Assessing the Correlation between Descriptors of Catalytic Materials and Their Performance

Martin Holena[a,b] [*], *Norbert Steinfeldt*[a], *Manfred Baerns*[c]

(a) Leibniz Institute for Catalysis, Albert-Einstein-Str. 29a, D-18059 Rostock, Germany

(b) Institute of Computer Science, Academy of Sciences, Prague, Czech Republic

(c) Fritz-Haber Institute, Faradayweg 4-6, D-14195 Berlin, Germany

ABSTRACT

In the paper, the methodology of correlating continuous descriptors of catalytic materials with their performance is addressed. Continuous descriptors are typically molar fractions of individual components, whereas the performance is represented most frequently by yields of reaction products or conversions or selectivity of educts. The existence of various correlation measures is recalled, designed specifically to express the correlation between given random variables with a single number. The paper suggests that in the analysis of catalytic experiments data, the application of correlation measures should complement the usually employed QSAR and similar models, which have a more ambitious objective of modeling the quantitative relationships between catalyst descriptors and performance, but usually suffer from the amount of data collected in the experiment being too small. In addition, it compares

[*] Corresponding author. E-mail: martin.holena@catalysis.de

correlation measures on the one hand with the analysis of variance, on the other hand with regression trees. The application of correlation measures and their comparison with the analysis of variance and with regression trees is illustrated by a detailed case study using data from high-temperature synthesis of hydrocyanic acid.

I. INTRODUCTION

The overall objective of the analysis of data from catalytic experiments is typically stated as follows: Find *correlations* between the *composition* of catalytic materials (and possibly *other descriptors* of their *properties*, or also of *reaction conditions*) and their *catalytic performance*, usually represented by yields of reaction products, conversions or selectivity of educts, or some more complicated performance function incorporating them[1–12]. In statistics, the correlation of a pair or a group of random variables denotes in general some departure from fully independent behavior of those variables, and it is taken as an indication of the existence of some relationship between them. Several kinds of such a departure exist, and various correlation measures have been proposed to quantify with a single number the extent to which a given pair or group of variables shows a correlation of the particular kind (cf. Section II). In publications stating the above objective of the analysis of catalytic data, however, more ambitious methods are usually employed than only the application of some correlation measure: modeling the quantitative relationships between catalyst descriptors and performance with some nonlinear regression model, frequently called *quantitative structure-activity relationship (QSAR), quantitative composition-activity relationship,* or *quantitative structure-property relationship* in this context[13–18]. Due to the highly nonlinear nature of those relationships, nonlinear regression models are the primary choice, most popular among them being *feed-forward artificial neural networks* of both main types – multilayer perceptrons and networks with radial basis functions[3–5,12–14,18]. Among other regression models, *polynomial* and *logistic* regression, *partial least-squares* regression and *regression trees* have been used to this end[6,7,10,14,17], in recent years also *Gaussian-process* regression (sometimes called *kriging*) and *support vector* regression[15,16].

The attractive feature of regression models is that they do not restrict the characterization of how catalyst descriptors correlate with performance to only a single number, like correlation measures do. Instead, they reveal how the performance changes with changing values of descriptors (e.g., with changing values of molar fractions of individual components). To this end, they use functions of many variables (as many as there are descriptors), which represent not only the influence of individual descriptors on the catalyst performance, but also the influence of interactions between theoretically arbitrary groups of descriptors. However, this feature is at the same time the source of a serious difficulty with regression models – compared to correlation measures, they need a very large amount of data. A general rule is that it increases exponentially with the number of considered descriptors. Hence, even if data about as few as 2 catalytic materials were sufficient to model with a desired reliability catalyst performance that depends solely on a single descriptor, then data about 2^{10}=1024 materials are needed to model with the same reliability performance that depends on 10 descriptors (a rather modest number in nowadays catalytic experiments). For practical reasons, so many catalytic materials are hardly tested in a single experiment – as a matter of fact, some published applications of artificial neural networks to catalysis involved only several dozens of catalysts (see refs. [22–28]). Therefore, it is important to complement the application of regression models with analysis of variance, which allows to fully control the involved interactions between descriptors, as well as with descriptor-wise application of various correlation measures. Although the information conveyed by such measures is restricted to a sigle-number characterization of the relationship betweeen the performance measure and a particular descriptor, it is more reliable than the information conveyed by regression models based on the same data. To survey the plethora of existing correlation measures and to illustrate their application to catalytic data is the objective of this paper.

An overview of available correlation measures of various kinds is given in the next section. Section III explains the relationship between results obtained with correlations measures and those obtained with the analysis of variance and regression trees, a regression model that also in a straightforward way indicates the influence of particular catalyst descriptors on its performance. Finally, an application of main correlation measures to investigating the correlation of catalyst descriptors with its performance is illustrated on data from high-temperature synthesis of hydrocyanic acid in Section IV.

II.IMPORTANT MEASURES OF CORRELATION BETWEEN RANDOM VARIABLES

As was already recalled in te introduction, different measures of correlation between two random variables exist, indicating different kinds of relationship between those variables, e.g., between a particular descriptor of the catalytic material and a variable representing its catalytic performance. In this section, all important correlation measures encountered in the literature will be explained. The explanation underlies two restrictions:

1. It deals only with ordinal data and ordinal variables (though not necessary continuous). This has two important reasons:

- Most usually, positive correlation between random variables is understood as a concordance between the tendencies of the correlated variables to assume high or low values. This is, of course, applicable only to ordinal variables. Below, also a more general approach based on general dependence will be explained. However, the commonly encountered measures corresponding to that approach, Spearman’s correlation coefficient and Schweizer and Wolff’s measure, are actually also used only for ordinal data[29–34]. Although their general definition covers also nominal variables, the methods for their estimation from data samples require the data to be ordinal, and some of their properties, which will be described below, actually hold only for continuous random variables.

- The variable representing catalytic performance (such as yield or conversion) is not only ordinal, but even continuous, and we are primarily interested in the correlation of performance with ordinal descriptors, mainly with fractions of individual components. To correlate performance with nominal variables, such as the kind of employed support, methods applicable to nominal data need to be used (see, e.g., refs. [35–38]), among which we have found the analysis of variance most useful, briefly recalled in Section III.

2. It concerns only pairs of random variables and is intended to be applied to the correlation between a descriptor (e.g., fraction of a particular component) and a catalytic performance (e.g., yield or conversion). Technically, most of the measures can be generalized to groups of three or more random variables[39–42], for some of them even various such generalizations are possible[41,42]. However, it is not appropriate to apply such measures to groups of more than two descriptors because the relationship between such descriptors (e.g., between the fractions of two components in catalyst) cannot be interpreted as a relationship between two random variables. The values of descriptors for particular catalytic materials are fixed during the design of those materials, thus apart from possible imprecision originating during the synthesis of the designed catalysts, no randomness is involved in a relationships between different descriptors after the design has finished, and such a relationship can be considered deterministic.

General dependence. Most generally, correlation between random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ only means some departure from their independence. From a stochastic point of view, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent if their joint distribution $\boldsymbol{H}$ is the product of the distribution $\boldsymbol{F}$ of $\boldsymbol{X}$ and the distribution $\boldsymbol{G}$ of $\boldsymbol{Y}$, i.e., if $\boldsymbol{H(x, y) = F(x)\cdot G(y)}$, or equivalently $\boldsymbol{H(x, y) - F(x)\cdot G(y)} = 0$, for all

$\boldsymbol{x}$ and $\boldsymbol{y}$. Hence, if $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y}) > 0$ or $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y}) < 0$ for any pair $\boldsymbol{x}, \boldsymbol{y}$, then this already indicates some correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

As an example, think of $\boldsymbol{X}$ as describing the molar fraction of Mo-oxides in the active shell of the catalyst, of $\boldsymbol{Y}$ as describing the yield of a particular reaction product. Let the ratio of the fraction of considered catalysts containing Mo in their active shell to that not containing Mo be 2:8 (i.e., $\boldsymbol{F}(0) = 0.8$), whereas the ratio of the fraction of considered catalysts with yield above 25 % to the fraction with yield up to 25 % be 4:6 (i.e., $\boldsymbol{F}(0.25) = 0.6$). If now the yield is independent of the fraction of Mo-oxides in the active shell of the catalyst, the set of all available catalysts divides as follows:

- the fraction of catalysts containing Mo and with yield above 25 % is 0.12;
- the fraction of catalysts containing Mo and with yield up to 25 % is 0.08;
- the fraction of catalysts not containing Mo and with yield above 25 % is 0.48;
- the fraction of catalysts not containing Mo and with yield up to 25 % is 0.32.

If any of these four fractions is different, then this indicates correlation between yield and the fraction of Mo-oxides. Moreover, the whole sequence of those four fractions is different in such a case: For example, if the fraction of catalysts containing Mo and with yield above 25 % is actually 0.15, then this sequence is 0.15, 0.05, 0.45, 0.35, indicating a positive correlation, whereas if that fraction is only 0.1, the sequence is 0.1, 0.1, 0.5, 0.3, indicating a negative correlation.

A simple way how to measure such a general dependence is to average the difference $\boldsymbol{H}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y})$ with respect to the random vector $(\boldsymbol{X}, \boldsymbol{Y})$, i.e., to compute the expectation

$$\mathbf{E}\big(\boldsymbol{H}(\boldsymbol{X},\boldsymbol{Y}) - \boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{Y})\big). \qquad (1)$$

Calculating this expectation always gives values in the interval [-1/12, 1/12]. The lower bound -1/12 is reached, e.g., for $\boldsymbol{Y} = -\boldsymbol{X}$, the upper bound 1/12 is reached, e.g., for $\boldsymbol{Y} = \boldsymbol{X}$

(cf. ref. [34]). Usually, a normalization by the constant 12 transforms the interval [-1/12, 1/12] to the interval [-1, 1], leading to the measure

$$\boldsymbol{\rho}_{X,Y} = 12\mathbf{E}\big(\boldsymbol{H}(\boldsymbol{X},\boldsymbol{Y}) - \boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{Y})\big), \tag{2}$$

called *Spearman's correlation coefficient* [30,31,33]. An unbiased estimate $\boldsymbol{r}_{X,Y}$ of $\boldsymbol{\rho}_{X,Y}$ based on a sample $(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)$ can be computed according to [30]:

$$\boldsymbol{r}_{X,Y} = 1 - \frac{6\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}}\big(\mathrm{xrank}(\boldsymbol{k})-\mathrm{yrank}(\boldsymbol{k})\big)^2}{\boldsymbol{n}(\boldsymbol{n}^2-1)}, \tag{3}$$

where an increasing ordering of the values $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ and an increasing ordering of the values $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ are considered, and xrank($\boldsymbol{i}$) is the position of $\boldsymbol{x}_i$ within the former, whereas yrank($\boldsymbol{i}$) is the position of $\boldsymbol{y}_i$ within the latter. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are, in addition to ordinality, even continuous, then Spearman's correlation coefficient has the following important properties [43]:

(i) $-1 \leq \boldsymbol{\rho}_{X,Y} \leq 1$;

(ii) $\boldsymbol{\rho}_{Y,X} = \boldsymbol{\rho}_{X,Y}$, whereas $\boldsymbol{\rho}_{-X,Y} = \boldsymbol{\rho}_{X,-Y} = -\boldsymbol{\rho}_{X,Y}$;

(iii) if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then $\boldsymbol{\rho}_{X,Y} = 0$;

(iv) if $\boldsymbol{f}$ is a function strictly increasing on the value set of $\boldsymbol{X}$ and $\boldsymbol{g}$ is a function strictly increasing on the value set of $\boldsymbol{Y}$, then $\boldsymbol{\rho}_{f(X),g(Y)} = \boldsymbol{\rho}_{X,Y}$;

(v) if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are with probability 1 strictly increasing functions of each other, then $\boldsymbol{\rho}_{X,Y} = 1$, whereas if they are with probability 1 strictly decreasing functions of each other, then $\boldsymbol{\rho}_{X,Y} = -1$.

