14 The Electrophilic Aromatic Substitution Reaction

learning objectives:

- structural factors influencing electrophilic aromatic substitution
- representation of molecular structures by a connection table
- three forms of structure representation:

  1) connection table

  2) specifying formal charges for all atoms of the ring

  3) use of electronic and steric parameters **specific to a reaction site**

14.1 The Problem

The substitution of a hydrogen atom of a monosubstituted benzene derivative by another group (e.g., nitro, halogen, acyl, alkyl) is a remarkable reaction on many grounds. First, it is of great industrial importance, many basic chemicals being produced by this reaction. Second, nearly all these reactions occur by the same fundamental mechanism: the attack of an electrophilic group, an agent with electron demand, on the benzene derivative.

This electrophilic substitution of a proton by another group can occur, in principle, in three different positions: *ortho* (*o*), *meta* (*m*), and *para* (*p*) (Figure 14-1).

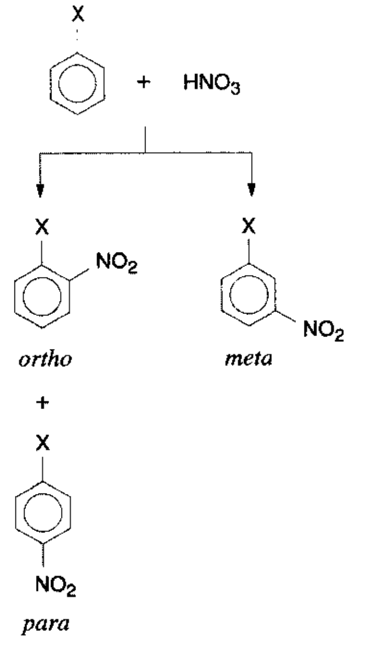


Figure 14-1: Product distribution in the electrophilic aromatic substitution reaction.

The rationalization for the relative reactivities of various monosubstituted benzene derivatives and for the distribution of *ortho*-*meta*-, and *para*-substituted products is a paradigm of the methods of physical organic chemistry. In undergraduate organic chemistry courses, it is a standard case for explaining the influence of various electronic effects, particularly the inductive and resonance effects, on the reactivity and selectivity of a chemical reaction. In spite of this apparently well-settled matter, there are still a lot of unanswered questions. The ratio of *ortho* to *para* product is hard to predict, although it is thought to be largely influenced by steric effects. Reaction conditions, particularly the solvent, can have a drastic influence on product distribution; this is hardly understood at all.

Furthermore, taking what we know about isomer distribution in the reaction of monosubstituted benzene derivatives, and using it to predict product ratios obtained from polysubstituted benzene derivatives is not very successful (Figure 14-2).

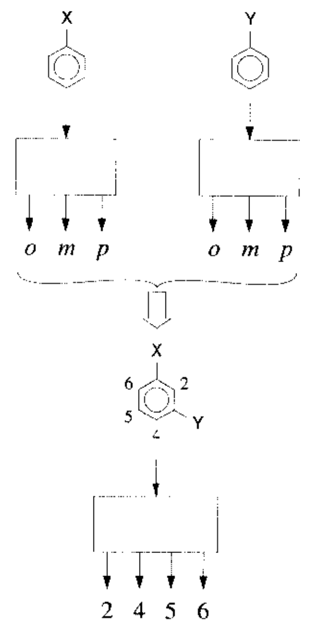


Figure 14-2: Predicting product distribution in the electrophilic aromatic substitution of disubstituted benzene derivatives, based on a knowledge of product distribution in the EAS of the two corresponding monosubstituted benzene derivatives.

Nevertheless, at an elementary level, electrophilic aromatic substitution (EAS) shows some distinct characteristics. Most substituents on the aromatic ring can be classified into two categories: groups that donate electron density, either by an inductive or a mesomeric effect (Figure 14-3) primarily give *ortho* and *para* substitution, while groups that are mesomeric electron acceptors (Figure 14-4) react by direct substitution at the *meta* position.

