ELSEVIER

Chemical Engineering Journal 141 (2008) 264–277

Chemical Engineering Journal

www.elsevier.com/locate/cej

Experimental design for the identification of hybrid reaction models from transient data

Marc Brendel [1], Wolfgang Marquardt *

Lehrstuhl für Prozesstechnik, RWTH Aachen University, 52064 Aachen, Germany

Received 31 July 2007; received in revised form 17 December 2007; accepted 26 December 2007

Abstract

Model identification from dynamic experimental data may involve the conduction of multiple experiments to identify a suitable model with adequate accuracy. In contrast to the *a priori* specification of design sets, an iteratively conducted model-based experimental design exploiting previous data promises significantly better results with lower effort. Yet, in dynamic systems, the inputs of the model to be identified often cannot be designed directly, but depend on the experimental degrees of freedom. To allow for model-based design for the identification of dynamic hybrid or fully unstructured models, a new design criterion is developed in this work. It is based on an input-space coverage approach and allows to simultaneously design the experiment for multiple models to be identified. Incremental identification is applied to efficiently construct the unknown models from data. The resulting iterative design and identification methodology is illustrated on a reaction kinetic model identification for the acetoacetylation of pyrrole in a simulation study.

Keywords: Experimental design; Dynamic model; Incremental identification; Hybrid modeling; Chemical reactor; Reaction network; Data-driven kinetics

1. Introduction

Mathematical models of chemical reaction processes are readily needed for a multitude of tasks including process design, analysis and optimization of process conditions [1] as well as model-based control, see e.g. [2]. Often, a reliable model of the process is unknown and needs to be identified from experimental data. Depending on the desired application and on the available process knowledge, such process models may be either *structured* (i.e. derived from physical knowledge), *unstructured* (e.g. fully data-driven), or *hybrid*, i.e. combining both physically motivated and data-driven model parts. The latter model building strategy has been shown to yield higher prediction accuracies compared to purely data-driven models [3]. It is a favorable option for the identification of unknown reaction kinetics from experimental data, if no structured model candidates for the description of the reaction kinetics are at hand. The algebraic, data-driven models of the unknown reaction kinetics are then identified from dynamic experimental data in a hybrid model structure consisting of the mole balances, reaction stoichiometries and the reaction kinetics, see e.g. [4–7]. Yet, the results of the identification process may prove unsatisfactory to meet the needs required when using data from a single experiment only. Then, several experiments need to be planned and realized to obtain dynamic data with high information content for the identification of the data-driven kinetics.

A number of design techniques exist for the identification of structured models from dynamic data to discriminate between model candidates and to estimate model parameters with high accuracy, see e.g. [8–10]. Yet, these design approaches for structured model identification are not applicable to hybrid or fully unstructured models due to the often vast number of non-uniquely identifiable parameters in data-driven model structures and the modeler's interest in prediction accuracy rather than parameter accuracy. No model-based design approach is known to the authors for the identification of data-driven parts in hybrid or fully unstructured models. A reason for the lack of appropriate design techniques to be embedded in an iterative design and identification process [11] may also be seen

Abbreviations: COV, Coverage design; FFC, Fractional factorial design; MIMO, Multiple input–multiple output; MISO, Multiple input–single output; RND, Random design.

* Corresponding author. Tel.: +49 241 80 94668; fax: +49 241 80 92326.
E-mail address: Wolfgang.Marquardt@avt.rwth-aachen.de (W. Marquardt).
[1] Present address: Evonik Degussa GmbH, Process Technology, 63457 Hanau, Germany.