The properties (ii) and (iii) actually hold even if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are only ordinal, and the properties (ii)–(iv) hold also for the unnormalized measure (1).

The implications in the properties (iii) and (v) cannot be reversed, i.e., from $\boldsymbol{\rho}_{X,Y} = 0$ cannot be concluded that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, and from $\boldsymbol{\rho}_{X,Y} = 1$ / $\boldsymbol{\rho}_{X,Y} = -1$ cannot be concluded that they are increasing / decreasing functions of each other. For example, the difference $\boldsymbol{H}(\boldsymbol{x},\boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y})$ can have large positive values in some areas and large negative values in other areas, but they still can average out to zero. Consequently, Spearman's correlation coefficient can serve neither as an indicator of the extent to which two variables (e.g., the fraction of a particular catalyst component and yield of a particular product) are independent, nor the extent to which they are increasing or decreasing functions of each other. For that purpose, a modification of Spearman's correlation coefficient has been proposed by Schweizer and Wolff in [32], and consists in replacing $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y})$ in (2) with its absolute value $|\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{F}(\boldsymbol{x})\cdot\boldsymbol{G}(\boldsymbol{y})|$. Hence, *Schweizer and Wolff's measure* is given by:

$$\boldsymbol{\sigma}_{\boldsymbol{X},\boldsymbol{Y}} = 12\mathrm{E}|\boldsymbol{H}(\boldsymbol{X},\boldsymbol{Y}) - \boldsymbol{F}(\boldsymbol{X})\boldsymbol{G}(\boldsymbol{Y})|.$$

For continuous $\boldsymbol{X}$ and $\boldsymbol{Y}$, Schweizer and Wolff's measure in addition to changing the implications in the properties (iii) and (v) to equivalences also modifies the remaining properties (i), (ii) and (iv), as follows:

(i') $0 \le \boldsymbol{\sigma}_{X,Y} \le 1$;

(ii') $\boldsymbol{\sigma}_{Y,X} = \boldsymbol{\sigma}_{X,Y}$;

(iii') $\boldsymbol{\sigma}_{X,Y} = 0$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent;

(iv') if $\boldsymbol{f}$ is a function either strictly increasing or strictly decreasing on the value set of $\boldsymbol{X}$ and $\boldsymbol{g}$ is a function either strictly increasing or strictly decreasing on the value set of $\boldsymbol{Y}$, then $\boldsymbol{\sigma}_{f(X),g(Y)} = \boldsymbol{\sigma}_{X,Y}$;

(v') $\boldsymbol{\sigma}_{X,Y} = 1$ if and only if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are either with probability 1 strictly increasing functions of each other or with probability 1 strictly decreasing functions of each other, or equivalently, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are with probability 1 strictly monotone functions of each other.

Due to the properties (i'), (iii') and (v'), Schweizer and Wolff's measure can be interpreted as an intensity of correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, in the sense of a distance of the relationship between them from the situation that they are completely independent, and closeness of that relationship to the situation that they are strictly monotone functions of each other. In the catalytic context, such an intensity of correlation between, e.g., yield and the fraction of a particular component means the distance from the situation that yield is completely independent of the fraction of that component and closeness to the situation that yield is a strictly increasing or strictly decreasing function of the fraction of a particular component.

An estimate $\boldsymbol{s}_{X,Y}$ of $\boldsymbol{\sigma}_{X,Y}$ based on a sample $((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ can be obtained through replacing $\boldsymbol{H}$, $\boldsymbol{F}$ and $\boldsymbol{G}$ with the corresponding empirical distribution functions. Consequently,

$$(4)\qquad \boldsymbol{s}_{X,Y} = \frac{\boldsymbol{12}}{\boldsymbol{n}^3}\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}}\sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n}}\left|\#\{\boldsymbol{k}:\mathrm{xrank}(\boldsymbol{k}) \leq \boldsymbol{i}, \mathrm{yrank}(\boldsymbol{k}) \leq \boldsymbol{j}\} - \frac{\boldsymbol{ij}}{\boldsymbol{n}}\right|,$$

where # stands for the number of elements.

Concordance. The probably most frequently encountered meaning of correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the concordance of higher and lower values between both variables: In the situation when $\boldsymbol{X}$ describes the fraction of some catalyst component and $\boldsymbol{Y}$ describes its catalytic performance, positive correlation means smaller fractions of that component are associated with lower performance and higher fractions with higher performance, whereas negative correlation means that higher fractions of the component are associated with lower performance and lower fractions with higher performance. A possible measure of that association is the probability that for two independent realizations $(\boldsymbol{x}_1,\boldsymbol{y}_1)$ and $(\boldsymbol{x}_2,\boldsymbol{y}_2)$ of $(\boldsymbol{X}, \boldsymbol{Y})$, $\boldsymbol{x}_1 < \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_2$ and $\boldsymbol{x}_1 > \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_2$ minus the probability that for such realizations, $\boldsymbol{x}_1 < \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_2$ and $\boldsymbol{x}_1 > \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_2$. In the context of the above example with $\boldsymbol{X}$ describing the fraction of Mo-oxides and $\boldsymbol{Y}$ describing the yield of reaction product, this is the probability that from two catalysts randomly obtained from the same population, the one with higher Mo fraction will also lead to a higher yield minus the probability that it will lead to a lower yield. That probability is called *Kendall's correlation coefficient* [29,30,32] and equals

$$(5)\qquad \boldsymbol{\tau}_{X,Y} = P[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{Y}_1 - \boldsymbol{Y}_2) > 0] - P[(\boldsymbol{X}_1 - \boldsymbol{X}_2)(\boldsymbol{Y}_1 - \boldsymbol{Y}_2) < 0],$$

where $(\boldsymbol{X}_1,\boldsymbol{Y}_1)$ and $(\boldsymbol{X}_2,\boldsymbol{Y}_2)$ are independent random vectors governed by the considered joint distribution $\boldsymbol{H}$. If at least one of the variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is continuous (which is always the case when one of them represents catalytic performance, as was already mentioned), then an unbiased estimate $\boldsymbol{t}_{X,Y}$ of $\boldsymbol{\tau}_{X,Y}$ based on a sample $((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ can be obtained through replacing the probabilities in (5) through frequencies of the corresponding events:

(1) $$\boldsymbol{t}_{X,Y} = \frac{2(\#\{(i,j):(x_i-x_j)(y_i-y_j)>0\}-\#\{(i,j):(x_i-x_j)(y_i-y_j)<0\})}{n(n-1)}.$$

Another possible measure of the concordance of higher and lower values between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the probability that a realization $\boldsymbol{x}$ of $\boldsymbol{X}$ differs from a particular summary statistic of $\boldsymbol{X}$ (such as expectation or median) in the same direction as a realization $\boldsymbol{y}$ of $\boldsymbol{Y}$ differs from the corresponding summary statistic of $\boldsymbol{Y}$. Most frequently, median is employed as the summary statistic, because of its robustness, in which case the measure is called *medial correlation coefficien*t or *Blomquist's measure*[34]. Similarly to (5), it equals

(2) $$\boldsymbol{\beta}_{X,Y} = \boldsymbol{P}[(\boldsymbol{X}-\boldsymbol{m}_X)(\boldsymbol{Y}-\boldsymbol{m}_Y) > \boldsymbol{0}] - \boldsymbol{P}[(\boldsymbol{X}-\boldsymbol{m}_X)(\boldsymbol{Y}-\boldsymbol{m}_Y) < \boldsymbol{0}],$$

where $\boldsymbol{m}_X$ and $\boldsymbol{m}_Y$ denote the medians of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Moreover, it can be shown[34] that (8) substantially simplifies to

(3) $$\boldsymbol{\beta}_{X,Y} = \boldsymbol{4H}(\boldsymbol{m}_X, \boldsymbol{m}_Y) - \boldsymbol{1}.$$

An estimate $\mathbf{b}_{X,Y}$ of $\boldsymbol{\beta}_{X,Y}$ based on a sample $((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ can again be obtained through replacing the distribution function $\boldsymbol{H}$ and the medians $\boldsymbol{m}_X$ and $\boldsymbol{m}_Y$ with their empirical counterparts. Consequently,

(4) $$\boldsymbol{b}_{X,Y} = \frac{\boldsymbol{4}}{\boldsymbol{n}}\#\{\boldsymbol{k}: \boldsymbol{x}_k \le \widehat{\boldsymbol{m}}_X, \boldsymbol{y}_k \le \widehat{\boldsymbol{m}}_Y\} - \boldsymbol{1},$$

where $\widehat{\boldsymbol{m}}_X$ and $\widehat{\boldsymbol{m}}_Y$ are the empirical medians of $\boldsymbol{X}$ and $\boldsymbol{Y}$, thus for example,

$$\widehat{\boldsymbol{m}}_X = \begin{cases} \frac{x_k+x_{k+1}}{2} & \text{if } \boldsymbol{n} = 2\boldsymbol{k}, \\ \boldsymbol{x}_{k+1} & \text{if } \boldsymbol{n} = 2\boldsymbol{k}+1. \end{cases}$$

Very important is that for $\boldsymbol{X}$ and $\boldsymbol{Y}$ continuous, both the Kendall's and the medial correlation coefficient can be shown to have the properties (i)–(v) listed above[34,43] (of course, with $\boldsymbol{\rho}_{X,Y}$ replaced by $\boldsymbol{\tau}_{X,Y}$ or $\boldsymbol{\beta}_{X,Y}$). This means that for continuous random variables, also the Spearman's $\boldsymbol{\rho}_{X,Y}$ actually measures the concordance of higher and lower values between them, in spite of being defined in a more general setting.

Correlation measure based on bounding probability distributions. Besides the Spearman's, Kendall's and medial correlation coefficients, there is one more ferequently encountered correlation measure that has the above properties (i)–(v), therefore can be interpreted as concordance of higher and lower values between the fraction of a particular catalyst component, and a variable describing the catalytic performance. This is the *Gini's coefficient*[34,43], defined

$$\begin{aligned}\gamma_{X,Y} &= \boldsymbol{P}[(\boldsymbol{X_1}-\boldsymbol{X_2})(\boldsymbol{Y_1}-\boldsymbol{Y_2})>\boldsymbol{0}]-\boldsymbol{P}[(\boldsymbol{X_1}-\boldsymbol{X_2})(\boldsymbol{Y_1}-\boldsymbol{Y_2})<\boldsymbol{0}]+\\ &+\ \boldsymbol{P}[(\boldsymbol{X_1}-\boldsymbol{X_3})(\boldsymbol{Y_1}-\boldsymbol{Y_3})>\boldsymbol{0}]-\boldsymbol{P}[(\boldsymbol{X_1}-\boldsymbol{X_3})(\boldsymbol{Y_1}-\boldsymbol{Y_3})<\boldsymbol{0}],\end{aligned} \tag{5}$$

where $(\boldsymbol{X}_1,\boldsymbol{Y}_1)$, $(\boldsymbol{X}_2,\boldsymbol{Y}_2)$ and $(\boldsymbol{X}_3,\boldsymbol{Y}_3)$ are independent random vectors, $(\boldsymbol{X}_1,\boldsymbol{Y}_1)$ governed by a particular joint distribution $\boldsymbol{H}$, $(\boldsymbol{X}_2,\boldsymbol{Y}_2)$ governed by the 2-dimensional distribution $\boldsymbol{U}(\boldsymbol{x},\boldsymbol{y})=\min(\boldsymbol{F}(\boldsymbol{x}),\boldsymbol{G}(\boldsymbol{y}))$, and $(\boldsymbol{X}_2,\boldsymbol{Y}_2)$ governed by the 2-dimensional distribution $\boldsymbol{L}(\boldsymbol{x},\boldsymbol{y})=\max(\boldsymbol{F}(\boldsymbol{x})+\boldsymbol{G}(\boldsymbol{y})-1{,}0)$. To understand the meaning of this definition, let us recall that $\boldsymbol{L}$ and $\boldsymbol{U}$ are bounding probability distributions for the joint distribution $\boldsymbol{H}$ [34]:

$$\boldsymbol{L}(\boldsymbol{x},\boldsymbol{y}) \le \boldsymbol{H}(\boldsymbol{x},\boldsymbol{y}) \le \boldsymbol{U}(\boldsymbol{x},\boldsymbol{y}).$$