Since the factors that determine the *ortho*/*para* product ratio are much less understood, the yields of these two products are quite often lumped together.

14.2 The Data

For the present example, we will use the work and data of Elrod, Maggiora, and Trenary (see Reference 14-1), who investigated the distribution of products in the nitration of a series of monosubstituted benzene derivatives. For reasons just mentioned, the yields of *ortho* and *para* product were combined; thus, they worked with the ratio of *meta* product to *ortho* plus *para* product.

Table 14-1 shows ten out of the 50 substituents used in this example, arranged in order of increasing percentage of *meta* product obtained in the electrophilic reaction.

One of the purposes of this example is to discuss the problem of finding an appropriate coding scheme for the structures of the compounds or substituents to be input into the neural networks. Two coding schemes are investigated.

Any coding scheme for this problem must, of course, be capable of representing these substituent effects. The first approach takes the *partial atomic charges* on the six carbon atoms of the ring (as

no.	substituent	yield of *meta* product
1	$-NH_2$	0
2	$-NHCOCH_3$	2
3	$-CH_3$	4
5	$-CH_2NH_2$	10
5	$-CH_2COOH$	22
6	$-SiMe_3$	40
7	$-CCl_3$	64
8	$-CONH_2$	70
9	$-COOH$	80
10	$-SO_2CH_3$	100

Table 14-1: Ten substituents, and the corresponding yield of the *meta* product.

calculated by a semiempirical quantum mechanical method, MOPAC – see Reference 14-3) as a representation of the electronic effects influencing product distribution (Figure 14-5).

In the other scheme, the structure of a substituent is represented by a (5 x 5)-connection table, in which each row represents one atom. The five entries are the atomic number, the ID number of this atom, the ID number of the atom it is bonded to, the bond order of this bond and the formal charge on this atom (Figure 14-6).

For each nonhydrogen atom of the substituent, we need a new row, starting with the atom directly bonded to the ring. If the substituent has fewer than five nonhydrogen atoms, the remaining rows are filled with zeros; if the substituent contains more than five atoms, the representation is cut after five.

Next, we need a canonicalization for the connection table, i.e., a set of rules valid in **all** cases to produce a **unique numbering** scheme for the atoms. Particularly in larger connection tables the atoms in a given substituent can be numbered in different ways, so that different inputs will be produced in the neural network, and consequently many different outputs will result.

14.3 The Network

With two essentially different representations for monosubstituted benzene derivatives, we will need two different architectures for the neural network. The first representation leads to an architecture with

X: OH, NH_2, Cl, Br, etc.

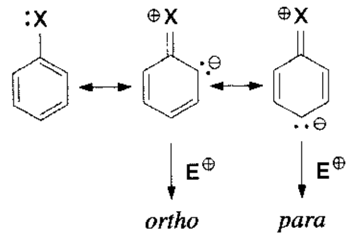


Figure 14-3: An atom X with a free electron pair can donate electron density to the *ortho* (o) and *para* (p) positions by a so-called *plus mesomeric effect*, and thus direct the attack of the electrophile primarily to the *ortho* and *para* positions.

Y=Z: H–C=O, RO–C=O, C≡N, NO_2, etc.

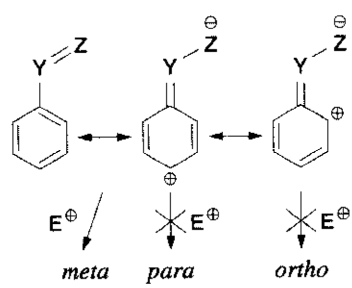


Figure 14-4: A group Y=Z with a multiple bond can reduce the electron density at the *ortho* (*o*) and *para* (*p*) positions by the so-called *minus mesomeric effect*. This makes the attack of the electrophile at these positions particularly difficult, leading to preferred *meta* substitution.

six input units, and the connection table representation requires 25 input units.

It turns out that one hidden layer is enough for both networks; in the first case (six input neurons), 10 neurons in the hidden layer are enough, while in the second case (25 input neurons), only five hidden neurons are necessary.

