J. Zupan and J. Gasteiger

Neural Networks in Chemistry and Drug Design

WILEY-VCH

Jure Zupan and Johann Gasteiger

Neural Networks in Chemistry and Drug Design

Second Edition

WILEY-VCH

Weinheim · New York · Chichester · Brisbane · Singapore · Toronto

Professor Jure Zupan
National Institute of Chemistry
P.O. Box 34-30
SI-1001 Ljubljana, Slovenia

Professor Johann Gasteiger
Computer-Chemie-Centrum
Inst. für Organische Chemie
Universität Erlangen-Nürnberg
Nägelsbachstr. 25
D-91052 Erlangen, Germany

This book was carefully produced. Nevertheless, authors, editors, and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

The cover shows a three-dimensional molecular model in front of a Kohonen neural network. The weight distribution in the various layers of the network is indicated by color coding.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data: applied for

Deutsche Bibliothek CIP-Einheitsaufnahme

Zupan, Jure : Neural networks in chemistry for drug and design / Jure Zupan ; Johann Gasteiger.
- 2., unveränd. Aufl., - Weinheim ; New York ; Chichester ; Brisbane ; Singapore ; Toronto : Wiley-VCH, 1999
1. Aufl. u.d.T.: **Zupan, Jure: Neural networks for chemists**
ISBN 3-527-29779-0 (Softcover)
ISBN 3-527-29778-2 (Hardcover)

Printed an acid-free and chlorine-free paper.

Prepress: Satzrechenzentrum Kühn & Weyh, D-79110 Freiburg
Printing: Rheinhessische Druckwerkstätte, D-55232 Alzey
Bookbinding: Wilhelm Osswald & Co., D-67933 Neustadt
Cover Design: Dipl. Chem. Markus C. Hemmer, Computer-Chemie-Centrum, Universität Erlanger-Nürnberg

Printed in the Federal Republic of Germany.

For Breda, Ulrike, Nina,
Julia, Eva, and Michael.

Preface

Almost all fields of human endeavor, science in particular, have seen dramatic changes in recent years. In all domains of our life we are swamped with data, be it on the stock market, medical diagnosis, car crash tests or, to come to the point, in chemistry.

Throughout history, until some decades ago, the main problem scientists were confronted with in their scientific endeavor was how to obtain data. Measurements were time-consuming, not sensitive enough, expensive, required permanent presence of the experimenter, manual recording, etc. In addition, there was the problem of preparation of the adequate material, lack of proper techniques, special purpose equipment and technical support. Scientists had to handle a lot of unpleasant routine work just to obtain a few numbers. Robert A. Millikan meticulously carried out tens of thousands of experiments to determine the charge of the electron. Marie Curie put in tremendous efforts to isolate a minute quantity of radium.

Today, "thanks" to the high amount of instrumentalization of science, the main problem is not how to gather data, but how to get rid of most of them. Unfortunately, only a small amount of data produced by the computerized instruments and stored in computer memory are really relevant to the problem. The sad truth is that the valuable information scientists are seeking can often be extracted much harder from the myriads of computer bits than from a small number of data measured in carefully planned and manually executed experiments.

Because everybody knows that the good old times are not going to return and the deluge of data will not disappear in our scientific work, it is better to be prepared to handle large quantities of data. In industrial laboratories thousands of products are analyzed each day, in production plants thousands of recipes are handled, monitored and slightly changed to meet the control requirements. Organic chemists are performing vast series of chemical reactions. Pharmaceutical research groups are inspecting thousands of potential drugs; biologists are analyzing long strings of amino acids in numerous proteins; spectroscopists are comparing huge databases of spectra or working on multidimensional images produced by sophisticated spectrometers, etc.

It is not hard to find examples where large amounts of data must be handled to extract the vital information. The only way to adapt ourselves to this flood of data is to acquire new knowledge, to rethink the methods we are using, and to adapt our skill for data handling to the new situation.