1385-8947/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cej.2007.12.027

Nomenclature

Roman symbols

c, $\mathbf{c}$	Concentration (scalar, vector)
d	Space dimensionality
E	Expected distance between data
f	Generic function
f^{r}, $\mathbf{f}^{\mathrm{r}}$, $\mathbf{F}^{\mathrm{r}}$	Reaction flux (scalar, vector, matrix)
H	Heaviside step function
k	Rate constant
$\mathbf{m}$	Generic model structure
n_{C}, n_{D}, n_{F}	Number of experimental designs
n_{E}	Number of experiments
n_{I}	Number of inputs
n_{O}	Number of outputs
n_{Q}	Number of samples
$\mathbf{N}$	Stoichiometric matrix
p	Probability distribution
q^{in}, q^{out}	Volumetric flow rates into and out of reactor
r, $\mathbf{r}$, $\mathbf{R}$	Reaction rate (scalar, vector, matrix)
t	Time
t_{f}	Batch time
t_{s}	Sampling interval
v, $\mathbf{v}$, $\mathbf{V}$	Volume (scalar, vector, matrix)
x, $\mathbf{x}$	Input data (scalar, vector/matrix)
$(\mathbf{x})_q$	The q th row vector of matrix $\mathbf{x}$
y, $\mathbf{y}$	Output data (scalar, vector)

Greek symbols

α_σ	Relative error level
Γ	Prediction ratio
δ^{cov}	Overall distance criterion (single output)
$\breve{\delta}^{\mathrm{cov}}$	Overall distance criterion (multiple outputs)
δ^{sst}	Distance criterion with respect to new data set
δ^{tot}	Distance criterion with respect to available data set
ϵ^{c}	Relative prediction error for concentrations
ϵ^{r}	Relative prediction error for reaction rates
$\boldsymbol{\theta}$	Vector of model parameters
ϑ	Distance measure
Λ	Prediction convergence ratio
ξ, $\boldsymbol{\xi}$	Scaled input (scalar, vector/matrix)
$(\boldsymbol{\xi})_q$	The q th row vector of matrix $\boldsymbol{\xi}$
σ	Standard deviation
τ	Dimensional scaling factor
$\boldsymbol{\varphi}$	Set of experimental conditions
$\boldsymbol{\varphi}^\star$	Optimal set of experimental conditions
ω, $\boldsymbol{\omega}$	Random point (scalar, vector)

Calligraphic symbols

$\mathcal{D}$	Data set
$\mathcal{S}$	Set of chemical species
$\mathcal{X}$	Input set

Subscripts, superscripts and accents

$(\cdot)_0$	Initial value
$(\cdot)^{(k)}$	Referring to k th output
$(\cdot)^{\mathrm{cov}}$	Referring to coverage design
$(\cdot)^{\mathrm{in}}$	Referring to feed
$(\cdot)^{\mathrm{max}}$	Maximum
$(\cdot)^{\mathrm{min}}$	Minimum
$(\cdot)^{\mathrm{pred}}$	Predicted (after identification process)
$(\cdot)^{\mathrm{rnd}}$	Referring to random design
$(\cdot)^{\mathrm{tot}}$	Referring to total set
$(\cdot)^{\mathrm{true}}$	True (simulated) quantity
$(\hat{\cdot})$	Estimated quantity
$(\bar{\cdot})$	Average
$(\cdot)^{\mathrm{T}}$	Matrix transpose

Mathematical notation

$\mathbb{R}$	Set of real numbers
$\mathbb{X}$	Data domain
∞	Infinity
$\in$	Element of
$\subseteq$	Subset
$\subset$	Proper subset
$\bigcup$	Generalized union

Chemical species

D	Diketene
DHA	Dehydroacetic acid
G	By-product
K	Pyridine
OL	Oligomers
P	Pyrrole
PAA	2-acetoacetyl pyrrole

in the complexity to flexibly identify generalizable models from dynamic data with conventional identification approaches [12].