Two output neurons are chosen, one for the combined percentage of *ortho* and *para* product, the other for the percentage of *meta* product. Since the total yield is 100%, it is enough to report only the percentage of *meta* product.

With two neurons in the output layer, the complete architecture of the first network is (6 x 10 x 2), with (6 + 1) x 10 + (10 + 1) x 2 = 92 weights. The second network, (25 x 5 x 2), has (25 + 1) x 5 + (5 + 1) x 2 = 142 weights. The architectures of both networks are shown in Figure 14-7.

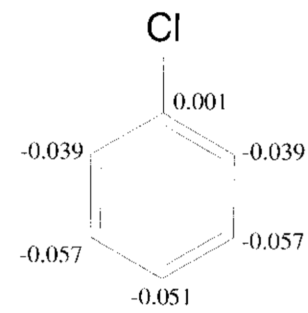


Figure 14-5: Partial charges on the six atoms of the benzenoid ring of *chlorobenzene* used as input representation.

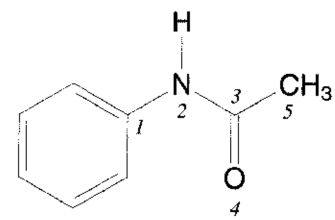
14.4 Learning and Results

The networks were trained by error back-propagation using the product ratios of the nitration of 37 monosubstituted benzene derivatives. The product ratios from 13 other compounds were used as a test set.

For both networks, Elrod et al. used 100,000 epochs (!) to reduce the errors in the training set to values as small as 0.3% in the better of the two networks. In spite of this excellent recall ability for the training set, the predictions had a much higher average error (12.1%); this is still not as good as we could desire.

In order to judge the quality of the results, the authors used two other approaches for estimating the amount of *meta* product: first, the results were compared with those obtained from CAMEO, an expert system for predicting the products of reactions (see Reference 14-4). Second, the 13 examples of the test set were given to three organic chemists, who were asked to predict the percentage of *meta* product. The predictions of these three chemists for all 13 compounds were averaged and compared with the results obtained by the two neural networks and by the CAMEO expert system. All this is summarized in Table 14-2.

Both neural networks gave better results than the expert system CAMEO. The average error of the three chemists is lower than that obtained from CAMEO and the neural network based on charge



atomic no.	bond atom 1	bond atom 2	bond order	charge
7	2	1	1	0
6	3	2	1	0
8	4	3	2	0
6	5	3	1	0
0	0	0	0	0

Figure 14-6: The substituent *acetanilide* and its connection table representation.

distribution. However, **the three chemists were outperformed by the network based on the connection table representation.**

system	training set	test set
neural network (6 x 10 x 2)	5.2	19.8
neural network (25 x 5 x 2)	0.3	12.1
CAMEO (expert system)	18.0	22.6
human experts		14.7

Table 14-2: Errors in recall and predictions for the amount of *meta* product by the two neural networks, by an expert system and by chemists (in percent).

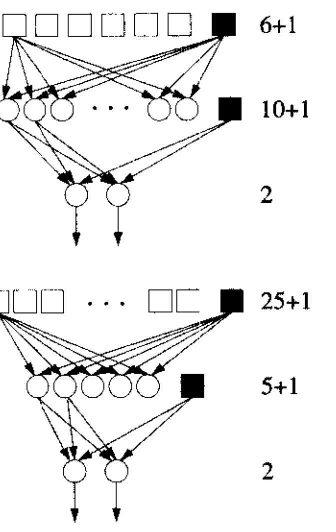


Figure 14-7: Two different neural network architectures for the two different structure-coding schemes.

Of the two neural networks, the one built on the connection table representation clearly shows the better results. This might be surprising, because this representation is much simpler to obtain than the one using partial atomic charges.

Why is the connection table network so good? A better question might be, why the other one is worse. In fact, it is not too surprising that the network based on the partial charges did not perform very well. The ground state charge distribution is only **one** of the various electronic factors influencing product distribution in electrophilic aromatic substitution, and, if considered **alone**, is clearly insufficient for representing the results of the nitrations.