According to (5), the Gini's coefficient is the sum of two differences of probabilities:

- the probability that for independent realizations $(\boldsymbol{x}_1,\boldsymbol{y}_1)$ and $(\boldsymbol{x}_2,\boldsymbol{y}_2)$ of random vectors governed by a particular joint distribution $\boldsymbol{H}$ and by the lower bounding distribution $\boldsymbol{L}$, respectively, $\boldsymbol{x}_1 < \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_2$ and $\boldsymbol{x}_1 > \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_2$ minus the probability that for such realizations, $\boldsymbol{x}_1 < \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_2$ and $\boldsymbol{x}_1 > \boldsymbol{x}_2$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_2$,
- and the probability that for independent realizations $(\boldsymbol{x}_1,\boldsymbol{y}_1)$ and $(\boldsymbol{x}_3,\boldsymbol{y}_3)$ of random vectors governed by a particular $\boldsymbol{H}$ and by the upper bounding distribution $\boldsymbol{U}$, respectively, $\boldsymbol{x}_1 < \boldsymbol{x}_3$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_3$ and $\boldsymbol{x}_1 > \boldsymbol{x}_3$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_3$ minus the probability that for such realizations, $\boldsymbol{x}_1 < \boldsymbol{x}_3$ coincides with $\boldsymbol{y}_1 > \boldsymbol{y}_3$ and $\boldsymbol{x}_1 > \boldsymbol{x}_3$ coincides with $\boldsymbol{y}_1 < \boldsymbol{y}_3$.

Using again the above example with $\boldsymbol{X}$ describing the fraction of Mo-oxides and $\boldsymbol{Y}$ describing the yield, this sum is the probability that from two catalysts, one of which was randomly obtained from a population obeying some joint distribution $\boldsymbol{H}$, and the other from a population obeying either the lower or the upper bounding distribution, the one with higher Mo fraction will also lead to a higher yield minus the probability that it will lead to a lower yield.

An estimate $\mathbf{g}_{X,Y}$ of $\boldsymbol{\gamma}_{X,Y}$ based on a sample $((\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n))$ can be obtained using the fact that it is possible to simplify (5) to[34]

$$\boldsymbol{\gamma}_{X,Y} = 2\mathbf{E}(|\boldsymbol{F}(\boldsymbol{X}) + \boldsymbol{G}(\boldsymbol{Y}) - 1| - |\boldsymbol{F}(\boldsymbol{X}) - \boldsymbol{G}(\boldsymbol{Y})|),$$

and then replacing the distributions $\boldsymbol{F}$ and $\boldsymbol{G}$ with their empirical counterparts. This leads to the estimate

$$\frac{2}{\boldsymbol{n}}\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}}\left[\left|\frac{\mathrm{xrank}(\boldsymbol{k})}{\boldsymbol{n}} + \frac{\mathrm{yrank}(\boldsymbol{k})}{\boldsymbol{n}} - 1\right| - \left|\frac{\mathrm{xrank}(\boldsymbol{k})}{\boldsymbol{n}} - \frac{\mathrm{yrank}(\boldsymbol{k})}{\boldsymbol{n}}\right|\right] =$$

(6)

$$= \frac{2}{\boldsymbol{n}^2}\sum_{\boldsymbol{k}=1}^{\boldsymbol{n}}[|\mathrm{xrank}(\boldsymbol{k}) + \mathrm{yrank}(\boldsymbol{k}) - \boldsymbol{n}| - |\mathrm{xrank}(\boldsymbol{k}) - \mathrm{yrank}(\boldsymbol{k})|].$$

The statistician C. Gini, when introducing this correlation measure, actually used a employed different estimate:

(7) $$\boldsymbol{g}_{X,Y} = \frac{\mathbf{1}}{\left\lfloor \frac{\boldsymbol{n}^2}{2} \right\rfloor}\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{n}}[|\mathrm{xrank}(\boldsymbol{k}) + \mathrm{yrank}(\boldsymbol{k}) - \boldsymbol{n} - 1| - |\mathrm{xrank}(\boldsymbol{k}) - \mathrm{yrank}(\boldsymbol{k})|],$$

where $\lfloor\ \ \rfloor$ denotes the integer part of a real number. However, the difference betweeen both estimates vanishes with increasing $\boldsymbol{n}$.

Linear dependence. According to the properties (v) and (v'), the correlation measures reviewed so far achieve their highest value, 1, whenever the dependence between the correlated variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, e.g., between a particular component fraction and yield, is described by a strictly increasing function. Similarly, the measures $\boldsymbol{\rho}_{X,Y}$, $\boldsymbol{\tau}_{X,Y}$, $\boldsymbol{\beta}_{X,Y}$, $\boldsymbol{\gamma}_{X,Y}$ achieve their lowest value, -1, whenever this dependence is described by a strictly decreasing fucntion ($\boldsymbol{\sigma}_{X,Y}$ =1 also in this case). Sometimes, the correspondence of the values 1 and -1 to such broad classes of fuctions can be disadvantageous. That disadvantage is for the measures $\boldsymbol{\rho}_{X,Y}$, $\boldsymbol{\tau}_{X,Y}$, $\boldsymbol{\beta}_{X,Y}$, $\boldsymbol{\gamma}_{X,Y}$ further increased through the already mentioned fact that the implication in the property (v) cannot be reversed, i.e., the measure can achieve the value 1 / -1 even if the variables are not incerasing / decreasing functions of each other. In such situations, another correlation measure is used, for which those values correspond to a more specific dependence, namely to linear dependence. This is the *linear correlation coefficient* of $\boldsymbol{X}$ and $\boldsymbol{Y}$ [46,47], also called *Pearson's correlation coefficient* and commonly denoted corr($\boldsymbol{X}$,$\boldsymbol{Y}$), which is defined

(8)
$$\operatorname{corr}(\boldsymbol{X},\boldsymbol{Y}) = \frac{\operatorname{cov}(\boldsymbol{X},\boldsymbol{Y})}{\sqrt{\operatorname{var}(\boldsymbol{X})\operatorname{var}(\boldsymbol{Y})}}$$

and has the following properties:

(i)–(iii), listed above for the Spearman's correlation coefficient (with $\boldsymbol{\rho}_{X,Y}$ replaced by corr($\boldsymbol{X}$,$\boldsymbol{Y}$));

(iv$^+$) corr($\boldsymbol{X}$,$\boldsymbol{Y}$) = 1 if and only if $\boldsymbol{Y}$ =a$\boldsymbol{X}$+$\boldsymbol{b}$ with probability 1, where $\boldsymbol{a}$ > 0, whereas corr($\boldsymbol{X}$,$\boldsymbol{Y}$) = -1 if and only if $\boldsymbol{Y}$ =a$\boldsymbol{X}$+$\boldsymbol{b}$ with probability 1, where $\boldsymbol{a}$ < 0.

For example, if again $\boldsymbol{X}$ describes the fraction of a particular component, and $\boldsymbol{Y}$ describes yield, then corr($\boldsymbol{X}$,$\boldsymbol{Y}$) = 1 indicates an increasing linear dependence of yield on that fraction, whereas corr($\boldsymbol{X}$,$\boldsymbol{Y}$) = -1 indicates a decreasing linear dependence.

An estimate $\mathbf{c}_{X,Y}$ of corr($\boldsymbol{X}$,$\boldsymbol{Y}$) based on a sample (($\boldsymbol{x}_1$,$\boldsymbol{y}_1$),…,($\boldsymbol{x}_n$,$\boldsymbol{y}_n$)) can be obtained through replacing the covariance and variances in (8) with their empirical counterparts:

(9)
$$\boldsymbol{c}_{X,Y} = \frac{\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_i\right)\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_i\right)}{\sqrt{\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_i\right)^2\left(y_i - \frac{1}{n}\sum_{j=1}^{n} y_i\right)^2}}.$$

III. COMPARISON WITH ANALYSIS OF VARIANCE AND REGRESSION TREES

As already mentioned in the introduction, correlation measures are not the only way how to quantify the strength of a relationship between random variables. Specifically, a quantitative analysis of the influence of random variables of arbitrary kinds (including nominal variables) on a dependent continuous variable is the objective of the *analysis of variance*. There is also an important kind of regression models, *regression trees*, the construction of which in a straightforward way incorporates quantitative information about the strength of influence of individual independent variables on the dependent variable.

Analysis of variance is an approach based on statistical hypotheses testing that quantitatively analyses the influence of varying the values of individual independent variables on the value of a continuous dependent variable [36,37]. In the area of catalysis, independent variables include most importantly variables describing the composition of the catalytic material, both qualitative (whether a particular component is or is not present in the material, or what has been used as support), and quantitative (fraction of a particular component in the material). Dependent variables, on the other hand, are the variables describing some kind of catalytic performance, notoriously exemplified by yield and conversion.

Analysis of variance assumes that each dependent variable follows some *basic statistical model*, in which the expectation of that variable is viewed as the sum of the effects of individual independent variables, called *main effects*, possibly superimposed by their *interactions* of various complexity[36,37]. The amount of available data for each combination of values of input variables determines how complex the basic model will be. The principle of the analysis of variance consists in testing the hypothesis that a particular main effect or interaction can be left out from that model without significantly changing the value of the output variable. If the tested hypothesis is valid, then both models will give the same error. Therefore, the ratio of both errors is computed in the analysis-of-variance method, and if that ratio differs significantly from the value 1, the tested hypothesis is rejected. Provided that the individual errors are normally distributed, also the distribution of the error ratio is known (it is called Fisher-Snedecor distribution). Using this distribution, the probability can be computed that the error ratio is as high as the value corresponding to the measured data, or even higher. That probability is called *achieved significance* of the test. The lower it is, the more unlikely it is that the measured data could occur if the simplified model is valid. Consequently, the more significant is then the effect/interaction that was left out from the model. For example, if the catalytic material consists of support and active components selected from a pool of 10 compounds, c1,…,c10, then in the basis model, the expectation $\mathbf{E}\boldsymbol{Y}$ of the yield $\boldsymbol{Y}$ equals the sum of main effects,

(10) $$\mathbf{E}\boldsymbol{Y} = \boldsymbol{\alpha}_{\text{support}} + \boldsymbol{\alpha}_{\text{c1}} + \cdots + \boldsymbol{\alpha}_{\text{c10}},$$

to which interactions of two variables can be added,

$$\mathbf{E}\boldsymbol{Y} = \boldsymbol{\alpha}_{\text{support}} + \boldsymbol{\alpha}_{\text{c1}} + \cdots + \boldsymbol{\alpha}_{\text{c10}} + \boldsymbol{\alpha}_{\text{support,c1}} + \cdots + \boldsymbol{\alpha}_{\text{support,c10}} + \boldsymbol{\alpha}_{\text{c1,c2}} + \cdots + \boldsymbol{\alpha}_{\text{c9,c10}},$$

or even interactions of more variables,

$$\mathbf{E}\boldsymbol{Y} = \boldsymbol{\alpha}_{\text{support}} + \boldsymbol{\alpha}_{\text{c1}} + \cdots + \boldsymbol{\alpha}_{\text{c10}} + \boldsymbol{\alpha}_{\text{support,c1}} + \cdots + \boldsymbol{\alpha}_{\text{c9,c10}} + \boldsymbol{\alpha}_{\text{support,c1,c2}} + \cdots.$$

The connection between analysis of variance and correlation is twofold:

1. If the statistical test in the analysis of variance rejects the hypothesis that a particular main efect (e.g., $\boldsymbol{\alpha}_{\text{cj}}$ for the coumpound cj in (10) can be left out, then this indicates a correlation between the corresponding independent variable and the dependent variable (in (10): between the fraction of cj and yield).
2. If an application of correlation measures reveals that that there is no correlation between certain independent variables and the dependent variable, then the main effects for such independent variables do not need to be included into the basic model. Consequently, the same amount of data allows more interactions among the remaining variables to be included instead of those main effects. Statistical testing whether each of the addded interactions can be left out from the model can provide valuable information that would not be available if they were not at first included – for example, the information that a

particular combination of active components or a combination of a particular active component and a particular support tend to increase yield or conversion.