Strategies to design experiments for the identification of data-driven models are reported in the data mining literature. The field of *active learning* (also *query learning*) is tightly connected with the theory of *design of experiments* (e.g. [8]) applied to establish structured models. Here, active learning approaches are widely recognized to make the learning process more efficient by actively selecting particularly salient data points $\mathbf{x}$ taken to generate the data $\mathrm{D} = (\mathbf{x}, \mathbf{y}(\mathbf{x}))$. For constructing unstructured data models, a large number of active learning strategies exist: On the one hand, all data points to be sampled can be designed beforehand. Inter alia, this concept is realized in the construction of random designs, (fractional) factorial designs [13] or space-filling designs (including latin hypercube sampling [14] and its variants [15]). Space-filling designs should be used when there is little or no information about the underlying effects of factors on responses. The aim is to spread the points as evenly as possible around the space of n_{I} feasible model inputs. The designs fill the n_{I}-dimensional space with points that are in some way regularly spaced. They are reported to be especially useful in conjunction with unstructured models [15]. Space filling designs depend

neither on the responses nor on the model used to approximate them.

Alternatively, new data can be selected adaptively in each step, depending on the set of already obtained data and/or their approximation. This is referred to as *sequential design*. Consider the following problem: Given a training set $D = \{(\mathbf{x}_q, y_q) \in \mathbb{X} \times \mathbb{R}, q = 1, \ldots, n_Q\}$, $\mathbb{X} \subset \mathbb{R}^{n_I}$ and a trained model, find a new input vector (the query) $\mathbf{x}_{n_Q+1} \in \mathbb{X}$ such that the expected information gain is maximal if $(\mathbf{x}_{n_Q+1}, y_{n_Q+1})$ is added to the training set. Here, y_{n_Q+1} denotes the result of measuring at $\mathbf{x}_{n_Q+1} \in \mathbb{X}$ and $\mathbb{X} \subset \mathbb{R}^{n_I}$ is the space of feasible input vectors. Then, the next query can be obtained by optimizing some query criterion on the whole space $\mathbb{X}$. In the query algorithms proposed in literature, the new data points are either chosen according to some heuristic (e.g. [16]) or by optimizing some objective function such as maximizing the expected information gain [17] or minimizing model uncertainty [18]. Cohn [18] and Cohn et al. [19] show that the expected generalization performance of active learning is significantly better than that of passive learning. Zhang [20] and Seung et al. [17] report similar improvements. A comprehensive review on active learning approaches has been presented by Hasenjäger and Ritter [21]. Most of the approaches are restricted to the case where only one new point at a time is queried, but the methods can be generalized to the more complex task of selecting multiple new data points.

Yet, in the identification of unknown models $y = f(\mathbf{x})$ from dynamic process data, model inputs $\mathbf{x}(t)$ (and outputs $\mathbf{y}(t)$) often cannot be chosen independently, but both depend on the experimental degrees of freedom $\boldsymbol{\varphi}$. Exemplarily, in the identification of reaction kinetic laws considered in this work, the individual concentration measurements over time – serving as inputs to the reaction kinetic model – depend on experimental settings such as initial concentrations, reactor design, feed rate and composition and, of course, the reaction itself that is unknown and needs to be modeled. Hence, an appropriate choice of these experimental degrees of freedom is required in order to guarantee reliable estimation of the unknown functional relations.

To allow for an efficient, model-based design of new experiments, for the identification of data-driven model parts in hybrid or fully unstructured models, a new design methodology is proposed in this work. A criterion is developed to fill the multi-dimensional space of inputs to the unknown functional relation(s) by appropriately selecting the experimental degrees of freedom while taking into account the system dynamics. The approach allows to simultaneously design the new experiment for subsequent identification of multiple unknown models. With a process model contrariwise forming the basis of any model-based experiment design technique, an iterative design and model identification procedure results.

To construct appropriate models from available experimental data in each step, an incremental identification approach [22] is used. The approach naturally exploits the hierarchical structure inherent to any process model. Known information (such as the dynamic mole balances) is incorporated to reduce the identification process to modeling uncertainties, i.e. the unknown reaction kinetic laws to be identified by purely algebraic regression. The approach is used for its computational efficiency and its flexibility to apply any (e.g. data-driven) model structure to construct the unknown reaction kinetics. In particular, training algorithms with inherent regularization (see e.g. [23]) can be directly employed to ensure generalizability of the data-driven models identified. The resulting dynamic hybrid model consisting of the previously known information such as the dynamic mole balances and the new information identified from data, i.e. the reaction kinetic laws, is then taken as a basis for the next design step.

