Lecture Transcripts

Principal Properties and Designs for Discrete Variations[1]

Rolf Carlson*
Department of Chemistry, Faculty of Science, University of Tromsoe, NO-9037 Tromsoe, Norway

Johan E. Carlson
*EISLAB, Department of Computer Science and Electrical Engineering,
Luleå University of Technology, SE-971 87 Luleå, Sweden*

Abstract:

A problem that is often encountered when a new synthetic reaction is developed is to determine suitable combinations of reagents, co-reagents, catalysts, solvents, etc. This contribution presents general strategies for designing experiments when the objective is to explore the discrete variations defined by different reagents, different catalysts, different solvents, etc. The concept of principal properties is introduced, and it is shown how the principal properties of the constituents of the reaction system can be used for the selection of suitable test systems. Chemical examples are provided by the following: the selection of test solvents in the reduction of an enamine; the selection of combinations of Lewis acids and amines in the synthesis of benzamides; the selection of ketone substrates, amines, and solvents in the Willgerodt−Kindler reaction; and the selection of ketone substrates, Lewis acid catalysts, and solvents for analysing the regioselectivity in the Fischer indole synthesis with dissymmetric ketones.

Introduction

Organic process chemistry is a challenging field of synthetic chemistry. It is an experimental science, and its objective is to furnish the chemist with simple and convenient methods for the construction of the desired target molecules from simple and easily available starting materials. All knowledge in synthetic chemistry is based on inferences from experimental observations. To develop new methods or to improve existing methods, it is therefore of tremendous importance that the experiments carried out have been designed in a proper manner. Hence, the concept of experimental design is vital to the organic process chemistry. This paper highlights some aspects of the design of experiments in synthetic chemistry when the objective is to determine which combination of substrate, reagent(s), co-reagent(s), catalyst, solvent, etc. will afford the most promising result and thus merit further exploration with respect to the adjustment of the detailed settings of the experimental conditions.

Problem Description

The Reaction Space. The essential feature of the problem can be described using the *reaction space* depicted in Figure 1. The *axes* of the reaction space define *variations* in the nature of the different constituents of the reaction system (*substrate, reagent, co-reagent, catalyst, solvent, etc.*) The entire reaction space is defined by the union of all possible combinations of these constituents.

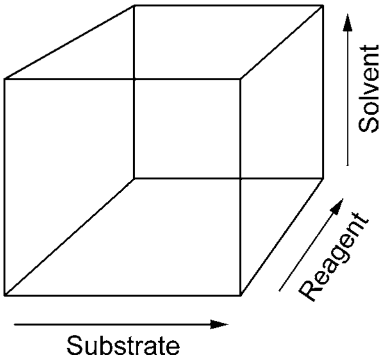


Figure 1. **The reaction space.**

An experimental design for exploring the reaction space defines a selection of test systems that covers the space as efficiently as possible. How such selections should be made is, of course, dependent on *which* questions we pose to our experimental system:

(1) Can the reaction be used for the conversion of all substrates containing the necessary functional group?

(2) Is the reaction sensitive to the nature of the solvent?

(3) Is there a combination of catalyst and solvent that can afford a selective transformation?

(4) *Which* properties of substrates, reagent(s), catalyst(s), solvent(s), etc. are critical?

(5) *Why* are they important?

(6) *How* do the properties of the reactants exert their influence?

(7) Which combination of reagent, catalyst, and solvent gives the best result?

To permit a systematic search of the reaction space, we must quantify the “axes”. This is accomplished by the *principal properties* of the classes of the constituents in the reaction space. A thorough discussion of the concept of principal properties and how principal properties can be used for exploring discrete variation is given in ref 2. A brief outline is given below.

* To whom correspondence should be addressed. E-mail: rolf.carlson@chem.uit.no.

[1] Presented at the 2nd International Symposium: Optimising Organic Reactions and Processes, Bergen, Norway, May 18−20, 2004.

[2] Carlson, R.; Carlson, J. E. *Design and Optimization in Organic Synthesis, Second Revised and Enlarged Edition*; Elsevier: Amsterdam, 2005.

Published on Web 08/13/2005
10.1021/op040022w CCC: $30.25

Principal Properties and the Reaction Space

When a molecule takes part in a chemical reaction or when it interacts with its surroundings, the properties at the molecular level determine the outcome. However, such intrinsic properties cannot be measured directly. What can be obtained are measures of macroscopic properties that can be linked to intrinsic properties by some physical−chemical model. For example, ^{13}C NMR shifts of different carbons in a molecule can be used to estimate the electron densities at these carbons; UV transitions can be used to estimate energy differences between the frontier orbitals. Some intrinsic properties can, of course, be estimated through quantum chemical modelling. Any chemical compound can be characterised by a very large number of measurable or computable property variables, *observables*. This means that we can collect observables for each "dimension" of the reaction space. The worst thing to do in this situation is to use commercially available software to generate a very large number of characterising variables and then use, for example, genetic algorithms or simulated annealing to select a few of them for modelling purposes with some criteria of model fit to justify the selection. We fully agree with the statement by Wold[3] that we should avoid large data sets and megavariate problems by designing our own data sets. What should be done is to analyse the chemical problem and then select *descriptors* that we know, believe, or suspect can pick up those molecular properties that will exert an influence on the reaction. Which descriptor to choose is therefore dependent on the chemical problem at hand, and there is not a universally valid set of descriptors that can always be used. If, for instance, we believe that the nucleophilic properties are critical, we should select descriptors related to this, for example, *refractive index, frontier orbital energy, ionisation potential, donor number,* pK_B, *proton affinity, ..., etc.*

In short, molecules can be characterised by their descriptors, and the descriptors can likely be assumed to be observable manifestations of intrinsic molecular properties. In several respects the members of the classes of constituents defining the "axes" of the reaction space can be assumed to be similar. For example, if a reaction is to be developed with ketones as substrates, the class of ketones defines the set of possible substrates. They have the same functional group and are in this respect similar. However, variations of the side chains make the various ketones slightly different with respect to their properties. Assuming that the descriptors can be used as probes of the intrinsic molecular properties, we can analyse the intrinsic properties by analysing the descriptors. Observable properties that depend on the *same* intrinsic property, can be assumed to be more or less correlated to each other. We can also assume that observable properties that depend on *different* intrinsic properties will not be strongly correlated. An analysis that takes these assumptions into account can be carried out by principal component analysis of the observed variation of the descriptors over a set of compounds. Assume that each compound has been characterised by a set of k descriptors $[q_1, q_2, ..., q_k]$. The set of n compounds will then define an $n \times k$ descriptor matrix $\mathbf{Q}$. For convenience, assume also that $\mathbf{Q}$ then has been transformed to $\mathbf{X}$ by mean centring each variables and scaling each variable to unit variance. Assume also that the descriptors have been transformed in a chemically meaningful way prior to centring and scaling. For example, UV transitions are better expressed as the reciprocal of the wavelength since this entity is proportional to energy. The principal component model can be written as

$$\mathbf{X} = \mathbf{T}\,\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_A]$ is the score matrix defined by the eigenvectors $\mathbf{t}_i$ ($in = 1,..., A$) of the correlation matrix $\mathbf{XX}^{\mathrm{T}}$; the matrix $\mathbf{P}^{\mathrm{T}}$ is the transpose of the loading matrix defined by the eigenvectors $\mathbf{p}_i$ ($i = 1,..., A$) of the variance−covariance matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$. A matrix of residuals always occurs when the number of components, A, included in the model is less than the number of original descriptor variables.