Before continuing, it is prudent to admit that this study (like all studies) has limitations:

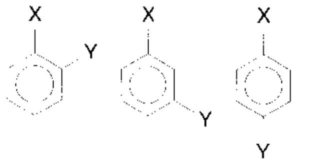
- The product ratio in the nitration of benzene derivatives depends strongly on reaction conditions, particularly on the concentration of sulfuric acid; this is not accounted for in the present study.
- The amounts of *ortho* and *para* product are combined, which prevents us from studying the important problem of the *ortho* effect. Furthermore, a separate treatment of the *ortho* and *para* distribution is a prerequisite for any attempt at predicting product distributions in di- and polysubstituted benzene derivatives.

This, of course, does not detract from the importance of the work by Elrod, Maggiora, and Trenary; but it is important to stress that the choice of input and output representations strongly determines the scope of an application and its prediction ability.

14.5 A Third Representation of Data

Satisfying as we might find the results obtained by the (25 x 5 x 2) neural network, they can not give an **explanation** of the effects influencing product ratios, since the connection table coding is arbitrary, and hides the chemical effects responsible for the product distribution.

This becomes particularly clear in the nitration of di- or polysubstituted benzene derivatives (Figure 14-8). Then, the above connection table representation is of no help at all in making generalizations. First, applying this representation to disubstituted benzene derivatives would require two (5 x 5)-connection tables; 50 input units would generate quite a different neural network architecture. Even worse, the two connection tables would be identical for the *ortho-*, *meta-*, or *para-*disubstituted benzene derivatives: they are insufficient to distinguish among these three different starting materials.



ortho *meta* *para*

Figure 14-8: *Ortho-*, *meta-*, and *para-*disubstituted benzene derivatives.

Clearly, we need a better coding scheme for di- or polysubstituted benzene derivatives. We will explore another representation that addresses the problem at the point where the reaction occurs, i.e., at the *ortho*, *meta*, or *para* position on the benzene ring.

The connection table representation is fine for predicting the yield ratio of *meta* to (*ortho* + *para*) product. However, if:

- the influence of sulfuric acid is to be accounted for
- the amounts of *ortho* and *para* product are to be distinguished
- the electronic effects governing product distribution in EAS reactions are to be deciphered, and
- the predictions on product ratios in the nitration of di- and polysubstituted benzene derivatives have to be made,

we need more general representations on the input side for the starting materials and reaction conditions, and on the output side for the product distribution.

For example, the input vector should be coded using a **reaction site-specific** representation; this means that the representation for a given substituent should be **different** for each output ring position, so that different variables may be input for different positions (while there is always only one answer: percent of the product corresponding to substitution at the position for which the variables are input).

In principle, there are five hydrogens on a monosubstituted benzene derivative that can be substituted by an electrophilic agent (e.g., nitro). However, because the molecule has a plane of symmetry, there are two equivalent *ortho* and two equivalent *meta* positions, so that only three different positions, *ortho*, *meta* and *para* have to be considered, but because of the symmetry, *ortho* and *meta* must be weighted twice as much as *para*.

Furthermore, we need one (or more) additional input units to account for one (or more) variables on reaction conditions; in this case the concentration of sulfuric acid should be included.

The problem of individually representing the three substitution positions of a monosubstituted benzene derivative was addressed by A. Fröhlich and coworkers of the Model Laboratory for Computer Chemistry at the Technical University of Munich (see Reference 14-5). They used **one** steric and **four** electronic variables.

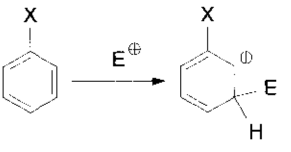


Figure 14-9: Intermediate state leading to the *meta* product in an electrophilic aromatic substitution ($E^{\oplus}$ is the electrophile).