Importantly, neither from 1. nor from 2. follows that a if there is a high correlation between the dependent variable and the independent variable corresponing to a particular compound cj in the basic statistical model (10), then the statistical in the analysis of variance has to reject the hypothesis that the main effect $\boldsymbol{\alpha}_{\mathrm{cj}}$ can be left out from (10). Indeed, the test takes into account the context of all the independent variables corresponding to any of the main effects in the basic model. On the other hand, as was shown in Section II, the correlation measures between the dependent variable and the independent variable corresponing to cj are computed only from values of those two variables, ignoring the context of the other independent variables.

Regression trees are regression models that, similarly to the analysis of variance, allow dependent variables of arbitrary kinds. Their principle consists in splitting the value set of some input variable into two parts $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ in such a way that the sum of squared errors, based on $(\boldsymbol{x}_1,\boldsymbol{y}_1),\ldots,(\boldsymbol{x}_n,\boldsymbol{y}_n)$, of the means of the regression variable $\boldsymbol{y}$ corresponding to $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$,

$$\sum_{\mathrm{x_i}\in \mathrm{S_1}}\left(\mathrm{y_i}-\frac{\sum_{\mathrm{x_i}\in \mathrm{S_1}}\mathrm{y_i}}{\#\{\mathrm{i}:\mathrm{x_i}\in \mathrm{S_1}\}}\right)^2+\sum_{\mathrm{x_i}\in \mathrm{S_2}}\left(\mathrm{y_i}-\frac{\sum_{\mathrm{x_i}\in \mathrm{S_2}}\mathrm{y_i}}{\#\{\mathrm{i}:\mathrm{x_i}\in \mathrm{S_2}\}}\right)^2 \tag{11}$$

is minimized over all possible splits $(\boldsymbol{S}_1,\boldsymbol{S}_2)$ of the value sets of all input variables. If the considered input variable is continuous, than only splits of the form $(\boldsymbol{S_1}=\{\boldsymbol{x}<v\},\ \boldsymbol{S_2}=\{\boldsymbol{x}\geq \boldsymbol{v}\})$ for some value $\boldsymbol{v}$ are considered. Both $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are then split again in the same way, possibly using different input variables. Such splitting continues as long as needed, forming a hierarchy of rectangular areas in the space of continuous input variables. The name of the method originated from the fact that such hierarchies can easily be visualised as tree-like graphs.

The sum of squared errors in (11) is actually the sum of sample variances of the regression variable $\boldsymbol{y}$ on the sets $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$. Since that sum is minimized over all input variables, the fact that a particular input variable has been selected in this minimization indicates its influence on the regression variable $\boldsymbol{y}$. More precisely, it indicates the influence of the selected variables in the sense that these variables allow to split the input space in a way leading to minimal sum of sample variances of $\boldsymbol{y}$ on both parts.

Depending on the number of such splits that are consecutively performed, trees of different sizes can be obtained. The most appropriate tree size is usually chosen using cross-validation:

- The set of available data about catalytic materials is randomly partitioned into k parts of approximately equal size.
- With each possible tree size, $\boldsymbol{k}$ trees are constructed, using for the construction of each of them one $\boldsymbol{k}$–1 parts, and leaving the remaining $\boldsymbol{k}$-th part to measure the sum of squared errors of predictions by the cnstructed tree $\boldsymbol{T}$,

$$\mathrm{SSE}(\boldsymbol{T}) = \textstyle\sum_{\boldsymbol{x_i} \in \boldsymbol{k}\text{-th part}} (\boldsymbol{T}(\boldsymbol{x_i}) - \boldsymbol{y_i})^2.$$

- To assess the appropriateness of each tree size, the SSE values for the test data are averaged over all $\boldsymbol{k}$ trees with that size.

IV. CASE STUDY WITH DATA FROM THE SYNTHESIS OF HCN

The correlation measures described in Section II, as well as their comparison with analysis of variance and regression, are now illustrated in a case study using data from the investigation of catalytic materials for the *high-temperature synthesis of hydrocyanic acid*. This investigation and its results were recently described in ref. [10]. The investigation was performed through high-throughput experiments in a circular 48-channel reactor. In most of these experiments, the composition of the materials was designed using a genetic algorithm developed specifically for the optimisation of solid catalysts[1,44,45].

Involved variables. The composition and preparation of the catalytic materials studied and the conditions to which they had been exposed have been described in detail in [10]. Here, only those facts are recalled that are important for understanding which variables are considered in this case study.

(i) All the materials tested consisted of a support and up to six metal additives. As support, 15 materials were tested: pure α-Al_2O_3 (alsint), as well as the compounds AlN, Mo_2C, TiB_2, TiN, Nb_2O_3, BN, ZrO_2, Sm_2O_3, SrO, CaO, MgO, TiO_2, SiC, and Si_3N_4, bound in an alumina matrix.

(ii) Eleven metal additives were used as active elements: Y, La, Zr, Mo, Re, Ir, Ni, Pt, Zn, Ag and Au. It is important to realize that the fractions of these compounds are not completely independent since their weight fractions sum up to the total weight fraction of the active part, which was fixed to 2.2 wt%.

(iii) As far as catalytic performance is concerned, the primary interest is in HCN yield. It was calculated form the CH_4 concentration and the reactor inlet and oulet, assuming HCN as the only product. The degree of conversion of NH_3 was considered uninteresting due to a low variability, the conversion of NH_3 being always nearly complete.

(iv) The inlet composition of the feed gas amounted to 10.7 vol. % $_{NH3}$, 9.3 vol. % CH_4 and 80 vol. % Ar. The reaction temperature was 1373K. At this temperature, it is thermodynamically possible to convert the introduced CH_4 completely to HCN.

For data analysis, the data collected in this case study are described by the following variables:

- A nominal input variable describing the support of the catalyst.
- Eleven continuous input variables describing the fractions of the metal additives Y, La, Zr, Mo, Re, Ir, Ni, Pt, Zn, Ag, and Au in the active shell of the catalyst.
- One continuous output variable describing the HCN yield.

Choice of support. Recall from Section II that the correlation measures can be applied only to the 11 continuous input variables and the 3 continuous output variables. Nevertheless, there is a simple way how to take ito account also te inflluence of the remaining, nominal variable, i.e. the influence of support: to apply the measures for each support separately. The distribution of the 696 available cataysts accordding to their support is depicted in Figure 1. We decided to restrict the subsequent illustration of applying correlation measures to the supports Si_3N_4 (160 catalysts) and SiC (123 catalysts). Figure 2 shows how frequently catalytic materials with those two supports contain each of the 11 active metals. It can be seen that, with the exception of La and Pt in SiC-supported catalysts, each of them was contained in less than 50% of the avilable catalytic materials, though with the exception of Zr in Si_3N_4-supported catalysts and Zr, Mo, Ni in SiC-supported catalysts, each of them occurred in at least 10% of the materials.

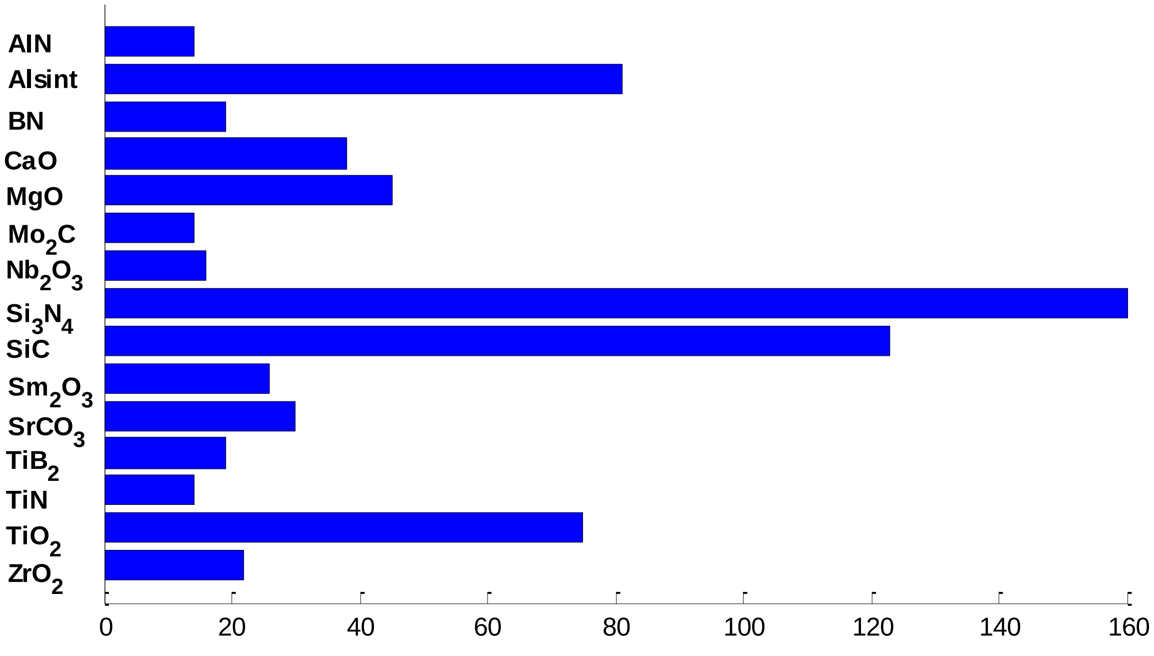

Figure 1. Distribution of the 696 available catalytic materials accordding to their support

Finally in Figure 3, percentiles of HCN yield values are shown for combinations of the supports Si_3N_4 and SiC and the metal additives occurring together with each of those supports in at least 10% of the avilable catalytic materials. As we have seen above, these are the additives Y, La, Re, Ir, Ni, Pt, Zn, Ag, Au.