The number if significant components, A, can be determined by cross-validation.[4] The principal component analysis constitutes a projection of the n object points in the k-dimensional descriptor space onto an A-dimensional hyperplane spanned by the eigenvectors, $\mathbf{p}_i$. The elements t_{ij} of the score vectors, $\mathbf{t}$, are orthogonal projections of the object j on the eigenvector $\mathbf{p}_i$. The principal component model describes the systematic variation of the properties over the set of compounds. The nonsystematic variation is collected in the residual matrix $\mathbf{E}$. The benefit of this procedure is that the number of variables we need to consider for taking the *systematic variation* into account is less than the number of original descriptors. We use the term *principal properties* to define the correlative pattern of the original descriptors as portrayed by the eigenvectors of property matrices. As the eigenvectors are orthogonal, we can assume that they portray different intrinsic properties. We can then use the eigenvectors to define the *axes* of the reaction space and the score values, t_{ij}, to quantify the variation along these axes. A plot of the score vectors against each other, e.g. $\mathbf{t}_1$ vs $\mathbf{t}_2$, will display the scatter of the object points projected onto the first two eigenvectors. Objects that are close to each other in the k-dimensional descriptor space will be projected close to each other in the score plot. Conversely, objects that are dissimilar to each other and therefore located far from each other in the descriptor space will be projected far from each other in the score plot. When there are only a few dimensions of the reaction space to consider, a selection of test candidates for experiments can be made by a visual inspection of the corresponding score plots.

Exploring the Reaction Space

Few Dimensions To Explore. The first example of this technique as a tool for selecting test items in organic synthetic chemistry was presented in a paper on the selection of solvents in organic synthesis.[5] Analyses of solvent properties by principal component analysis and factor analysis had previously been described,[6] but the use of such methods in conjunction with screening designs in organic synthesis was

[3] Wold, S.; Berglund, A.; Kettaneh, N. *J. Chemom.* **2002**, *16*, 377.

[4] Wold, S. *Technometrics* **1978**, *20*, 397.

[5] Carlson, R.; Lundstedt, T.; Albano, C. *Acta Chem. Scand.* **1985**, *B 39*, 79.

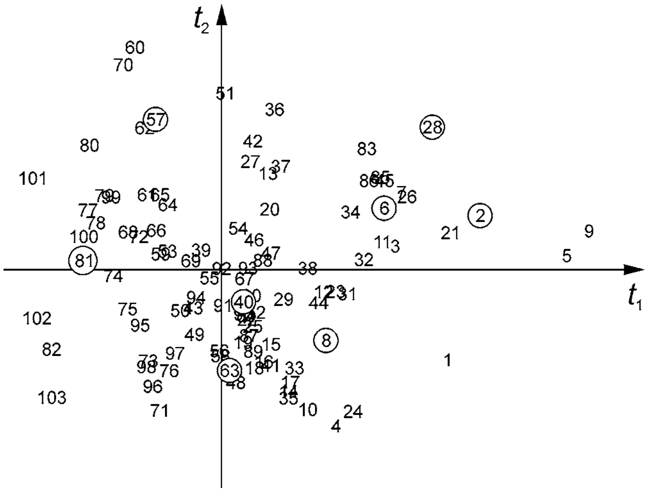


Figure 2. **Selected solvents in the enamine reduction study.**

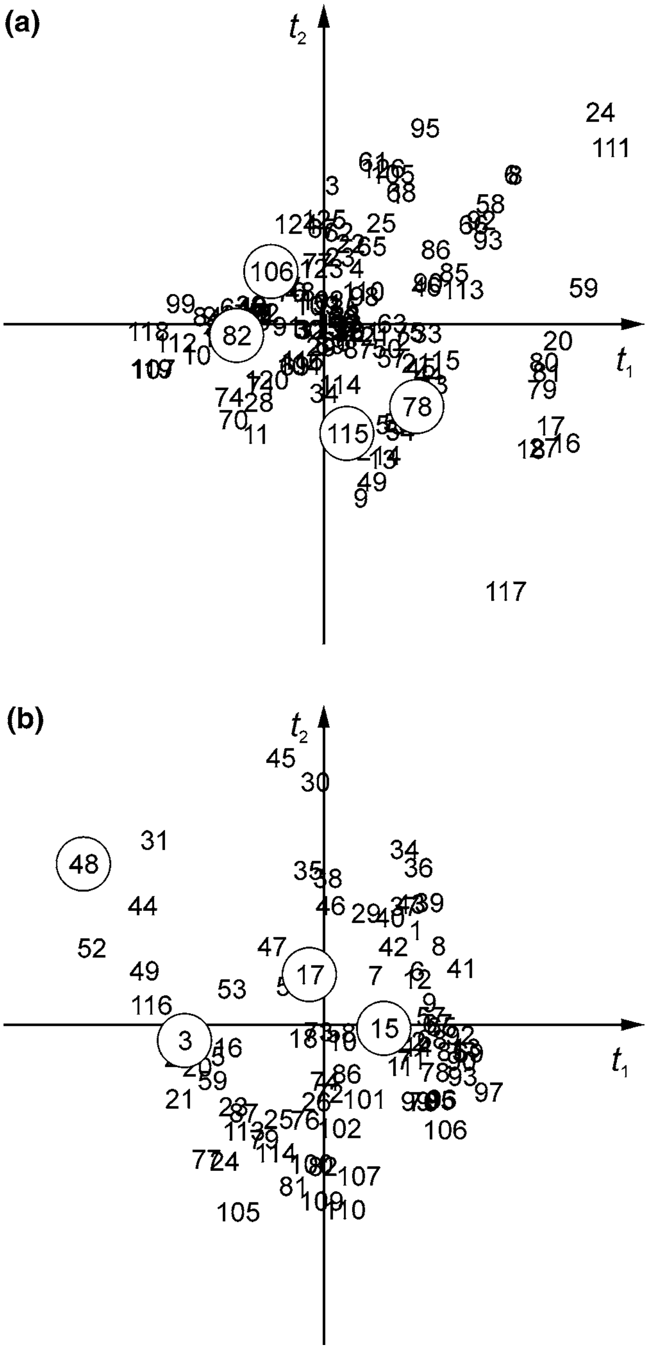


Figure 3. **(Top) Selected amines: benzylamine (*78*), butylamine (*82*), morpholine (*115*), dipropylamine (*106*). (Bottom) Selected Lewis acids: BF$_3$ (*44*), TiCl$_4$ (*3*), AlCl$_3$ (*17*), ZnI$_2$ (*99*), ZnCl$_2$ (*15*).**

not known when ref 5 was published. Different design strategies were discussed. If the problem is to determine wether the solvent properties influence the result, a selection of solvents that are projected at the periphery of the score plot should be selected. This ensures a maximum spread in the properties of the selected solvents. Today, in the era of combinatorial chemistry, such a design would be called a *diversity design.* If the problem is to determine whether there is a gradual change in the performance of the reaction that can be attributed to the properties of the solvent, a selection of test candidates that are uniformly distributed in the score plot should be made. In combinatorial language, this corresponds to a *grid search*. It was also discussed how a D-optimal design can be used for the selection of test candidates. In retrospect, this paper was indeed prophetic.