Figure 14-9 shows the intermediate formed in an electrophilic aromatic substitution; we can use electronic variables for the carbon atom where the electrophile E is bonded as controlling parameters: the σ-electronegativity, χ_σ; the π-electronegativity, χ_π; and both the average inductive stabilization, $\chi_\sigma^{av}(o,p)$, and the resonance stabilization, R^+, of the positive charge generated at this carbon atom (Figure 14-9). These parameters can be calculated by empirical methods.

Two additional input units are provided: one for an estimate of the amount of steric hindrance, *Ster*, at the reaction position obtained from the van der Waals radii of the atoms, and one for the concentration of sulfuric acid, $[H_2SO_4]$ (Figure 14-10).

The values of the input parameters for each of the three sites of phenol are given in Table 14-3.

With a site-specific representation of the starting material, we need only one output neuron, the amount of product at the site being considered (*meta* position, in Figure 14-10). Seven hidden neurons complete the architecture of the neural network in this study (Figure 14-10).

Note that the neural network is trained with data for each individual position separately. This means that if we input the six parameters χ_σ, χ_π, $\chi_\sigma^{av}(o,p)$, R^+, *Ster*, and $[H_2SO_4]$ for the ***para***

site	χ_σ [eV]	χ_π [eV]	$\chi_\sigma^{av}(o,p)$ [eV]	R^+ [1/eV]	*Ster* [Å^3]	[H_2SO_4] [%]	yield [%]
ortho	8.45	5.41	8.66	21.1	5.85	74.1	18.0
meta	8.25	5.34	8.38	6.0	5.40	74.1	0.5
para	8.23	5.34	8.66	21.1	5.04	74.1	63.0

Table 14-3: Site-specific input and output parameters for the three different sites of phenol. (Note that the total yield is 18 x 2 + 0.5 x 2 + 63 = 100)

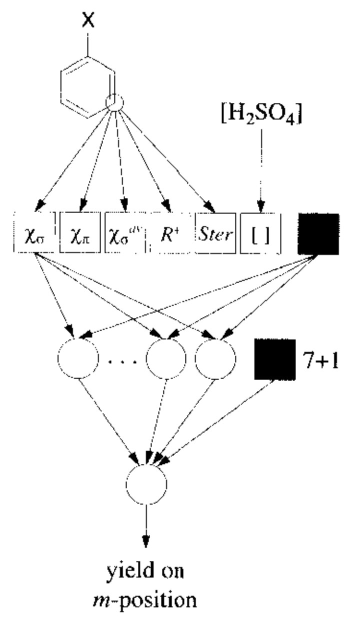


Figure 14-10: Architecture for prediction of the yield of product at a specific site (*meta*, in this case).

position, then as a target for training the network, the percentage of ***para*** product should be given. However, if data for the ***ortho*** or ***meta*** positions are used, the network outputs only half the expected yield, since in actuality, each ring has two such positions ("statistical factor", or "symmetry factor").

Thus, one single neural network is trained to predict the yield of substitution products at **each individual** position.

This network was trained by back-propagation of errors, using as a training set the product distributions in the nitration of 23 monosubstituted benzene derivatives at various concentrations of sulfuric acid. Altogether, these 23 compounds provided 159 data on the yields of substitution products at different positions, for different concentrations of sulfuric acid. The data are easily learned, with an average error of 6% on recall.

Three disubstituted benzene derivatives were then used for testing the predictive performance of the neural network. Since each disubstituted benzene derivative contains four sites for potential substitution, predictions for each of the twelve sites for nitration were individually made by inputting the electronic and steric variables of each site together with a preset concentration of sulfuric acid (Figure 14-11). The average error in the prediction of the yield for substitution at the various positions amounts to 10%.

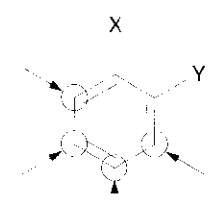


Figure 14-11: The four sites for potential substitution in an *ortho*-disubstituted benzene derivative.

Figure 14-12 summarizes the three different approaches for setting up a multilayer neural network capable of predicting the regio-selectivity in electrophilic aromatic substitution, and for coding a monosubstitued benzene derivative as input.