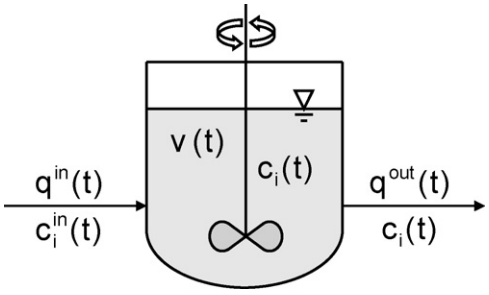

Fig. 1. Generic reactor scheme.

The paper is organized as follows: First, a short outline of the fundamentals of incremental identification to construct the unknown reaction kinetic laws is given in Section 2. To allow for model-based sequential design for the identification of data-driven model parts in dynamic process models, a new design criterion is presented in Section 3. The criterion selects the experimental degrees of freedom such as to cover the multi-dimensional space of inputs of the sought functional relation(s) in a space-filling manner. The proposed concept is illustrated on an industrially relevant reaction, the acetoacetylation of pyrrole with diketene in Section 4. Finally, the conclusions are summarized in Section 5.

2. Hybrid model construction using incremental identification

Consider the generic, ideally mixed, homogeneous and isothermally operated reactor depicted in Fig. 1. If the number and type of occurring reactions, their stoichiometries, kinetic laws and corresponding reaction parameters are known, a dynamic model of the reaction system can be constructed, capable of predicting the reactor behavior over time. The mole balance equations are set up first,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[v(t)\mathbf{c}(t)\right] = q^{\mathrm{in}}(t)\mathbf{c}^{\mathrm{in}}(t) - q^{\mathrm{out}}(t)\mathbf{c}(t) + \mathbf{f}^{\mathrm{r}}(t), \tag{1}$$

with the reactor volume $v(t)$, the volumetric feed rate $q^{\mathrm{in}}(t)$ and the molar concentration vectors $\mathbf{c}(t)$ and $\mathbf{c}^{\mathrm{in}}(t)$ of the species in the reactor and the feed, respectively. In the balance equation (1), the reaction fluxes $\mathbf{f}^{\mathrm{r}}(t)$ for the various species need further description. Using the stoichiometric matrix $\mathbf{N}$, containing the stoichiometric relations of the reaction network, a constitutive equation is set up to express these reaction fluxes in terms of the reaction rate vector $\mathbf{r}(t)$,

$$\mathbf{f}^{\mathrm{r}}(t) = v(t)\mathbf{N}^{\mathrm{T}}\mathbf{r}(t). \tag{2}$$

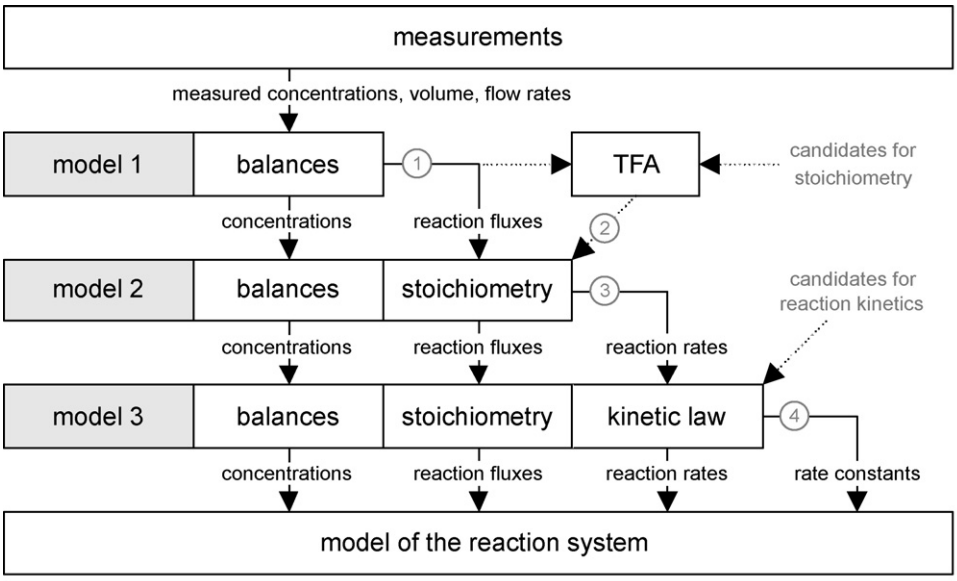

Fig. 2. Scheme of the incremental identification approach.