Example: Reduction of an Enamine. It was found that D-camphor could be converted to the corresponding enamines in high yields by adding camphor to a TiCl$_4$−amine complex.[7] These enamines were almost quantitatively converted to the saturated bornylamines when they were treated with 98% formic acid. The reaction was highly stereoselective and the *endo* isomer was the dominating product. The following *endo/exo* ratios were observed: *N*-bornylmorpholine (*93/7*), *N*-bornylpiperidine (*92/8*), *N*-bornylpyrrolidine (*85/15*). The reactions were carried out by adding a stoichiometric amount of formic acid to the neat enamine at 100 °C. The question was whether the stereoselectivity could be improved by running the reaction in solution, and in that case, which solvent should be used? The pyrrolidine enamine was used as model substrate since this compound yielded the poorest stereoselectivity using the neat enamine. The following test solvents were selected from the principal property score plot so that they afforded a maximum spread (diversity) design. The following solvents were chosen (the numbers refer to the score plot in Figure 2): formamide (*2*), sulfolane (*28*), 1,2-dichlorobenzene (*57*), cyclohexane (*81*), tetrahydrofuran (*63*), methoxyethanol (*8*), methanol (*4*), 2-methyl-2-butanol (*40*), and diethylene glycol (*6*).

When the reaction was run in these solvents (one equivalent of formic acid, 100 °C or reflux), the amounts of the *endo* isomer were in the range 85−87%, and no improvement of the selectivity was observed. Since the selected solvents covered a very large variation of the solvent properties, it was concluded that the probability of finding a solvent in which the reaction is stereoselective must be very low. Instead, a method for isolating the *endo* isomer by recrystallisation of the hydro-

[6] (a) Bohle, M.; Kollecker, W.; Martin, D. *Z. Chem.* **1977**, *17*, 161. (b) Chastrette, M. *Tetrahedron* **1979**, *35*, 1441. (c) Cramer, R., III. *J. Am. Chem. Soc.* **1980**, *102*, 1837, 1849. (d) Elguero, J.; Fruchier, A. *Anal. Quim. Ser. C.* **1983**, *79*, 72. (e) Svoboda, P.; Pytela, O.; Vecera, M. *Collect. Czech. Chem. Commun.* **1983**, *48*, 3287.

[7] Carlson, R.; Nilsson, Å. *Acta Chem. Scand.* **1985**, *B 39*, 181.

[8] Nordahl, Å.; Carlson, R. *Acta Chem. Scand.* **1988**, *B 42*, 28.

Table 1. **Yields obtained with different combinations of amines and Lewis acids**

amine	Lewis Acid	yield of carboxamide/%	amine	Lewis Acid	yield of carboxamide/%
benzylamine	BF_3-etherate	86	morpholine	BF_3-etherate	91
	$TiCl_4$	15		$TiCl_4$	81
	$AlCl_3$	63		$AlCl_3$	72
	ZnI_2	0		ZnI_2	0
	$ZnCl_2$	0		$ZnCl_2$	0
butylamine	BF_3-etherate	81	dipropylamine	BF_3-etherate	45
	$TiCl_4$	19		$TiCl_4$	44
	$AlCl_3$	68		$AlCl_3$	8
	ZnI_2	trace		ZnI_2	0
	$ZnCl_2$	0		$ZnCl_2$	0

Table 2. **Preparative scale synthesis of benzamides**

amine	Lewis acid	reaction time/h	yield/%
benzylamine	BF_3-etherate	96	89
butylamine	BF_3-etherate	96	83
morpholine	$TiCl_4$	10	86
dipropylamine	$TiCl_4$	10	86

chloride salt was developed. The solvent for this was also determined from the principal properties, but this is another story.

Example: Lewis Acid-Mediated Synthesis of Benzamides. This example is taken from ref 8. Carboxamides can be prepared by a large number of methods. Most often, the carboxylic acid is converted to a more reactive intermediate, e.g. the acid chloride which then is allowed to react with an amine. For practical reasons it is preferable to form the reactive intermediate in situ. A carboxamide-forming reaction of this type attracted our interest in the context of our studies on Lewis acid-catalyzed reactions, viz. the reaction between carboxylic acids and primary or secondary amines catalysed by Lewis acids and tertiary amines. Scattered examples of this reaction had been found in the literature, but no systematic study had been undertaken.

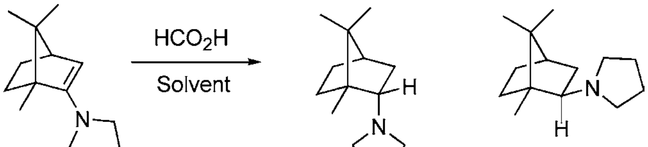
Benzoic acid was used as model substrate. Triethylamine was used as the tertiary amine. Four amines and five Lewis acid to be tested were selected from the score plots to ensure a sufficient spread in their properties, see Figure 3. Significant experimental variables were identified from a two-level fractional factorial design. The significant variables were then further investigated by an augmented two-level design to also determine significant interaction effects. We will not go into details on the experimental variables. We will focus on the results obtained in the study of the reaction space.

Table 1 summarises the yields obtained with the selected reaction systems.

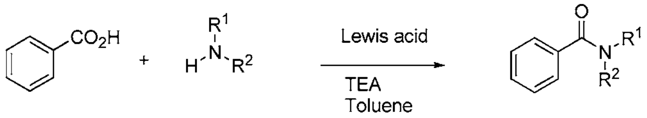
The results in Table 1 show that boron trifluoride afforded good yields, and also the carboxamide with all four amines. Aluminum trichloride afforded lower yields than boron trifluoride. Titanium tetrachloride afforded high yield only with the secondary amines. The zinc halides were useless. There was a striking difference in the reaction times required for obtaining the maximum yield in preparative scale runs as shown in Table 2.

This study shows that a selection of reagents by their principal properties made it possible to identify suitable combinations of reactants and reagents for preparative-scale synthesis of benzamides.

Scheme 1



Scheme 2



Several Dimensions of the Reaction Space To Explore. In the general case, the complexity of the problem increases, however, when the number of dimensions of the reaction space increases. One possibility to cope with the problem is to use a two-level fractional factorial design to explore the reaction space. However, this yields a selection of test items that are sparsely distributed in the reactions space, and such designs are only useful for screening purposes. A fractional factorial design for exploring the reaction space is shown below with an example of the Willgerodt−Kindler reaction.

