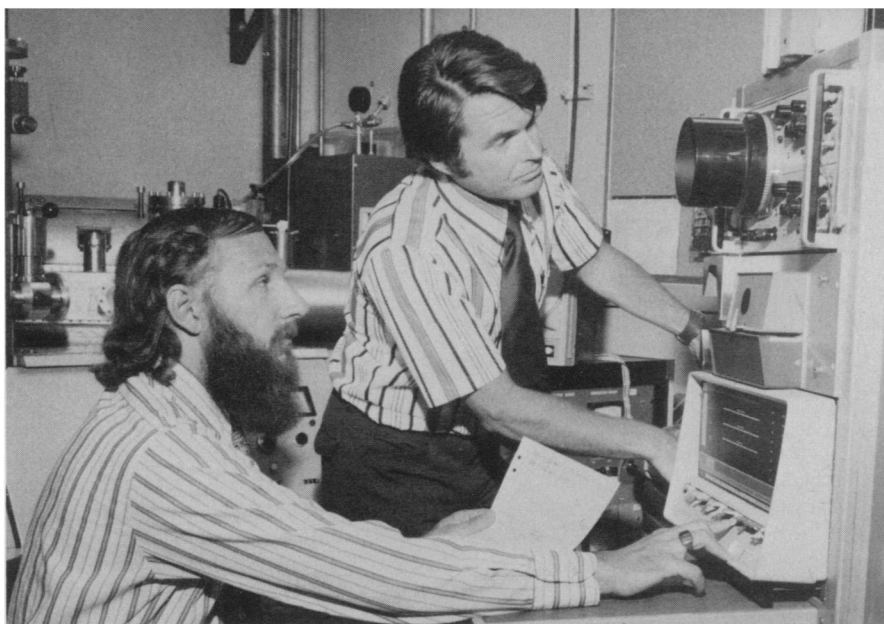
Predicting chemical properties by computer

One of the amazing things about computing machinery is that although it is designed and programmed by human minds, its abilities often surpass those of its inventors. A striking example is the application of artificial intelligence to pattern recognition in the analysis of chemical data. This approach is being applied at the Lawrence Livermore Laboratory by Bruce R. Kowalski (professor of chemistry at Colorado State University and a consultant to LLL) and Charles F. Bender. They believe, as they state in the Aug. 9 JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, that it will find wide application in the solution of problems that now provide much heavy labor for chemists.

The method confronts a problem basic to all sciences. The scientist desires to know something about a physical or chemical system that he cannot measure directly (example: The structure of a molecule or the reactivity of a group of solutions) but which he believes he can find out from a set of measured quantities. One approach to the question is analytic or algebraic. One can search by using algebraic methods for relationships among the known quantities and the unknown ones that can be expressed as mathematical functions or formulas. The formulas can then be solved for the unknowns. Sometimes the analysis is helped by a known or hypothesized theoretical relationship between knowns and unknowns, but often, especially in chemistry, even theory is lacking. Kowalski and Bender present the starkest case: "Given a set of objects and a list of measurements made on those objects is it possible to find and/or predict a property of the objects that is not directly measurable but is known to be related to the measurement *via* some *unknown* relationship?"

Pattern recognition can do it. The technique is essentially a graphic method for solving problems. Humans continually use graphic methods to solve problems in two or three dimensions. Suppose someone wished to find whether there was a relationship between population and land area for a selected group of cities. He could make a graph on which one axis was population and the other area, and plot his cities. Any pattern, trend or relationship would show up immediately on the graph, and, if the investigator wished, he could then go back and derive the exact algebraic relation if there was one. If not, a great deal of labor would have been saved.

But human powers of visualization fail above three dimensions. Since every different kind of data input represents a separate dimension in this kind of



Lawrence Livermore Laboratory

Kowalski and Bender using computer for pattern-recognition analysis of data.

problem-solving, human operators are restricted to only the simplest problems. They must attack multidimensional problems algebraically. Computers, in contrast, can visualize in ten, a hundred or a thousand dimensions, and artificial intelligences are now good enough that they can be taught or teach themselves to recognize patterns in these multidimensional problems and to make classifications and draw conclusions from them.

In the JACS article Kowalski and Bender present a six-dimensional test example. Given are 68 elements of the periodic table. Each has a representative oxide that is either basic, acidic or amphoteric (capable of reacting as either acid or base). The test was to see whether the artificial intelligence, using six pieces of information about each element—including such things as its most important valence and its melting point—could find a pattern that separated the acids from the bases and classified the amphoteric oxides on one side or the other. None of the six properties alone could make the separation, but taken together they could. The computer successfully solved the problem.

For chemists who like to stay in charge of their own work Kowalski and Bender point out that the method does not remove the scientist from data analysis. The chemist can supervise the computer's work by a technique known as nonlinear mapping that translates the many-dimensional patterns into two dimensions. The important thing in determining patterns is the distance between two points. The mapping from many dimensions to two cannot be made without some error in the distances, but there are techniques for minimizing the error and thus transferring the basic outline of the pattern. Because of the inherent error, however, the investigator has to be careful in

classifying borderline cases from the two-dimensional map.

The great advantage of pattern recognition is speed. "In chemistry today, data are compiled in enormous quantities," notes Kowalski. "Some instruments are capable of producing 10,000 to 20,000 units of information per second." Problems that used to be insoluble can now be solved, and those that used to take days can now be done in minutes.

Another advantage is that the computer can recognize which data are redundant or not relevant to the question at hand and throw them out. Investigators working carefully by analytical methods usually had to carry all the data inputs through the calculation. To determine the structure of hydrocarbon molecules, for example, conventional methods employ 500 different parameters. The artificial intelligences at LLL have been scoring 95 percent accuracy using only five to ten.

Though the method is still in an experimental stage, it has been used successfully in a number of trial cases. In one such it helped some archaeologists from the University of California determine the origins of some Indian artifacts found in northern California. The archaeologists had tried unsuccessfully to correlate concentrations of elements in the artifacts with those in known source sites. The LLL pattern-recognition programs eliminated a lot of the information as irrelevant and used the remainder to find positive correlations. The result helped trace old Indian trade routes.

Kowalski and Bender expect pattern recognition will have a broad application in chemistry and other fields. One indication they are right is an interest on the part of scientists in LLL's biomedical division, who want to use it to identify and classify chromosomes. □