With this book we want to introduce you into a rediscovered and reshaped old method called "neural networks". No doubt, the name is provocative. But so was the term "artificial intelligence" which decades ago raised the blood pressure of many otherwise very reasonable scientists. Things have settled down and today the term artificial intelligence is, if not fully admired and supported, at least accepted. The situation with artificial neural networks is not quite so settled and peaceful yet.

The first word of the term "neural networks", has a clear link with a "neuron", a nerve cell. This points further towards that part of the human body considered by many to be the one most distinguished in mankind: the brain. In this book we would like to make clear that the emphasis in the term "neural network" is not on the word "neural" but rather on the word "network".

Networking is one of the most important things in any organizational scheme: running a railroad company, a bank, a production plant, or handling scientific data. Networking the flow of data is a familiar concept to the electrical engineer, but is not so familiar to the chemist. In the context of this book, the word network always means an assembly of "little devices" called "neurons" which perform the same set of simple operations all the time. Hence the final output is not primarily the result of these simple operations but rather of the way how these "little devices" are linked together and how they change their internal parameters to adapt their individual outcomes to some external control or competition among other "neurons".

One source of the flexibility for handling data by neural networks stems from their "architecture", i.e., from the number of neurons it is composed of, how these neurons are interconnected, etc. The second reason why neural networks are so adaptable to different problems and applications is the possibility to implement both basic ways of learning: learning by oneself, unsupervised learning, and learning by a tutor, supervised learning.

These two concepts are probably the most important aspects of learning. Throughout the book, especially in the Part IV "Applications", great emphasis is put on the notion that **both** types of learning, supervised and unsupervised, are needed to accomplish most of the tasks met in applications of neural networks.

The book is composed as a textbook and we therefore first introduce the concepts of neural networks. The ultimate goal of the book is to bring you to a level of understanding that you will be able to apply neural networks to your problems either with a commercial neural network package or with a self-made program.

The first part of the book deals with the concept of neuron, transfer functions, "bias", etc. A comparison is made with the classical linear learning machine. Then, the linking of neurons to layers and the linking of layers of neurons among themselves is described. In the second part, one-layer neural networks are discussed. First, the Hopfield network and the ABAM (Adaptive Bidirectional Associative Memory) are explained and then the Kohonen network is explained in more detail. Kohonen learning is the most important unsupervised, or self-organizing learning scheme offered by neural networks. The two-dimensional map formed by the labels of objects that enter the training procedure as well as the maps which are formed within the individual Kohonen network of layers of weights are the most informative results of the Kohonen learning.

In Part III, multilayer neural networks and learning in those networks is discussed. The counter-propagation of targets and back-propagation of errors learning schemes are introduced. The counter-propagation network consists of two layers: the upper layer performing the Kohonen learning and the output layer performing the adaptation of weights to the targets which are input to the network from the counter-part side of the network, i.e., from the output side. The back-propagation of errors is the most widely used method of neural network learning. Today, at least in chemistry, more than 90 percent of all applications are made by back-propagation of error learning. Therefore, this method is discussed in more detail.

For each learning strategy or neural network architecture we worked out a simple example which in most cases is not from the field of chemistry. In this way, by selecting examples not linked to any specific type or field of application we have been pursuing two goals:

first, such a broad choice gives us much more flexibility in selecting appropriate examples, and second, the book might attract a much wider audience if the methods are accompanied by examples that are not tied to only one field of application – chemistry in our case.

Actually, the same strategy for the selection of examples was followed in the "Applications" part. This time the field was of course confined to be chemistry, but we have tried to show you applications from the various domains of chemistry. Today (November 1992) the Chemical Abstract Services lists nearly 500 articles published under the code word "neural networks" and are related to chemistry. There are many interesting applications and the choice of examples must, of course, be an arbitrary one. Other authors might have made a completely different choice. However, there are a few general subfields in chemistry where neural networks are applied more than in others.