For fine-tuning the reaction system it will, however, be necessary to select test items in such a way that it is possible to evaluate if there are gradual changes in the outcome of the reaction and if these changes can be traced back to the properties of the constituents of the reaction system. For this, the test items for each dimension of the reaction space should be selected in such a way that the test points form a uniformly distributed grid of points in the corresponding score plots. A brute force example with the Fischer indole syntheis is shown below, and it is then shown that the number of selected test items can be considerably reduced by using a statistical design based upon singular value decompositions of candidate model matrices.

Fractional Factorial Design for Screening the Reaction Space. These principles were presented in the context of the Willgerodt−Kindler reaction,[9] see Scheme 3. The question posed to the reaction system was whether it would have been possible to assess the scope and limitation of the Willgerodt−Kindler reaction from a small number of experiments.

[9] Carlson, R.; Lundstedt, T.; Shabana, R. *Acta Chem. Scand.* **1986**, *B 40*, 694.

Scheme 3

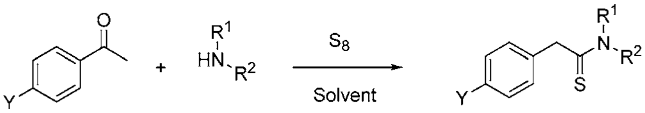


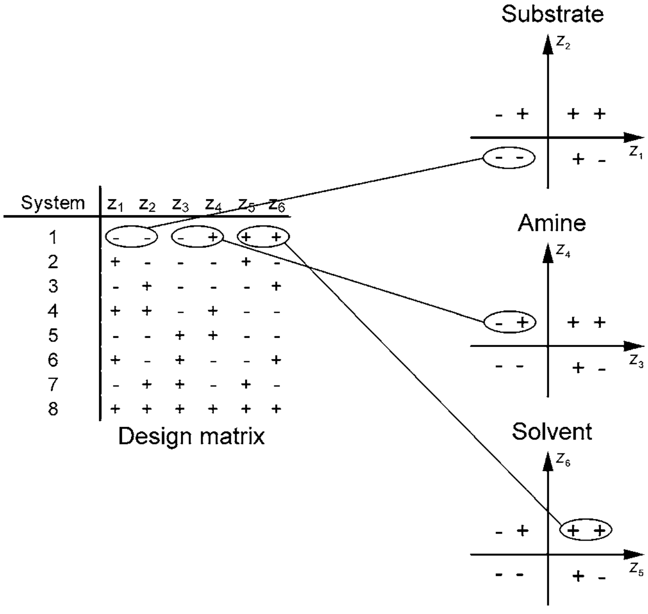
Table 3. **Selected items from the score plots in the Willgerodt−Kindler study**

	assignment				
items	z_i	z_j	substituent Y	amine	solvent
1	−	−	Cl−	isopropylamine	benzene
2	+	−	H−	morpholine	ethanol
3	−	+	MeO−	diethylamine	quinoline
4	+	+	PhO−	dipentylamine	triethylene glycol (TEG)

The perturbations to consider were: (a) the nature of the substrate as described by the properties of the substituents, *Y*, (b) the nature of the amine co-reagent, (c) the nature of the solvent. For each of these, two principal components were sufficient to describe the systematic variations. Thus, the reaction space was six-dimensional, each dimension was described by two score vectors. To span the variation, a fractional factorial design, 2^{6-3}, was used, letting the variables two by two define selections from different quadrants of the score plots, see Figure 4.

The selections from the score plots are summarised in Table 3 and the final design is shown in Table 4. The optimum experimental conditions for each selected system were determined by response surface methodology. This was actually the very first statistically designed combinatorial library in synthetic organic chemistry.

For each selected system, the experimental conditions giving the maximum yield were determined by response surface techniques. From these results, it was then shown that a PLS model could be used to predict the optimum conditions for new reaction systems.



System	z_1	z_2	z_3	z_4	z_5	z_6
1	-	-	-	+	+	+
2	+	-	-	-	+	-
3	-	+	-	-	-	+
4	+	+	-	+	-	-
5	-	-	+	+	-	-
6	+	-	+	-	-	+
7	-	+	+	-	+	-
8	+	+	+	+	+	+

Design matrix

Figure 4. **Selection of test items by a fractional factorial design.**

Table 4. **Selected reaction systems in the Willgerodt−Kindler study**

	reaction system substituent	amine	solvent
1	Cl−	Et_2NH	TEG
2	H−	*i*-$PrNH_2$	quinoline
3	MeO−	*i*-$PrNH_2$	EtOH
4	PhO−	Et_2NH	benzene
5	Cl−	Pe_2NH	benzene
6	H−	morpholine	EtOH
7	MeO−	morpholine	quinoline
8	PhO−	Pe_2NH	TEG

The conclusion from this study was that a small number of suitably selected test items can reveal the general scope of synthetic reactions. These principles have later successfully been used in medicinal chemistry for designing combinatorial libraries in the context of QSAR modelling.[10]

Brute Force and Singular Value Decomposition

To understand these principles we shall look at the underlying principles and discuss modelling of the reaction space.

Modelling. The outcome of a chemical reaction is determined by the energy changes during the reaction. The course of the reaction can be described as a movement on a potential energy surface, from one energy minimum representing the starting materials to another energy minimum representing the products. The energy difference between the minima determines the position of chemical equilibria. The energy barrier to surmount going from one minimum to another, the activation energy, determines the rate of the reaction. How deep the minima are, how high the energy barriers are will depend on the detailed experimental conditions. This holds both for variations in the reaction space and for variations in the experimental space. For example, modelling how changes of the reaction systems influence rates and equilibria, i.e. how the trajectories over the potential energy surface differ in shape, forms the basis of all linear free energy relationships.[10]

Two responses are of general interest in organic synthesis, the yield and the selectivity.

The yield is actually the integral of the reaction rate over time, i.e.

$$\text{yield} = \int \text{rate } \mathrm{d}t$$

The selectivity can be described as the ratio of the rates of the reactions forming the different products, *A* and *B*, as

$$\text{selectivity} = \frac{\text{rate } (A)}{\text{rate } (B)}$$

If the reaction is rapidly reversible, the selectivity is determined by the equilibrium constant, K_{eg}.

$$K_{\text{eg}} = [A]/[B]$$

[10] See, for instance: Sjöström, M.; Eriksson, L. Application of Statistical Experimental Design and PLS Modeling in QSAR. In *Chemometric Methods in Molecular Design*; vand der Waterbeemb, H.; Ed.; VCH: Weinheim, 1995.