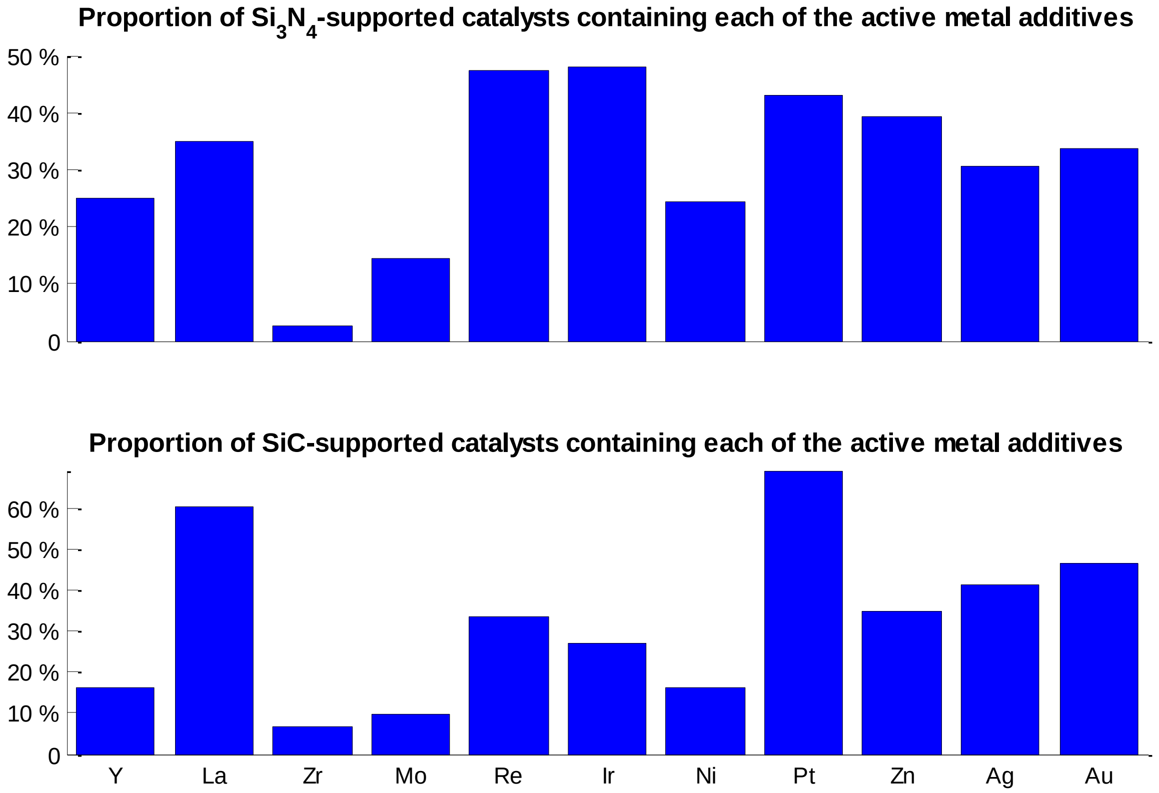

Figure 2. Number of catalysts with supports Si_3N_4 and SiC containing each of the active metal additives

Results of applying the correlation measures. The 6 correlation measures introduced in Section II were used to find the correlations between the molar fractions of the 11 metal additives and the HCN yield for data on the 283 catalytic materials supported by Si_3N_4 or SiC. For each metal additive, the correlation measures were applied only to those catalysts that contained, in addition to the respective support, also that additive. The reason for this restriction is that most properties of correlation measures are valid only if both random variables for which the correlation is calculated are continuous. Whereas it is quite natural to view values of yield as realizations of a continuous random variable, this is not the case for the values of fractions of metal additives if they should include also the value 0 (meaning "that additive is not at all contained in the catalytic material"). Indeed, the above mentioned high number of catalysts not containing the additive would mean equally high number of realizations equal 0, which is extremely unlikely for a continuous variable, even if its values have a very low discernibility. It is much more natural to view the value 0 as the realization of a discrete random variable describing whether the additive is contained in the material or not. Only for the subpopulation of the catalytic materials in which it is contained, there is a continuous random variable providing the fraction of the considered additive in the material. To achieve sufficient reliability, correlation measures have been applied only to those additives that were contained in at least 10 % of the available catalytic materials. These were all additives except Zr (i.e., 10 metal additives) in the case of Si_3N_4-supported catalysts, and all additives except Zr and Mo (i.e., 9 metal additives) in the case of Si_3C-supported catalysts.

The results for catalytic materials with support Si_3N_4 are depicted in Figure 4. They can be sumarized as follows:

a) The fraction of Pt has a positive correlation with the HCN yield according to all measures except medial correlation coefficient. In all such cases, the correlation of its fraction is also the highest positive one. In addition, the fraction of Ir has consistently a low positive correlation with the HCN yield (including medial correlation coefficient), and with the exception of linear correlation coefficient, the same is also true for the fraction of Au. Therefore, an increasing molar fraction of these two elements might improve HCN yield, too. Moreover, the values of the correlation measures, most importantly the values of the Schweizer & Wolff's measure, which reflects the intensity of the correlation betweeen a molar fraction and the HCN yield (cf. Section II) imply that the positive impact of the Pt molar fraction on the HCN yield is higher than that of the Au or Ir molar fraction.

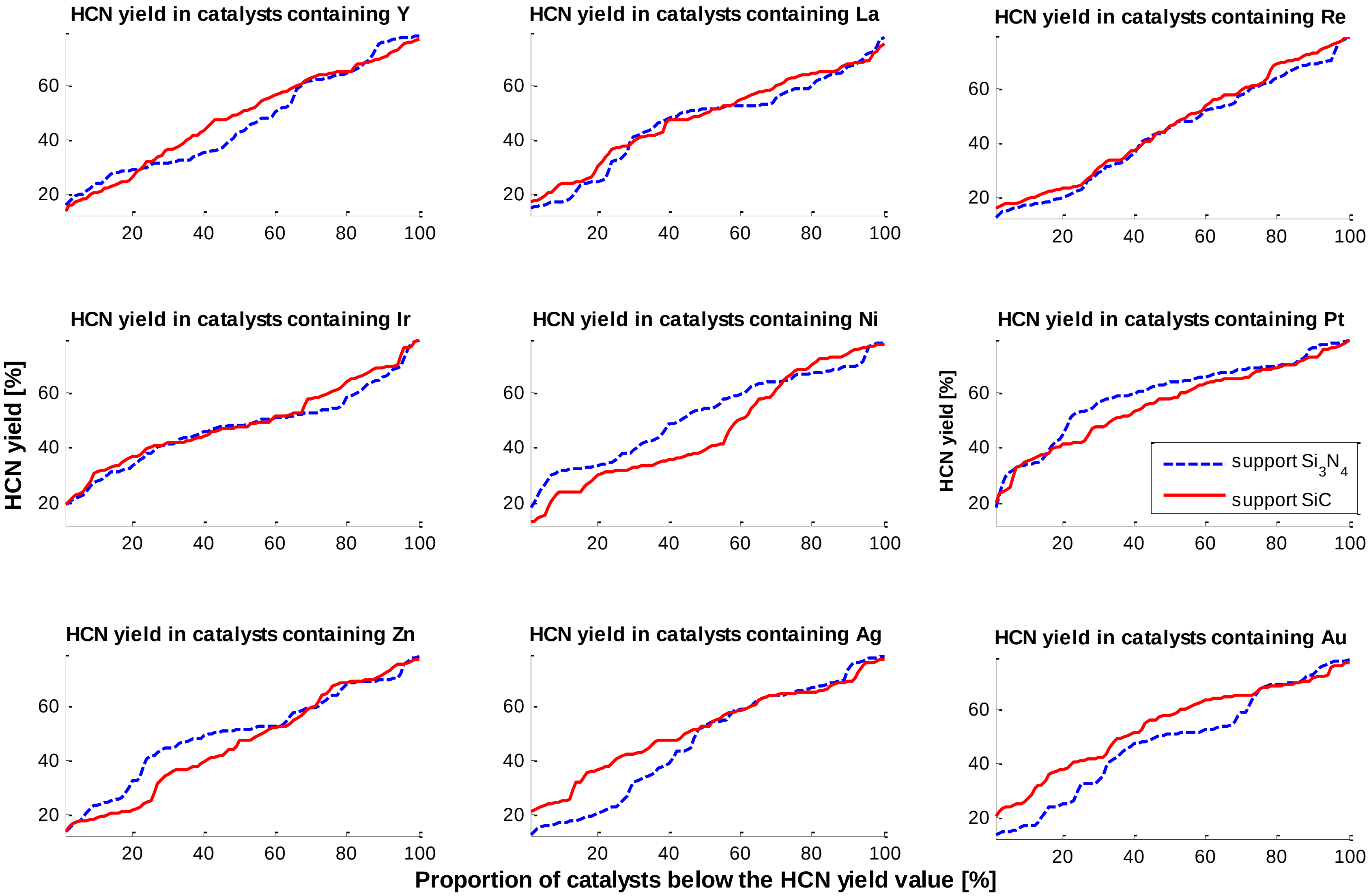

Figure 3. Percentiles of HCN yield values in catalytic materials with supports Si_3N_4 and SiC and metal additives except Zr and Mo

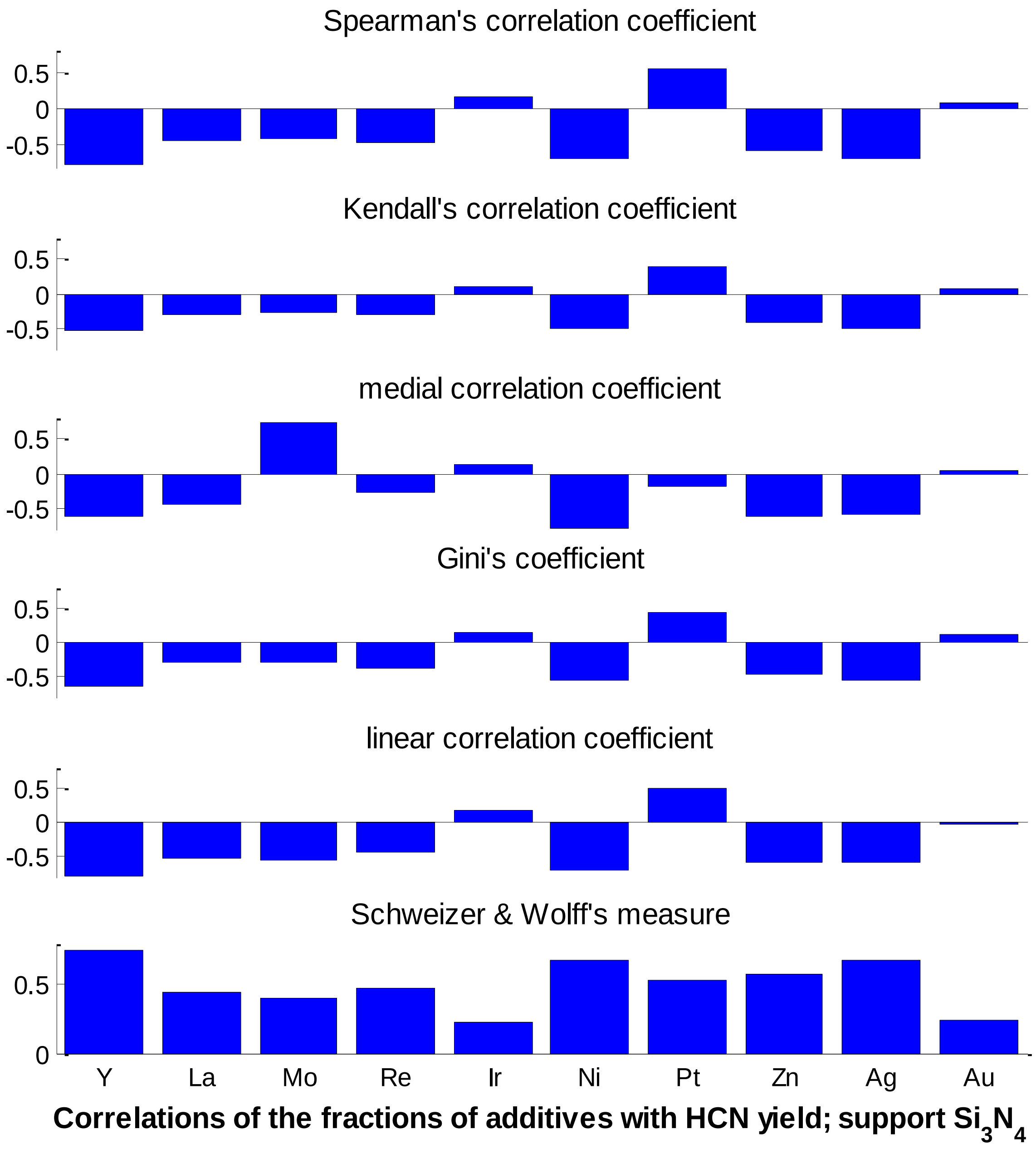

Figure 4. Correlations between fractions of metal additives and HCN yield for catalytic materials supported by Si_3N_4 and the 10 additives contained in at least 20 such materials

b) The most negative correlation with the HCN yield according to the correlations measures indicating the direction of correlation (all measures except the Schweizer & Wolff's, cf. Section II) was obtained for the fractions of elements Y, Ni, Zn and Ag. Moreover, the fractions of Y, Ni, Zn and Ag have also the highest intensity of correlation according to the Schweizer and Wolff's measure. Those facts indicate that catalyst compositions with high molar fraction of these elements should be avoided, whereas small molar fractions of them have a positive influence on the HCN yield. Consequently, Y, Ni, Zn and Ag may act as promoters, which is in agreement with findings in [10].

c) The fractions of La, Mo and Re have less negative correlation with the HCN yield than the fractions of Y, Ni, Zn and Ag, and also the intensity of their correlation with the HCN yield according to the Schweizer and Wolff's measure is lower. Therefore, the negative influence of high molar fraction of these elements on the HCN yield is low, and their fraction in catalysts with a high HCN yield can be larger compared to Y, Ni, Zn and Ag. Indeed, in one of the catalysts with highest yield reported in ref [10], the fraction of Re was 36 %.