The first among such fields is process control and fault detection in processes. The complex relationships between the controlled and the manipulated variables of the process, together with the many data available from the constant monitoring of a process offer an ideal "playground" for testing and applying neural networks. The concepts of "moving window" and of the binary representation of discrete state variables are introduced. These many aspects have made the chapter on process control and fault detection rather large.

The other field where the back-propagation of error learning has seen a fast and early momentum is the prediction of the secondary structure of proteins. The first article by Qian and Sejnowski was already published in 1988. There is no doubt that the introduction of the neural network approach to the determination of protein structure is an achievement by itself. First, because it "transplanted" a method from one scientific discipline (speech reproduction) to a completely different field, showing the power of interdisciplinary knowledge and reasoning. Secondly, because the newly introduced method improves the previously obtained predictions, and thirdly, because it opens new perspectives in posing new questions in the field. For example: What is the most appropriate representation of amino acid sequences? How can the long-range influences be taken into account? What variety of problems can be solved by the method? etc.

The third area in chemistry where the neural network approach seems to be very promising, is structure elucidation by different spectroscopic methods. The problem of establishing reliable correlations between different types of spectra (infrared, mass, ^1H NMR, ^{13}C NMR, etc.) and the chemical structure of the corresponding compound has been around so long, that there is no wonder that the neural network approach immediately attracted the attention of spectroscopists and those working on computerized structure elucidation processes. It is interesting to note that spectroscopy was one of the first areas where not only the back-propagation of errors learning, but other learning methods, like Kohonen learning, and neural network architectures like ABAM and Hopfield networks were used.

Other examples in the book come from different areas such as classification of objects into several categories, optimization of recipes and procedures, quantitative structure-activity relationship (QSAR) studies, reactivity of bonds, aromatic electrophilic reactions, and mapping of electrostatic potential into a two-dimensional plane.

We hope that the variety of examples and the flexibilities of the neural network learning methods and their architectures will encourage the user to think about:

- how to select the best neural network learning method or even to simultaneously try several of them,
- what neural network architecture and which parameters to choose to obtain as much information from the data available as possible, and above all:
- how to interpret the results obtained to pinpoint the required information best.

The above points are much more important than to think about how to find a dataset that is suitable for a neural network package from the shelf in order to publish a paper as quickly as possible.

At the end we would like to extend our special thanks to four of our coworkers: Ms. Vera Simon, Ms. Marjana Novic, Mr. Xinzhi Li and Mr. Marjan Tusar who helped us by working out many examples and calculations presented in this book. Of course, the thanks go as well to all other members and coworkers of the Laboratory for Chemometrics at the National Institute of Chemistry in Ljubljana and the Laboratory for Computer Chemistry at the Organisch-Chemisches Institut der Technischen Universität München who helped us by supporting the friendly and stimulating atmosphere in the laboratories while exchanging their ideas and sharing their enthusiasm with us.

Special thanks go to Miss Natalia Berryman who restlessly transferred our sometimes non-executable ideas into her fine drawings. Altogether she managed to complete more than 220 final pictures. This actually means at least twice that number of sketches and provisional ideas.

Thanks are also due to Miss Elisabeth Lohof who read the entire manuscript and made a number of valuable suggestions in English style. In addition, she composed the final version of the text. We also want to thank Dr. Hans Lohninger, Technische Universität Wien, for making a number of valuable comments on style, contents and didactic improvements of the manuscript.

Our families have suffered most from our excessive absence, and it is our sincere hope that the sacrifices they made will be at least partially rewarded by us being with them more in the future.

The roots of this book go back to a meeting in the cafeteria of Beijing airport between a physicist from Slovenia and an organic chemist from Bavaria. This was followed by long scientific discussions in the tiny Tyrolian village of Hochfilzen.

The idea for writing this book developed from a course on neural networks J. Z. gave as visiting professor at the Technische Universität München. The Bundesminister für Forschung und Technologie, Germany is gratefully acknowledged for supporting J. Z.'s three-month/year position for three consecutive years at the Laboratory for Computer Chemistry in Garching. Our thanks go to Dr. Jan Michael Czermak of BMFT and Dr. Volker Schubert of GMD-PTF for arranging for this position of a project professor.