For exploring the experimental conditions it is reasonable to assume that the outcome of the reaction, y, is dependent on the experimental conditions and that we can assume that there is some kind of functional relation, f, between them, i.e.

$$y = f(\text{experimental conditions})$$

The experimental conditions are defined by the settings of the experimental variables, $x_1, x_2, ..., x_k$, and we can assume the following functional relation:

$$y = f(x_1, x_2, ..., x_k)$$

For exploring the reaction space, it is reasonable to assume that the outcome, y, of a reaction is dependent on the properties of the reaction system and that we also in this case can assume some functional relation between the properties of the system and the outcome, that is

$$y = f(\text{properties of the reaction system})$$

Assume that the reaction space has been defined by the principal properties of the constituents. We can therefore assume a functional relation between y and the principal properties:

$$y = f(\text{principal properties of the reaction system})$$

If x_i denotes the score value of a principal property, we can assume the following functional relationship:

$$y = f(x_1, x_2, ..., x_k)$$

In general, we do not know an analytical expression of f, and it will be difficult to derive it from chemical theory. The function f is determined by the shape of the potential energy surface and how this shape is altered by experimental perturbations. Provided that the experimental perturbations are not too large so that a totally different reaction mechanism begins to operate, we can assume a smooth change of the shape of the potential energy surface as a result of the experimental variations. We can therefore assume that f is smooth and several times differentiable. Under these conditions, we can obtain an approximation of f by a Taylor expansion.

Let **0** be the centre point of the domain to be explored, $x_1 = x_2 = ... = x_k = 0$. A Taylor expansion around the centre point will be:

$$\begin{aligned} y = f(\mathbf{0}) &+ \partial f(\mathbf{0})/\partial x_1\, x_1 + \partial f(\mathbf{0})/\partial x_2\, x_2 + ... + \partial f(\mathbf{0})/\partial x_k\, x_k + \\ &+ {}^1/_2 \partial^2 f(\mathbf{0})/\partial x_1 \partial x_2\, x_1\, x_2 + ... + {}^1/_2 \partial^2 f(\mathbf{0})/\partial x_i \partial x_j\, x_i\, x_j + ... \\ &+ {}^1/_2 \partial^2 f(\mathbf{0})/\partial x_1^2\, x_1^2 + ... + {}^1/_2 \partial^2 f(\mathbf{0})/\partial x_k^2\, x_k^2 + \\ &\qquad \text{higher-order terms} \end{aligned}$$

In most cases, a sufficiently good approximation is obtained if the Taylor polynomial is truncated after the inclusion of the second order terms. The truncated Taylor expansion is more conveniently written as,

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \beta_{12} x_1 x_2 + ... + \beta_{ij} x_i x_j + \\ \beta_{11} x_1^2 + ... + \beta_{kk} x_k^2 + e \end{aligned}$$

The error term, e, contains contributions from the omitted terms of the Taylor expansion.

Such models are often called response surface models or response surfaces since they describe a surface in the space spanned by $\{y, x_1, x_2, ..., x_k\}$. The coefficients of the response surface model describe how the settings of the experimental variables are linked to the response. We can therefore analyse the roles played by the variables from estimates of their coefficients. Such estimates can be obtained from properly designed experiments. The model is linear in the coefficients and least-squares estimates of the coefficients, $b_0, b_1, ..., b_k, b_{12}, ..., b_{ij}, b_{11}, ..., b_{kk}$, can be obtained by fitting the polynomial to known experimental results by multiple linear regression.

Experimental Design

From the above discussion it is seen that a fairly detailed description of the roles played by the experimental variables can be obtained from the response surface model. Before designing the experiment we must therefore decide: *How detailed is the information required?*

* *Linear models:* A model with only linear terms will seldom give a very precise description, but such models are very useful in screening experiments with many variables with a view to determining which variables are important.

* *Interaction models:* If we include the cross-product terms in the model, it is possible to assess interaction effects. It is always advisable to consider possible interaction effects.

* *Quadratic models:* Such models can describe nonlinear dependencies between the response and the variables. Close to an optimum, for instance, the maximum yield, the response surface is curved. To describe the response surface in the near optimum region it will be necessary to take the curvature of the surface into account. This is accomplished by the quadratic terms.

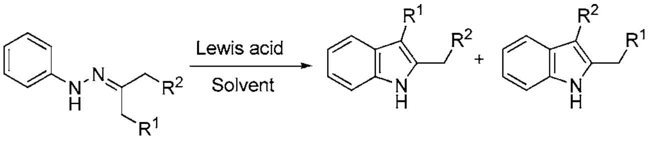
Fischer Indole Synthesis

This example is included to show how the reaction space can be investigated by experimental designs in the principal properties to determine which properties are critical. A general strategy for the construction of optimal designs is outlined. The example is taken from results published in ref 12.

When phenylhydrazones from aldehydes and ketones containing α-methylene groups are heated in the presence of acid, they undergo a rearrangement and a ring closure to form indoles. The reaction was discovered by Emil Fischer in 1883[13] and has ever since been a workhorse in the field of heterocyclic chemistry for the synthesis of indoles. An extensive monograph over the reaction has been published.[14] A problem with the Fischer indole synthesis is that regioisomeric indoles are formed in the reaction of phenylhydrazones from dissymmetric ketones containing α- and α′-methylene groups.

The Problem, the First Brute Design and Conclusion from It. Which properties of the Fischer indolisation reaction systems are critical for obtaining a regioselective reaction in the reaction of phenylhydrazones from dissymmetric bis-

Scheme 4



methylene ketones, see Scheme 4. Are there certain combinations of Lewis acid catalysts and solvents that can give regioselective reactions? With a view to answering these questions, 5 ketones, 12 Lewis acids, and 10 solvents were selected from their principle properties. The score plots are shown in Figure 5. The selected items are summarized in Table 5.