Similarly, the results for SiC-supported materials, depicted in Figure 5, can be sumarized as follows:

a) Positive correlation of the molar fraction with the HCN yield for catalysts with support SiC was found for the elements Pt and Au using the Spearman's, Kendall's and Gini's correlation coefficient. As was the case for catalysts with support Si3N4, the linear correlation coefficient is positive, but the medial correlation coefficient is negative for the fraction of Pt. In contrast, both the linear and the medial correlation coefficient are negative for the fraction of Au. The values of the correlations measures indicating the direction of correlation, as well as the intensity of correlation according to the Schweizer and Wolff's measure show that the molar fraction of Pt has a dominating positive influence on the HCN yield. Compared to Pt, the influence of the molar fraction of Au is only marginal.

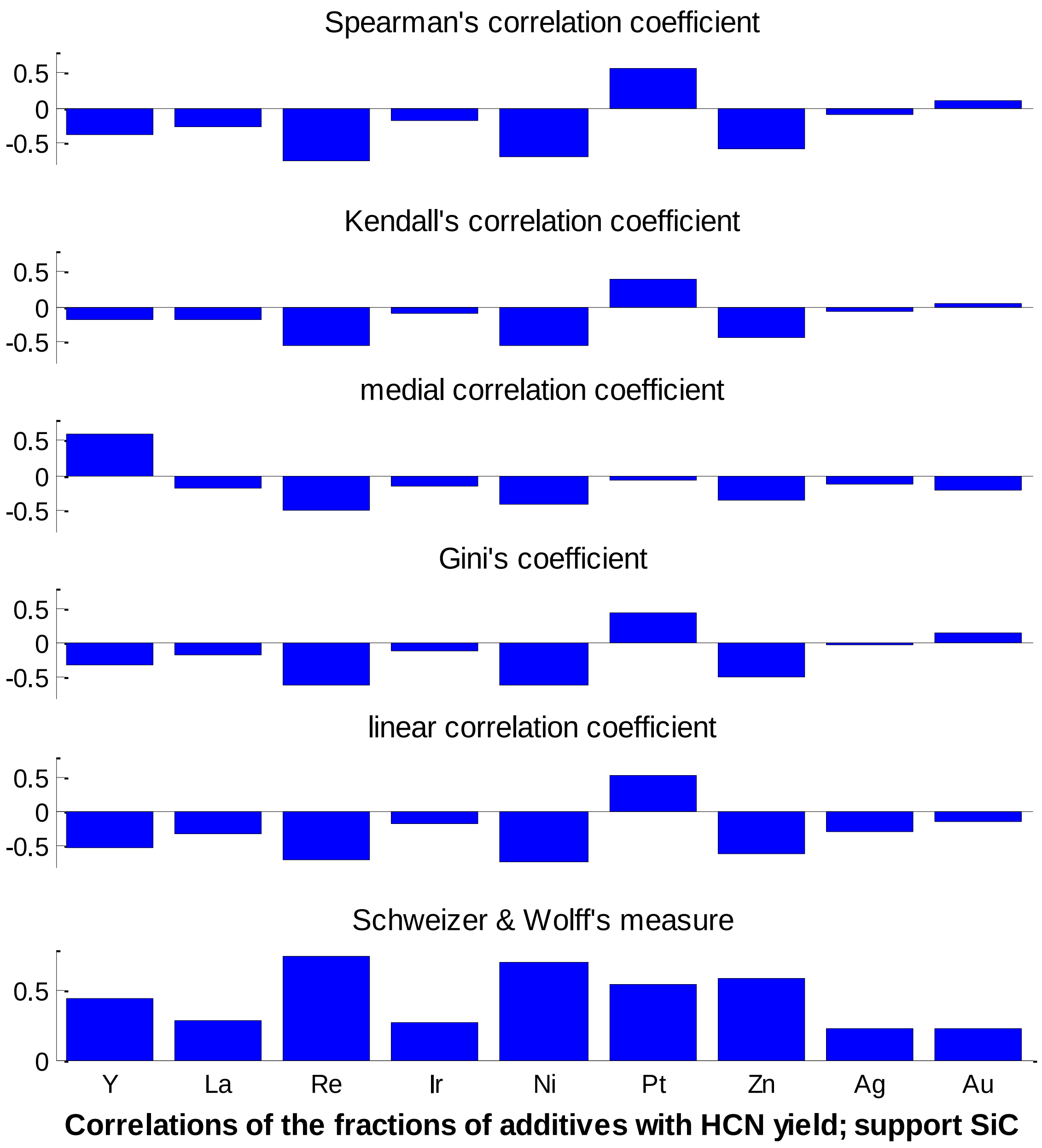

Figure 5. Correlations between fractions of metal additives and HCN yield for catalytic materials supported by SiC and the 9 additives contained in at least 20 such materials

b) The most negative correlation with the HCN yield according to the correlations measures indicating the direction of correlation was obtained for the fractions of Re, Ni ad Zn, for which have also the highest intensity of correlation according to the Schweizer and Wolff's measure. Therefore, high molar fraction of these elements should be avoided. Recall that a strong negative correlation between the fraction of Ni and Zn and the HCN yield was observed already in the case of Si3N4-supported catalysts. It might be concluded that independently of the support, high molar fraction of tose two elements is inapproapriate for achieving high HCN yield. On the other hand, the strong negative correlation between the fractions of Re, Ni, Zn and the HCN yield implies also that the presence of these elements in low molar fractions has a positive influence on the HCN yield.

c) The fractions of La, Ir and Ag have less negative correlation with the HCN yield than the fractions of Re, Ni and Zn, and also the intensity of their correlation with the HCN yield according to the Schweizer and Wolff's measure is lower. Therefore, it is expected that the fraction of La, Ir and Ag in catalysts with a high HCN yield can be larger compared to Re, Ni and Zn.

Comparison with results obtained using the analysis of variance. In [10], analysis of variance of the HCN yield was performed using data about all 696 catalytic materials. For comparison with the application of correlation measures, we now performed it only with data about the 283 materials supported by Si_3N_4 or SiC, separately for each of those supports. Moreover, the choice of independent variables differs from [10] in two respects:

(i) Since support is now the same for all materials in each of the performed analyses of variance, it does not any more serve as an independent variable.

(ii) In the analysis of variance in [10], the metal additives were described with 2-valued independent variables that only indicate the presence of the respective additive in the catalyst. Needless to say, those independent variables substantially differ from the fractions of individual additives in the material, which are used in the calculations of correlation measures. On the other hand, it is not possible to use directly those fractions as independent variables in the analysis of variance because they can have many different values in the data (theoretically as many as the number of considered catalysts) and would impose more parameters to the basic statistical model underlying the analysis than can be estimated from the data (in statistical terms, they would add too many *degrees of freedom* to the model). Therefore, we used auxiliary variables as independent variables describing the metal additives in the analysis of variance, each of which adds at most 5 degrees of freedom to the basic statistical model, at the same time attempting to mimic the behavior of the fraction of the respective metal additive. For a metal additive $\boldsymbol{a}$, such an auxiliary independent variable $\boldsymbol{X_a}$ was defined as follows:

1. If the set of values of the fraction of $\boldsymbol{a}$, $\boldsymbol{F_a}$, in data has at most 5 elements,

   $|\{\boldsymbol{F_a}(\boldsymbol{x}): \boldsymbol{x} \in \boldsymbol{C}\}| \leq 5,$

   where $\boldsymbol{C}$ denotes the set of considered catalysts (thus $\boldsymbol{C}$ has 160 elements if support is Si_3N_4, and 123 elements if support is SiC), then

   $\boldsymbol{X_a}(\boldsymbol{x}) = \boldsymbol{F_a}(\boldsymbol{x})$ for $\boldsymbol{x} \in \boldsymbol{C}$.

2. Oherwise, construct the sets $\boldsymbol{C}_1,\ldots,\boldsymbol{C}_5$ as

   $\boldsymbol{C_1} = \{\boldsymbol{x} \in \boldsymbol{C}: 0 < \boldsymbol{F_a}(\boldsymbol{x}) \leq \boldsymbol{d}_1\}, \ \boldsymbol{C_j} = \{\boldsymbol{x} \in \boldsymbol{C}: \boldsymbol{d_j} - 1 < \boldsymbol{F_a}(\boldsymbol{x}) \leq \boldsymbol{d_j}\}, \boldsymbol{j} = 2, \ldots, 5,$

   where $\boldsymbol{d}_1,\ldots, \boldsymbol{d}_5$ are the even deciles of the set $\{\boldsymbol{F_a}(\boldsymbol{x}): \boldsymbol{x} \in \boldsymbol{C} \ \& \ \boldsymbol{F_a}(\boldsymbol{x}) \neq 0\}$, and then define

   $$\boldsymbol{X_a}(\boldsymbol{x}) = \begin{cases} 0 & \text{if } F_a(x) = 0 \\ \text{mean}\{F_a(x): x \in C_j\} & \text{if } x \in C_j, j = 1, \ldots, 5 \end{cases}$$

The results of the analysis are listed in Table 1. In Si_3N_4-supported catalysts, Pt, Ir and Mo were significant at the 1% level of significance (i.e., highly significant), whereas Ni and Ag were significant only at the 5% level but not at the 1% level. In SiC-supported catalysts, on the other hand, Pt, Au, were significant at the 1% level in SiC-supported catalysts. Observe that for SiC-supported catalysts, the achieved significances are cleraly lower than for catalysts with support Si_3N_4. Therefore, it can be assumed that the interaction of SiC with Pt and Au is lower than the interaction of Si_3N_4 with Pt and Ir.

Table 1. Results of the analysis of variance of the HCN yield on the presence of the 11 considered metal additives for the data about the 283 materials supported by Si_3N_4 or SiC (highly significant main effects: ■, significant main effects: ░)

Dependent variable	Main factor	Achieved significance	
		support Si_3N_4	support SiC
HCN yield	Y	0.16	0.32
	La	0.32	0.65
	Zr	0.95	0.83
	Mo	$3.6\cdot10^{-4}$	0.83
	Re	0.35	0.073
	Ir	$4.7\cdot10^{-7}$	0.1
	Ni	0.016	0.37
	Pt	$2.0\cdot10^{-12}$	$4.5\ \cdot10^{-5}$
	Zn	0.058	0.18
	Ag	0.042	0.65
	Au	0.19	$5.0\cdot10^{-3}$

There is one clear correspondence between the results of analysis of variance and results obtained with correlation methods: The fraction of Pt, which is in combination with both considered supports the most significant metal additive at all (achieved significance 10^{-12} in Si_3N_4-supported catalysts, 10^{-5} in SiC-supported catalysts), has also irrespectively of support the highest positive correlation with the HCN yield, according to all measures except medial correlation coefficient.

On the other hand, this is also the only unquestionable correspondence. Whereas the fractions of the metal additives Ni and Ag, which are significant in Si_3N_4-supported catalysts, have in those catalysts a high intensity of correlation according to Schweizer and Wolff's measure, the same is not true for the fractions of the highly significant additives Mo and Ir in Si_3N_4-supported catalysts, neither for the fractions of the highly significant Au in SiC-supported catalysts. At the same time, the fraction of Ni the has a high intensity of correlation also in SiC-supported catalysts, although Ni is not at all significant in them (achieved significance 0.37).