J. G. would like to express his thanks to the Ministry for Science and Technology of Slovenia which enabled him to have this close cooperation and a one month leave in Ljubljana to work on the project.

Jure Zupan Johann Gasteiger

December 1992

Preface to the Second Edition

It attests to the importance of neural networks in general, and their application to chemical problems in particular, that a second edition of this book becomes necessary. The field has grown: see Table 9-1, that gives the number of publications for the application of neural networks in chemistry. And it has matured: after the initial hype where we saw many problems that could just as well have been tackled with more traditional statistical or pattern recognition methods, the specific advantages of neural networks have become clearer. We also see the use of a more diverse set of neural network methods, not only feed-forward networks trained by the back-propagation algorithm but also other methods, particularly, Kohonen networks.

From the feedback that we have obtained we felt that not much needs to be done to Parts I – III (Chapters 1 – 8). We found a few places that needed minor corrections and some clarification.

However, we have quite heavily extended Part IV „Applications" (Chapters 9 – 21) to account for the surge of new applications. In this endeavor we have largely concentrated on work from our two research groups because much is an extension of work that was emerging in the first edition and has now come to full bloom. Heavy emphasis is placed on examples from drug design because of the strong interest in this area.

A book is somehow outdated when it appears in print. In order to do something against this fate, we have decided to establish a web-site (*http://www2.ccc.uni-erlangen.de/ANN-book/*) that allows us to keep the reader updated on recent developments in the area of neural networks in chemistry. This gives us also the possibility to provide additional material such as giving access to programs and data sets.

We were fortunate that quite a few people have shared our enthusiasm for neural networks. In particular, our coworkers and collaborators have ventured with us into this exciting world of applying neural networks to chemical problems.

In the past six years since the first edition of this book has appeared both our groups have held many collaborations with various Laboratories in Slovenia, Germany as well as in the rest of Europe.

J. Z. would like to thank his closest coworkers Dr. Marjana Novic, Dr. Marjan Vracko-Grobelsek, and Dr. Marko Perdih for their contributions to various aspects of their common research. It is a pleasure to thank Dr. Darinka Brodnjak-Voncina from the University of Maribor, Dr. Itziar Ruisanchez and Prof. Xavier. F. Rius from University Rovira i Virgili, Tarragona, Prof. Giuseppina G. Gini from Politecnico di Milano, and Dr. Emilio Benfenati from Instituto Mario Negri, Milano, for sharing the common interests in artificial neural networks. For a scientist, the most rewarding moment comes when the research laboratories in the industry are implementing the new methods in their daily work. In this respect the thanks go to Dr. Nineta Majcen and M.Sc. Karmen Rajer-Kanduc from Cinkarna, doo. Celje, to Dr. Livija Tusar and M.Sc. Nevenka Leskovsek from Color, doo. Medvode, and M.Sc. Ales Brglez, Gorenje doo, Velenje.

J. G. wants to thank his coworkers Dr. Bruno Bienfait, Dr. Lingran Chen, Sandra Handschuh, Markus Hemmer, Dr. Xinzhi Li, Oliver Sacher, Dr. Jens Sadowski, Christof Schwab, Dr. Jan Schuur, Dr. Paul Selzer, Dr. Valentin Steinhauer, Andreas Teckentrup, and Dr. Markus Wagener for their work in showing the broad scope of applications that neural networks can find in chemistry. In addition, he appreciated the collaboration with Dr. Soheila Anzali, Darmstadt, Prof. Ulrike Holzgrabe, Bonn and Dr. Jarek Polanski, Katowice, Poland that led to important new contributions in methodology and applications.

Again, the book could directly be produced from the Postscript file that we submitted thanks to the careful editing efforts of Angela Döbler and Oliver Sacher.

Jure Zupan Johann Gasteiger

January 1999

Contents

Part II One-Layer Networks

Part III Multilayer Networks

Part IV Applications