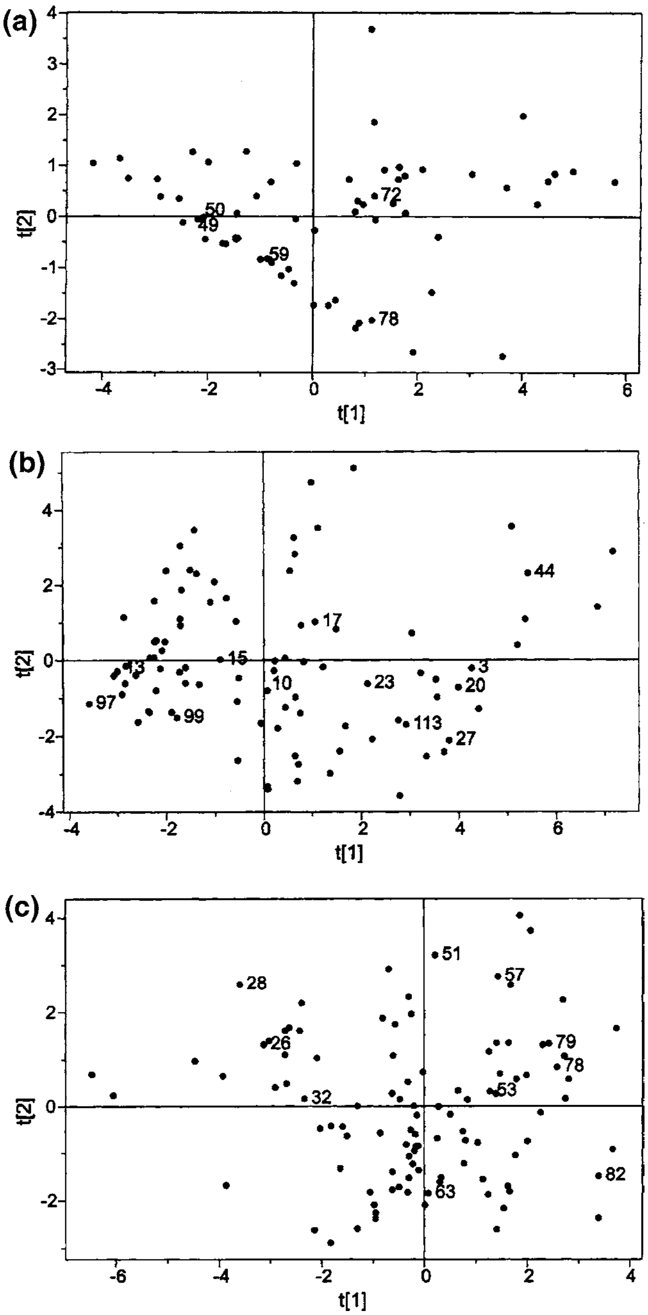


Figure 5. **Score plots used for selecting test items in the Fischer indole synthesis study: ketones (top), Lewis acids (middle), solvents (bottom).**

Table 5. **Selected items in the Fischer indole study**

ketones	Lewis acids	solvents
3-hexanone	BF_3	sulfolane
2-hexanone	CuCl	carbon disulfide
3-undecanone	ZnI_2	*N*,*N*-dimethylacetamide
1-phenyl-2-butanone	$TiCl_4$	quinoline
5-methyl-3-heptanone	$ZnCl_2$	1,2-dichlorobenzene
	$SbCl_5$	dimethyl sulfoxide
	PCl_3	carbon tetrachloride
	CuI	chloroform
	$FeCl_3$	tetrahydrofuran
	$SiCl_4$	hexane
	$AlCl_3$	
	$SnCl_4$	

The number of possible combinations of these items is 600. Of these, 296 were tested at the bench. The reactions were monitored by high-resolution gas chromatography for 48 h to ensure that the isomers were stable and did not equilibrate. Of these, 162 systems afforded the Fischer indole reaction; the remaining systems failed. A list of the 162 successful reaction systems is given refs 2 and 12 and is not reproduced here.

To analyse the results, PLS modelling was used. The regioisomeric excess, re, was used as the response variable. The X-block contained the principal property scores of the selected items. For the ketones the υ parameter given by Charton[15] was included to take steric effects of the side chains into account. The X-block was expanded by including all columns of the cross-product and the squared variables. A three-component PLS model was significant according to cross-validation and accounted for 87% of the variance of the response. Analysis of the PLS weight plots afforded the following conclusions as to which factors influence the regioselectivity:

ketones: steric bulk of the side chain

solvent: polarisabilty and lipophilic properties

Lewis acid: none

The chemical conclusion drawn was that to increase the selectivity, the effects of steric congestion should be amplified. One way to accomplish this would be to use an acid bound to a solid matrix. To make a long story short, regioselective indolisation can be obtained with zeolites as acid catalysts, and this afforded a new and highly efficient procedure for the reaction.[16]

General Strategy for Selection of Test Systems. A fairly large number of experimental runs were used in the Fischer study shown above. The question was whether it would have been possible to reach the same conclusions from a considerably smaller set of test system. The answer is *yes*, and such a strategy is outlined below.

The basis of the strategy is the principal properties of the reaction space. The first thing to do is to clearly state the

[11] See, for example: Hammett, L. P. *Physical Organic Chemistry*, 2nd ed.; McGraw-Hill: New York, 1970.

[12] Prochazka, M. P.; Carlson, R. *Acta Chem. Scand.* **1989**, *43*, 651.

[13] Fischer, E.; Jourdan, F. *Ber. Dtsch. Chem. Ges.* **1883**, *16*, 2241.

[14] Robinson, B. *The Fischer Indole Synthesis*; Wiley: Chichester, 1982.

[15] Charton, M. *Top. Curr. Chem.* **1982**, *117*, 57.

[16] Prochazka, M. P.; Eklund, L.; Carlson, R. *Acta Chem. Scand.* **1990**, *44*, 610; Prochazka, M. P.; Eklund, L.; Carlson, R. *Acta Chem. Scand.* **1990**, *44*, 614.

objectives of the study and then to analyse the chemical problem. From this analysis, determine which descriptors should be used for determining the principal properties for each "dimension" of the reaction space so that the relevant intrinsic molecular properties are portrayed. Collect descriptor data for sets of possible test candidates. These data sets can contain a large number of possible test candidates. Determine the principal properties. From the score plots for each dimension of the reaction space make a selection of a subset of chemically relevant test items so that the interesting range of variation of the properties is spanned by the selected candidates.

The next step is to determine how detailed the information is required and then to determine the type of model that should be used:

* a linear model
* an interaction model
* a full quadratic model

Construct the full combinatorial library matrix containing all combinations of the selected test items. This will actually be a full factorial design in the discrete settings. Then, convert the combinatorial library into the matrix, **D**, containing the corresponding principal property scores as variables.

Construction of the Design. Construct the candidate model matrix, **C**, by expanding **D** with columns for the cross-products and the squared variables included in the model. The space spanned by the columns in **C** is called the *reaction model space*. From **C**, experiments are then selected to define the model matrix **X** so that the experiments selected span the row space of **C**, which is equivalent to saying that the experiments selected should span the reaction model space. How this can be accomplished in shown below. A complete treatment with all mathematical details is given in ref 16. Here follows only a brief account of the general principles.

Before we treat the Fischer indole synthesis, we will show the principles using a simpler example, viz. the selection of test solvents. Assume that we wish to determine how the properties of the solvent influence the outcome, *y*, of a reaction and that a quadratic model will be necessary. Assume also that the properties of the solvents are adequately described by their principal properties. We will use the solvent data given in ref 2. This data set contains 103 solvents characterised by nine property descriptors, and it gives two significant principal components. The score vectors, $\mathbf{t}_1$ and $\mathbf{t}_2$ display the between-solvent variation in the principal property space. The score values t_{1r} and t_{2r} are the coordinates of solvent *r* when the descriptors are projected onto the two first principal components. A quadratic model that relates the outcome of the experiment to the principal properties will therefore be

$$y = \beta_0 + \beta_1 t_1 + \beta_2 t_2 + \beta_{11} t_1^2 + \beta_{22} t_2^2 + \beta_{12} t_1 t_2 + e$$

The response model can be written

$$y = \mathbf{X}\beta + \mathbf{e}$$

where **X** is the model matrix, β is the vector of model coefficients to be estimated, and **e** is a vector of errors. The best estimate **b** of the coefficient vector β in the least-squares sense

$$\mathbf{b} = \mathbf{X}^{\dagger} y$$

where $\mathbf{X}^{\dagger}$ is the pseudo-inverse of **X** as given by *singular value decomposition*, SVD. If **X** has full column rank, the pseudo-inverse simplifies to

$$\mathbf{X}^{\dagger} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

which yields the ordinary least-squares estimate of **b**. It is, of course, possible to fit the model by PLS. A good design for fitting the model by the pseudo-inverse will also be a good design for fitting the model by PLS. The problem is now reduced to find test solvents that define **X** so that the coefficients can be estimated. The model matrix **X** is obtained by the following procedure.