Figure 14-12a) shows the representation used by Elrod, Maggiora and Trenary, in which the charges at the six positions of a monosubstituted benzene derivative are input in order to predict the yields of *meta* and (*ortho* + *para*) products by using **two** output neurons. In Figure 14-12b), the structure of the **substituent** is coded by a (5 x 5) connection table; again, the *meta* and (*ortho* + *para*) product yields are obtained on **two** outputs.

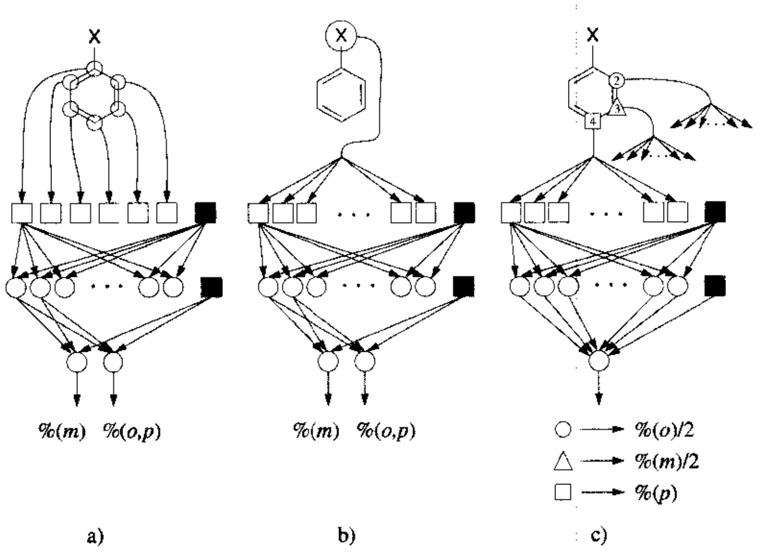


Figure 14-12: Three different architectures and input representations for learning the regioselectivity of the electrophilic aromatic substitution reaction: a) global charge vector, b) connection table of the substituent, c) local electronic, steric, and reaction condition representation.

Figure 14-12c) diagrams the site-specific neural network that takes as input one steric and four electronic variables for one ring site, as well as the concentration of sulfuric acid. **One** output neuron suffices to predict the amount of product at that site.

14.6 Concluding Remarks

Go back and read the quotation of Bernhard Widrow at the very end of Chapter 9. The important message to carry away from this study is that your method of representing information determines the scope of the predictions that can be made. A **global** representation of the substituent in a benzene derivative by a connection table only allows us to make global predictions, e.g., the amount of *meta* product vs. the sum of *ortho* and *para* product. A *local* representation of the influence of a substituent on each individual ring position allows us to make **local** predictions: the amount of product at each individual position. Furthermore, only a local representation of a monosubstituted benzene derivative can be generalized to make predictions about product ratios in the nitration of di- and polysubstituted derivatives.

14.7 References and Suggested Readings

14-1. D. W. Elrod, G. M. Maggiora and R. G.Trenary, "Applications of Neural Networks in Chemistry. ***1***. Prediction of Electrophilic Aromatic Substitution Reactions", *J. Chem. Inf. Comput. Sci.* **30** (1990) 477 – 484.

14-2. D. W. Elrod, G. M. Maggiora and R. G. Trenary, "Applications of Neural Networks in Chemistry. ***2***. A General Connectivity Representation for the Prediction of Regiochemistry", *Tetrahedron Comput. Methodol.* **3** (1990) 163 – 174.

14-3. MOPAC program, available from Quantum Chemistry Program Exchange (QCPE 455, Version 6.0).

14-4. M. G. Bures, B. L. Roos-Kozel and W. L. Jorgensen, "Computer-Assisted Mechanistic Evolution of Organic Reactions (CAMEO). ***10***. Electrophilic Aromatic Substitution", *J. Org. Chem.* **50** (1985) 4490 – 4498.