Sometimes, a new insight into the influence of catalyst composition on the HCN yield can be obtained through combining results obtained with the correlation measures with the results of analysis of variance. For example, the very high sigificance of Ir in Si_3N_4-supported catalysts, combined with the fact that the fraction of Ir has a low positive correlation with the HCN yield in that case, indicates that for most of the possible values of the above defined auxiliary variable $\boldsymbol{X_{Ir}}$, the variance of the HCN yield is quite low. Consequently, the influence of the fractions of the remaining elements is mostly low once the fraction of Ir is fixed. For similar reasons, the variance of the HCN yield is quite low for most of the possible values of the above defined auxiliary variable $\boldsymbol{X_{Au}}$ in the case of SiC-supported catalysts. In such catalysts the influence of the fractions of the remaining elements is thus mostly low once the fraction of Au is fixed.

Comparison with results obtained using regression trees. Also a regression tree with the HCN yield as dependent variable was for the HCN data constructed already in [10], taking into account all 696 catalytic materials. For comparison with the application of correlation measures, we now constructed regression trees using only data about the 283 materials to which they were applied, a separate tree for each support. We employed the Matlab implementation [49] of the original regression trees proposed in [50], putting a single restriction on the resulting trees, namely the minimal size of data in leafs to be 10.

The constructed regression tree for Si_3N_4-supported catalytic materials is visualized in Figure 6. Observe that the primary splits are according to the fraction of Pt, and the secondary splits for catalysts with lower fractions of Pt according to the fraction of Ir, whereas for those with higher fractions of Pt once again according to the fraction of Pt. The tertiary splits are according to the fractions of Mo, Zn, once again Ir, and Au. The regression tree shows that on the Si_3N_4 support, especially the combination of Pt with the elements Au and Pt can give high HCN yield. Furthemore, that fact that Pt determines the primary and one secondary split, and that Ir determines one secondary and one tertiary split, imply a great impact of those two elements on HCN formation, a conclusion suppported also by high significance of those elements according to the analysis of variance.

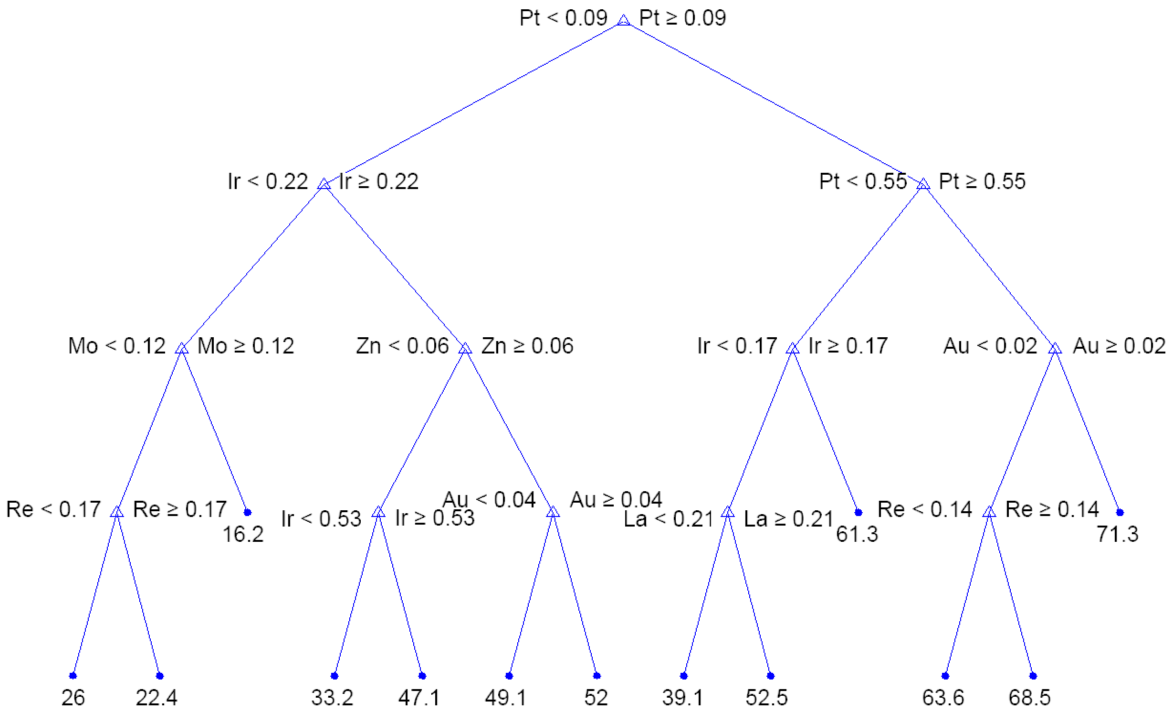

Figure 6. Regression tree for HCN yield in catalysts with support Si_3N_4

The regression tree for SiC-supported catalysts is visualized in Figure 7. Its apparently smaller size (i.e., lower number of nodes) compared to the tree for Si_3N_4-supported catalysts is a consequence of the different number of catalysts with both kinds of support in the available data. Also for SiC-supported catalysts, the primary splits are according to the fraction of Pt, and the secondary splits according to the fraction of Ir for catalysts with lower fractions of Pt, and again according to the fraction of Pt for those with higher fractions of Pt. The tertiary splits are, for this support, only according to fractions of two elements: Zn and Au. The regression tree shows that also for the SiC support, the addition of further elements, here in particular Ir and Au, can lead to high HCN yield.

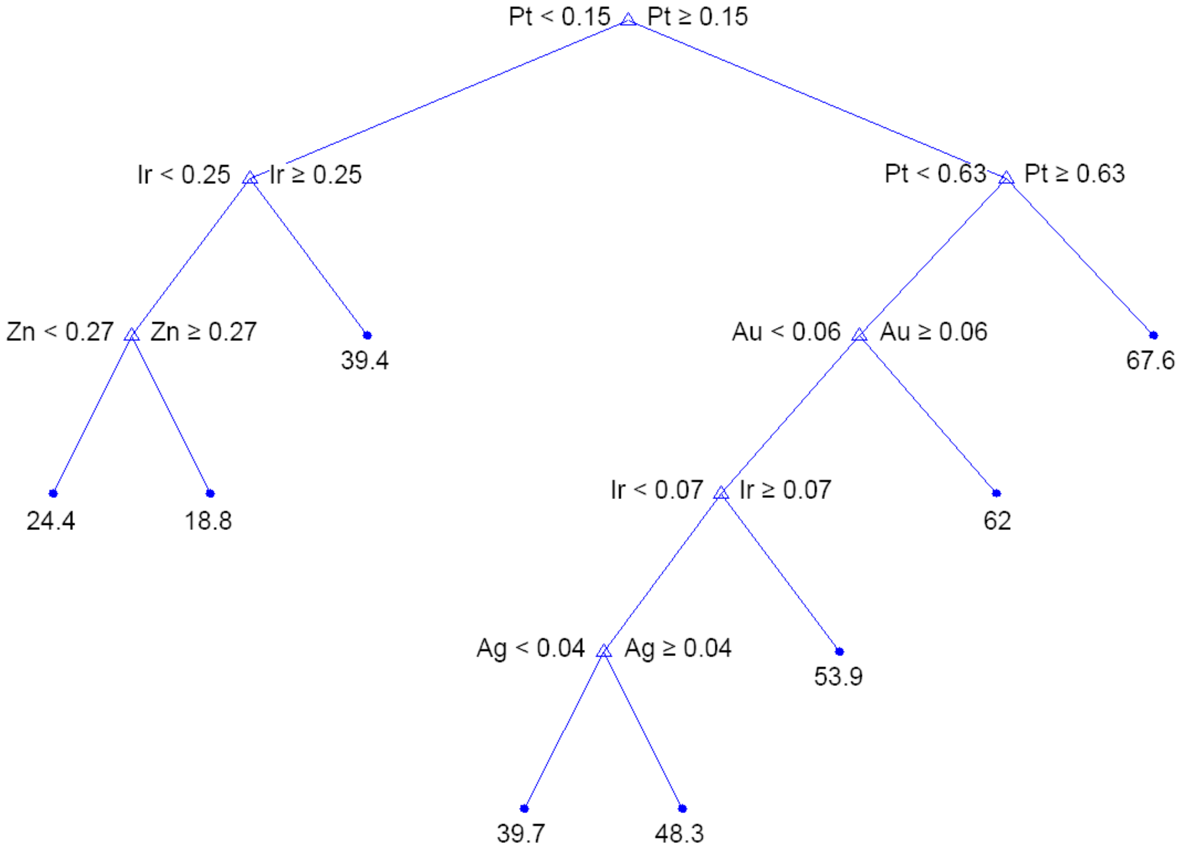

Figure 7. Regression tree for HCN yield in catalysts with support SiC

Also for regression trees, like for the analysis of variance, a clear correspondence with the results obtained with correlation methods can be found only in emphasizing the importance of the fraction of Pt. That fraction determines not only the primary splits of HCN yield and of the conversion of CH_4, but also the secondary splits in the branches corresponding to higher values. In Si_3N_4-supported catalysts, two more fractions determining splits at the first three levels of the tree, the fractions of Ir and Au, tend to have a positive correlation with the HCN yield. The fraction of Ir is actually the only one for which that correlation is consistently positive. In SiC-supported catalysts, however, this is not true, although the fractions of Ir and Au determine for them splits on the same layers. Even more importantly, the proportion of Zn, which also determines splits on the third layer, has on the contrary a consistently negative correlation with the HCN yield.

V.CONCLUSION

In the paper, the methodology of correlating continuous descriptors of catalytic materials (most typically, the fractions of individual components) with their catalytic performance was addressed. We attempted to draw reader's attention to the existence of various correlation measures, and to the importance of complementing the commonly used QSAR and similar models with the application of such measures. These measures are particularly relevant if the final goal of our interest in correlation between the composition of catalytic materials and their performance is a decision which elements should be included in a pool considered for the design of new catalysts (typically using some combinatorial design method[51]). For such a decision, it is completely sufficient to know the fraction of which elements are positively correlated and the fracitons of which are negatively correlated with catalyst performance. A QSAR model is actually not needed, nor is the large amount of data that reliable modelling with them requires. Finally, the paper also pointed out the similarities and dissimilarities between the results obtained with correlation measures, and those obtained with two other approaches, based on different principles: the analysis of variance and regression trees.

The presented case study using data from high-temperature synthesis of HCN[10] has shown that correlation measures can help to increase our insight into to the properties and behaviour of the studied catalytic materials. The metal additives that were, in this case study, found most important for HCN yield by the correlation measues and by the two other approaches, are once more summarized in Table 2 below. They are in agreement with existing knowledge about HCN synthesis. The catalytic components should be suited for breaking hydrogen-carbon bonds of methane as well as hydrogen-nitrogen bonds from ammonia. The remaining CH_x and NH_x should then couple on the surface resulting finally in HCN. M. Diefenbach et al.[52] have shown in experimental and computational studies that C-N bond formation is mediated by Pt^+. CH_4 is first dehydrogenated to $PtAuCH2^+$ while dehydrogenation of ammonia does not occur on Pt due to its endothermicity. The major pathway is found via PtH and $CH_2NH_2^+$. Compared to other transition metal cations like Fe*, Co^+, Rh^+, W^+, Os^+, Ir^+ and Au^+, Pt^+ is assumed to be unique for its ability to activate methane and to mediate C-N bond formation as a precursor for HCN. According to the present results it is likely that also Ir, Au, and Re as well as possibly Mo facilitate HCN formation in a similar manner. In a later paper, K. Koszinowski et al.[53] argued that CN coupling of methane and ammonia might ooccur on $Pt_mAu_n^+$ clusters; however, in further work they proved that only the dinuclear carbene complex $PtAuCH_2^+$ mediates C-N formation. Au itself was shown not to form HCN. This is not in contradiction to the present work, in which Pt and a support was present in the catalytic reaction. Tentatively, it may be assumed that also Ir and Re follow similar mechanisms for the mechanisms of the HCN-formation reaction.