After the expansion of **D** to **C** by including a column of ones (corresponds to the constant term in the model), columns for the cross-product and the squared variables, the (103 × 6) candidate model matrix, **C**, is obtained. The next step is to factor **C** by a singular value decomposition

$$\mathbf{C} = \mathbf{USV}^{\mathrm{T}}$$

where $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2\ ...\ \mathbf{u}_6]$ is defined by the eigenvectors $\mathbf{u}_i$ of the correlation matrix $\mathbf{CC}^{\mathrm{T}}$; $\mathbf{V}^{\mathrm{T}}$ is the transpose of $\mathbf{V} = [\mathbf{v}_1\ \mathbf{v}_2\ ...\ \mathbf{v}_6]$ which is defined by the eigenvectors of the variance−covariance matrix $\mathbf{C}^{\mathrm{T}}\mathbf{C}$, and **S** is a diagonal matrix in which the diagonal elements, $\sigma_{1,}$, ..., σ_6, are the square roots of the eigenvalues of $\mathbf{CC}^{\mathrm{T}}$ (or $\mathbf{C}^{\mathrm{T}}\mathbf{C}$, they are equal). The next step is to find the row vector $\mathbf{c}_m$, in **C** that is most parallel to the first singular vector $\mathbf{v}_1$, i.e the vector corresponding to the largest singular value. It is not likely that we can find a solvent for which the corresponding row in **C** is perfectly parallel to $\mathbf{v}_1$, but the best choice would be the row for which the absolute value of the scalar product $|\mathbf{c}_m \cdot \mathbf{v}_1|$ is as large as possible. The corresponding solvent vector points in the direction showing the largest variation in the model space. The first selected solvent by these principles is *N*-methylacetamide, a highly polar aprotic solvent. The corresponding row is added as the first row in the designed model matrix **X**. To select the next solvent, we remove from **C** the properties already taken into account by the first selected one. This produces a new matrix **C′** by the following transformation

$$\mathbf{C}' = \mathbf{C}[1 - (\mathbf{c}_m{}^{\mathrm{T}}\mathbf{c}_m/\mathbf{c}_m\mathbf{c}_m{}^{\mathrm{T}})]$$

The procedure is then repeated using **C′** as input. The second solvent thus selected is iodobenzene, a highly polarisable, nonpolar solvent. Removing its properties from **C′** and repeating the procedure affords the following selection of six solvents that span the model space: *N*-methylacetamide, iodobenzene, sulfolane, pentane, methanol, and 1,1,1-trichloroethane. This selection is chemically reasonable, it spans the polarity−polarisability properties and corresponds to a selection that could have been made by mere intuition. It is, however, not likely that estimates of the coefficients of the quadratic model will have a high precision when determined

Table 6. **Reaction systems in the Fischer indole study selected by the SVD algorithm**

entry	(Id)[a]	ketone	Lewis acid	solvent	re[b]
1	(33)	2-hexanone	BF_3	carbon disulfide	100.0
2	(48)	3-undecanone	BF_3	sulfolane	23.6
3	(28)	2-hexanone	$TiCl_4$	sulfolane	100.0
4	(27)	2-hexanone	ZnI_2	sulfolane	100.0
5	(98)	1-phenyl-2-butanone	BF_3	THF	60.0
6	(11)	3-hexanone	$SbCl_5$	carbon disulfide	54.0
7	(55)	3-undecanone	BF_3	carbon disulfide	21.0
8	(89)	1-phenyl-2-butanone	BF_3	1,2-dichlorobenzene	62.0
9	(56)	3-undecanone	ZnI_2	carbon disulfide	34.2
10	(52)	3-undecanone	PCl_3	sulfolane	31.0
11	(155)	1-phenyl-2-butanone	$TiCl_4$	hexane	56.0
12	(128)	3-hexanone	CuCl	carbon tetrachloride	43.2
13	(144)	2-hexanone	$SbCl_5$	THF	100.0
14	(125)	3-hexanone	$TiCl_4$	Quinoline	34.0
15	(151)	3-undecanone	PCl_3	hexane	32.0
16	(42)	2-hexanone	$AlCl_3$	1,2-dichlorobenzene	100.0
17	(81)	1-phenyl-2-butanone	$AlCl_3$	sulfolane	28.0
18	(93)	1-phenyl-2-butanone	PCl_3	1,2-dichlorobenzene	52.0
19	(34)	2-hexanone	ZnI_2	carbon disulfide	100.0
20	(47)	2-hexanone	$AlCl_3$	THF	98.0
21	(150)	3-undecanone	$ZnCl_2$	hexane	14.0
22	(85)	1-phenyl-2-butanone	$ZnCl_2$	carbon disulfide	70.0
23	(71)	3-undecanone	$TiCl_4$	THF	40.0
24	(104)	5-methyl-3-heptanone	BF_3	sulfolane	100.0
25	(99)	1-phenyl-2-butanone	ZnI_2	THF	88.0
26	(112)	5-methyl-3-heptanone	ZnI_2	carbon disulfide	100.0
27	(133)	3-hexanone	BF_3	chloroform	43.2
28	(115)	5-methyl-3-heptanone	$TiCl_4$	1,2-dichlorobenzene	100.0
29	(111)	5-methyl-3heptanone	BF_3	carbon disulfide	100.0
30	(123)	5-methyl-3-heptanone	PCl_3	THF	100.0
31	(129)	3-hexanone	ZnI_2	carbon tetrachloride	48.2
32	(6)	3-hexanone	$AlCl_3$	sulfolane	63.2
33	(124)	5-methyl-3-heptanone	$AlCl_3$	THF	100.0
34	(13)	3-hexanone	$AlCl_3$	carbon disulfide	60.4
35	(24)	3-hexanone	$FeCl_3$	THF	58.0

[a] The numbers within parentheses refer to the run number in the entire data set (162 runs) given in refs 2, 17. [b] Percent regioisomeric excess.

from such a small selection of test solvents. One way to increase the precision would be to make replicate runs of the experiments, but this will not yield any new information as to the roles played by the properties of the solvent. Instead, we suggest another way to increase the precision of the estimated coefficients. The confidence limits of the estimated coefficients are proportional to $1/\sigma_i$, the reciprocal of the singular values of the model matrix $\mathbf{X}$. To increase the precision we should therefore select additional solvents from the candidate matrix $\mathbf{C}$ in such a way that the smallest singular value of $\mathbf{X}$ (the model matrix of already chosen experiments) is increased. This can be accomplished by selecting new solvents for which the corresponding row in $\mathbf{C}$ is most parallel to the singular vector $\mathbf{v}_i$ that corresponds to the smallest singular value σ_i of $\mathbf{X}$. A full account of these aspects is given in ref 17. By this principle, additional solvents can be selected to improve the precision of the estimated model parameters, and the following next six complementary solvents are: morpholine, pyridine, 1,2-dichloroethane, *cis*-decaline, 1,2-diaminoethane, and diglyme.

This was just an example to illustrate the principles. In a real case, we should of course only include really relevant solvents in the candidate matrix.

The SVD Algorithm and the Fischer Indole Experiments. The question was, wether the conclusions drawn after the experiments with 162 different reaction systems could have been reached with fewer experiments. Singular value decomposition of the candidate model matrix corresponding to a full quadratic model (44 columns) and 162 rows afforded 35 distinct singular values. Hence, the model matrix does not have a full column rank. To span the row space 35 experiments were selected by the singular vectors. These experiments are summarised in Table 6. A PLS model was then fitted to link the regioisomeric excess to the model matrix. A two-component model was significant (cross-validation) and explained (70 + 23 = 93% of the variance of the response, $Q^2 = 0.519$, and 0.416, respectively. Normal probability plots of the PLS loading weights of the $\mathbf{X}$ block clearly identified the important variables as outliers from the noise line, Figure 6. These variables are exactly the same as were found from the PLS model established from all of the 162 experiments.

It is interesting to see how the quality of the predictions by the model varies when the number of included experiments increases. The models were fitted using the pseudo-inverse, and in Figure 7 the prediction error sum of square (PRESS) for the entire data set (162 experiments) is plotted

(17) Carlson, R.; Carlson, J.; Grennberg, A. *J. Chemom.* **2001**, *15*, 455.
(18) Björk, Å. *Numerical Methods for Least Squares Problems*; SIAM: Philadelphia, PA, 1996.

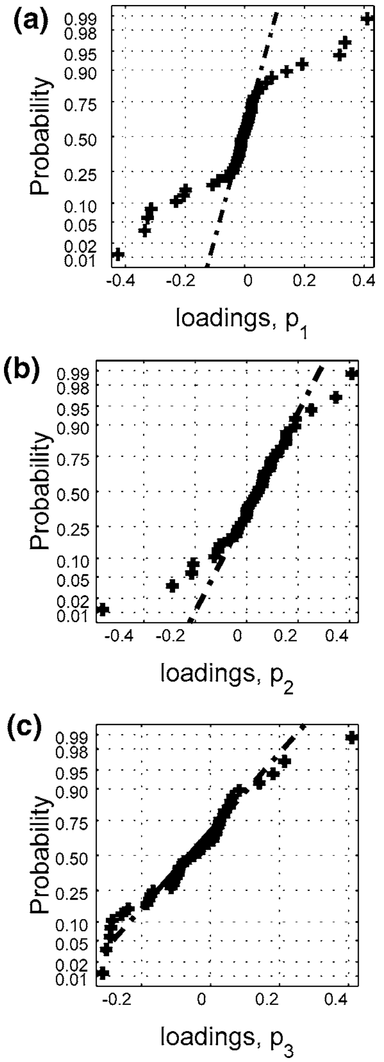


Figure 6. **Cumulative normal probability distributions of the PLS X-block weights. The significant variables have weights that appear as outliers from a normally distributed noise.**

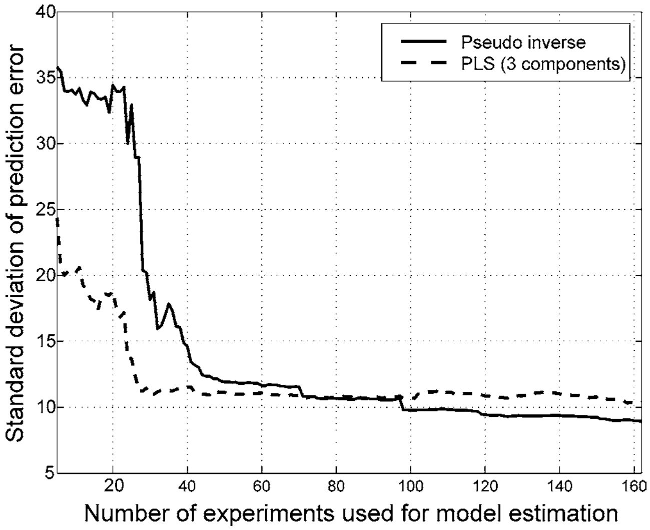


Figure 7. **PRESS vs the number of experiments in the design.**

vs the number of included experiments in the model matrix. It is seen that there is a rapid drop in the prediction errors when the maximum rank is reached (35 experiments); after that there are only minor improvements of the predictions by the models. Running more than ca. 40 experiments would therefore be a waste of time.

A Note on D-Optimal Designs. It can be argued that the design problem discussed above can be solved by first assigning the model and then determining a D-Optimal design by a search among the candidate experiments. However, when we run chemical experiments, we must take our background chemical knowledge into account and use this in the design process. It is often the case that we know beforehand that certain combinations of substrate, reagents, and solvent will not work. Such combinations should therefore be excluded from the candidate experiments. Such truncations due to our chemical knowledge of the possible candidates impose restrictions to the possible variation in the reaction model space. By this, the candidate model matrix runs the risk of being singular. In such cases, any attempt to establish a D-Optimal design by searching combinations of candidate experiments will fail. This problem is overcome by the SVD algorithm.

Conclusions

The starting point of any experimental study is a clear statement of the objectives. The next step is to identify the problems that have to be solved before we can reach the goal. In this process we have to use all our background knowledge and previous experience. When the problems have been identified, we can pose detailed questions to our experimental systems. These questions are the starting point for the design of experiments so that the results obtained in these experiments are likely to provide the answers to our questions. Multivariate models make it possible to determine the roles played by the experimental variables and the properties of the reaction systems. This gives clues to a better understanding of the chemistry involved.

Statistically designed experiments, principal component analysis, and PLS modelling are very powerful tools that open up efficient strategies for experimental studies. Commercially available software packages at low prices make these methods readily accessible to experimenters. Hence, there is no excuse to publish results and "new synthetic methods" that have been obtained in poorly designed experiments.

Acknowledgment

We thank the organisers of this symposium for their kind invitation. We thank all past and present students and coworkers who have contributed to the field of synthetometrics. A special thank-you is directed to those who have contributed to the examples given in this paper. Their names appear in the reference list. We thank the *Journal of Chemometrics*, *Elsevier Science Publishers*, and *Acta Chemica Scandinavica* for their kind permission to reproduce the examples in this paper. Financial support to R.C. from the Norwegian Research Council is gratefully acknowledged.

Received for review November 15, 2004.

OP040022W