Table 2. Metal additives most important for HCN yield according to correlation measures, analysis of variance, and regression trees.

support	fraction of the additive has a strong positive or negative correlation with HCN yield	presence of the additive is significant according to the analysis of variance of HCN yield	fraction of the additive determines splits at the three highest levels of the regression tree for HCN yield
Si_3N_4	Pt, Y, Ni, Ag	Pt, Ir, Mo, Ni, Ag	Pt, Ir, Mo, Zn, Au
SiC	Pt, Re, Ni, Zn	Pt, Au	Pt, Ir, Zn, Au

The case study also showed not only what results can be obtained with individual measures and approaches, but also how the differences between results obtained with each of them can be explained by means of the differences in their principles and properties.

REFERENCES

1. Buyevskaya, O.V.; Bruckner, A.; Kondratenko, E.V.; Wolf, D.; Baerns, M. Fundamental and combinatorial approaches in the search for and optimization of catalytic materials for the oxidative dehydrogenation of propane to propene, *Catal. Today* **2001**, *67*, 369–378.

2. Ehrich, H.; Berndt, H.; Pohl, M.M.; Jähnisch, K.; Baerns, M. Oxidation of benzene to phenol on supported Pt-Vo_x and Pd-Vo_x catalysts. *Appl. Catal., A: General* **2002**, *230*, 271–280.

3. Holeňa, M.; Baerns, M. Feedforward neural networks in catalysis: A tool for the approximation of the dependence of yield on catalyst composition, and for knowledge extraction. *Catal. Today* **2003**, *81*, 485–494.

4. Farrusseng, D.; Klanner, C.; Baumes, L.; Lengliz, M.; Mirodatos, C.; Schüth, F. Design of discovery libraries for solids based on QSAR models. *QSAR Comb. Sci.* **2005**, *24*, 78–93.

5. Omata, K.; Kobayashi, Y.; Yamada, M. Artificial neural network-aided development of supported Co catalyst for preferential oxidation of CO in excess hydrogen. *Catal. Commun.* **2005**, *6*, 563–567.

6. Du, G.; Yang,Y.; Qiu, W; Lim, S; Pfefferle, L.; Haller, G.L. Statistical design and modeling of the process of methane partial oxidation using V-MCM-41 catalysts and the prediction of the formaldehyde production. *Appl. Catal., A: General* **2006**, *313*, 1–13.

7. Čukić, T.; Kraehnert, R.; Holeňa, M.; Herein, D.; Linke, D.; Dingerdissen, U. The influence of preparation variables on the performance of Pd/Al_2O_3 catalyst in the hydrogenation of 1,3-butadiene: Building a basis for reproducible catalyst synthesis. *Appl. Catal., A: General* **2007**, *323*, 25–37.

8. Gobin, O.C.; Martinez, J.A.; Schüth, F. Multi-objective optimization in catalytical chemistry applied to the selective catalytic reduction of NO with C_3H_6. *J. Catal.,* **2007**, *252*, 205–214.

9. Sieg, S.C. *Modelling Quantitative Composition Activity Relationships (QCARs) for Heterogeneous Catalysts by Kriging and a Multilevel B-Spline Approach.* Thesis. Universität des Saarlandes, 2007.

10. Moehmel, S.; Steinfeldt, N.; Engelschalt, S.; Holeňa, M.; Kolf, S.; Baerns, M.; Dingerdissen, U.; Wolf, D.; Weber, R.; Bewersdorf, M. New catalytic materials for the high-temperature synthesis of hydrocyanic acid from methane and ammonia by high-throughput approach. *Appl. Catal., A: General* **2008**, *334*, 73–83.

11. Suh, C.; Sieg, S.C.; Heying, M.J.; Oliver, J.H.; Maier, W.F.; Rajan, K. Visualization of High-Dimensional Combinatorial Catalysis Data. *J. Comb. Chem.* **2009**, *11*, 385–392.

12. Valero, S. ; Argente, E. ; Botti, V. ; Serra, J.M. ; Serna, P. ; Moliner, M.; Corma. A. DoE framework for catalyst development based on soft computing techniques. *Comput. Chem. Eng.* **2009**, *33*, 225–238.

13. Klanner, C.; Farrusseng, D.; Baumes, L.; Lenliz, M.; Mirodatos, C.; Schüth, F. The development of descriptors for solids: Teaching "catalytic intuition" to a computer. *Angew. Chem. Int. Ed.* **2004**, *43*, 5347-5349.

14. Corma, A.; Serra, J.M.; Serna, P.; Moliner, M. Integrating high-throughput characterization into combinatorial heterogeneous catalysis: unsupervised construction of quantitative structure/property relationship models. *J. Catal.* **2005**, *232*, 335-341.

15. Sieg, S.; Stutz, B.; Schmidt, T.; Hamprecht, F.; Maier, W.F. A QCAR-approach to materials modeling. *J. Mol. Model.* **2006**, *12*, 611-619.

16. Serra, J.M.; Baumes, L.A. ; Moliner, M. ; Serna, P.; Corma, A. Zeolite synthesis modelling with support vector machines: a combinatorial approach. *Comb. Chem. High Throughput Screening* **2007**, *10*, 13–24.

17. Rothenberg, G. Data mining in catalysis: Separating knowledge from garbage. *Catal. Today* **2008**, *137*, 2–10.

18. Serna, P.; Baumes, L.A..; Moliner, M.; Corma A. Combining high-throughput experimentation, advanced data modeling and fundamental knowledge to develop catalysts for the epoxidation of large olefins and fatty esters. *J. Catal.* **2008**, *258*, 25–34.

19. Farrusseng, D.; Clerc, F.; Mirodatos, C.; Rokotomalala, R. Virtual screening of materials using neuro-genetic approach: Concepts and implementation. *Comput. Mater. Sci.*, **2009**, *45*, 52-59.

20. Faghihi, E.M. and A.H. Shamekhi, *Development of a neural network model for selective catalytic reduction (SCR) catalytic converter and ammonia dosing optimization using multi objective genetic algorithm.* Chem. Eng. J. (Amsterdam, Neth.), 2010. **165**(2): p. 508-516.

21. Nandi, S., et al., *Hybrid process modeling and optimization strategies integrating neural networks/support vector regression and genetic algorithms: study of benzene isopropylation on Hbeta catalyst.* Chem. Eng. J. (Amsterdam, Neth.), 2004. **97**(2-3): p. 115-129.

22. Kito, S.; Hattori, T.; Murakami, Y. Estimation of catalytic performance by neural network – product distribution in oxidative dehydrogenation of ethylbenzene. *Appl. Catal., A: General* **1994**, *114*, L173–L178.

23. Hou, Z.Y.; Dai, Q.; Wu, X.Q.; Chen, G.T. Artificial neural network-aided design of catalyst for propane ammoxidation. *Appl. Catal., A: General* **1997**, *161*, 183–190.

24. Huang, K.; Feng-Qiu, C.; Lü, D.W. Artificial neural network-aided design of a multi-component catalyst for methane oxidative coupling. *Appl. Catal., A: General* **2001**, *219*, 61–68.

25. Corma, A.; Serra, J.M.; Argente, E.; Botti, V.; Valero, S. Application of artificial neural networks to combinatorial catalysis: modeling and predicting ODHE catalysts. *ChemPhysChem* **2002**, *3*, 939–945.

26. Tompos, A; Margitfalvi, J.L.; Tfirst, E.; Végvári, L. Information mining using artificial neural networks and "holographic research strategy". *Appl. Catal., A: General* **2003**, *254*, 161–168.

27. Tompos, A.; Margitfalvi, J.L.; Tfirst, E.; Végvári, L. Evaluation of catalyst library optimization algorithms: Comparison of the Holographic Research Strategy and the Genetic Algorithm in virtual catalytic experiments. *Appl. Catal., A: General* **2006**, *303*, 72–80.

28. Günay, M.E.; Yildirim, R. Neural network aided deisgn of Pt-Co-Ce/Al_2O_3 catalyst for selective CO oxidation in hydrogen-rich streams. *Chem. Eng. J.* **2008**, *140*, 324–331.

29. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; Wiley: New York, 1973.

30. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*; Holden-Day: San Francisco, 1975.

31. Maritz, J.S. *Distribution-Free Statistical Methods*; Chapman & Hall: London, 1981.

32. Schweizer, B; Wolff, E.F. On nonparametric measures of dependence for random variables. *Ann. Statist.* **1981**, *9*, 879–885.

33. Myers, J.L.; Well, A.D. *Research Design and Statistical Analysis*; Lawrence Erlbaum: New Jersey, 2003.

34. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, 2006.

35. Reynolds, H.T. *Analysis of Nominal Data*; Sage: Thousand Oaks, 1977.

36. Scheffé, H. *The Analysis of Variance*; Wiley: New York, 1999.

37. Sahai, H.; Ageel, M.I. *Analysis of Variance: Fixed, Random and Mixed Models*; Birkhäuser: Boston, 2000.

38. Agresti, A. *Categorical Data Analyis*, 2. Edition; Wiley: New York, 2002.

39. Block, H.W.; Ting, M.L. Some concepts of multivariate dependence. *Comm. Statist. A – Theory Methods* **1981**, *10*, 749–762.

40. Wolff, E.F. N-dimensional measures of dependence. *Stochastica* **1981**, *4*, 175–188.

41. Joe, H. Multivariate concordance. *J. Multivariate Anal.* **1990**, *35*, 12–30.

42. Joe, H. *Mutivariate Models and Dependence Concepts*; Chapman & Hall: London, 1997.

43. Scarsini, M. On measures of concordance. *Stochastica* **1984**, *8*, 201–218.

44. Wolf, D.; Buyevskaya, O.V.; Baerns, M. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Appl. Catal., A: General* **2000**, *200*, 63–77.

45. Rodemerck, U.; D. Wolf, D.; Buyevskaya, O.V.; Claus, P.; Senkan, S.; Baerns, M. High-throughput synthesis and screening of catalytic materials: Case study on the search for a low-temperature catalyst for the oxidation of low-concentration propane. *Chem. Eng. J.* **2001**, *82*, 3–11.

46. Cox, D.R.; Hinkley, D.V. *Theoretical Statistics*, Chapman & Hall: London, 1974.

47. Kendall, M.G.; Stuart, A. *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*, Griffin: London, 1979.

48. Holeňa, M.; Baerns, M. Computer-Aided Strategies for Catalyst Development. In: *Handbook of Heterogeneous Catalysis*, Ertl, G. et al., Eds.; Wiley-WCH: Weinheim, Germany, 2008, pp 66–81.

49. *Statistics Toolbox 7.3*, The MathWorks, Inc.: Natick, 2010.

50. Breiman, L.; Friedman, J.H.; Olshen, R.A; Stone, C.J. *Classification and Regression Trees*, Wadsworth: Belmont, 1984.

51. Baerns, M.; Holeňa, M. Combinatorial Development of Solid Catalytic Materials. World Scietific – Imperial College Press: London, 2009.

52. Diefenbach, M.; Brönstrup, M.; Aschi, M.; Schröder, D.; Schwarz, H. HCN Synthesis from Methane and Ammonia: Mechanisms of Pt+-Mediated C–N Coupling. *J. Am. Chem. Soc.* **1999**, *121*, 10614–10625.

53. Koszinowski, K.; Schröder, D.; Schwarz, H. C–N Coupling of Methane and Ammonia by Bimetallic Platinum–Gold Cluster Cations, *Organometallics* **2004**, *23*, 1132–1139.

FOR TABLE OF CONTENTS USE ONLY:

Measuring the Correlation of Catalyst Properties with Its Catalytic Performance

Martin Holena, Norbert Steinfeldt