The reaction rates can finally be described by a set of constitutive equations as functions of the concentrations $\mathbf{c}(t)$ and the reaction parameters $\boldsymbol{\theta}$:

$$\mathbf{r}(t) = \mathbf{m}\left(\mathbf{c}(t), \boldsymbol{\theta}\right). \tag{3}$$

If suitable models for the reaction kinetics (3) are unknown, unstructured models are often taken to construct the relation between rates $\mathbf{r}$ and concentrations $\mathbf{c}$. The resulting overall reactor model, consisting of a structured part (e.g. mass balances, stoichiometries) and the unstructured reaction kinetics is hybrid, see e.g. [3,24,25].

To identify the unknown reaction kinetics from data, i.e. model parameters $\boldsymbol{\theta}$ in a given structure, but preferably also the model structure $\mathbf{m}$ to ensure generalizability [26], measurements over time are supposed to be available. They include data for the reactor volume v and the concentrations c_i of some or all of the species i involved in the reaction network. The flow rates q^{in} and q^{out} as well as the feed concentrations c_i^{in} are set by the experimental procedure and are therefore known as functions of time. Measurements taken are always corrupted with noise.

Incremental identification [22] may be applied to efficiently set up the unstructured reaction kinetics. The identification procedure is schematically depicted in Fig. 2. It exploits the hierarchical model structure inherent to any process model [27], providing stepwise identification of quantities as they are used in the modeling process. Incremental identification includes the following steps, as marked in the figure:

(1) The reaction fluxes $\hat{f}_i^{\text{r}}(t)$ are estimated individually from concentration data for each measured species i using mole balances only.
(2) If the reaction stoichiometric model $\mathbf{N}$ is unknown, target factor analysis [28] is used to test possible stoichiometries and to determine the number of occurring reactions.
(3) With the stoichiometric information, the reaction rates $\hat{\mathbf{r}}(t)$ are then calculated from the fluxes $\hat{\mathbf{f}}^{\text{r}}(t)$.
(4) Kinetic laws $\mathbf{r} = \mathbf{m}(\mathbf{c})$ are obtained by regressing the time-variant estimates of concentrations $\hat{\mathbf{c}}(t)$ and rates $\hat{\mathbf{r}}(t)$ with candidate kinetic model structures $\mathbf{m}$.

A sequence of decoupled identification problems results. Due to decoupling, the number of possible model candidates in each step is drastically reduced. In addition, kinetics identification is restricted to the solution of purely algebraic regression problems as process dynamics are considered in the flux estimation and can be omitted subsequently. This allows the use of standard training algorithms to estimate a suitable model structure and the corresponding parameters, as the inputs and outputs of the unstructured reaction kinetics have been explicitly calculated. In contrast, for simultaneous model identification approaches (see e.g. [29]) specific training algorithms or extensive validation on a set of potential structures are required to guarantee generalizable models [12]. Overall, incremental identification leads to a drastic increase in efficiency and robustness, compared to conventional simultaneous parameter estimation. Generally, incomplete measurements can be handled by the approach. Within incremental identification, a bias is introduced in the flux estimation step [22], which propagates to the estimated parameters in the kinetic laws. Despite the bias introduced, Bardow and Marquardt [30] show that the precision is largely retained in incremental identification.

Having identified suitable models of the kinetic laws, a dynamic, hybrid reaction kinetic model is constructed using the balances and reaction stoichiometries [6]. The hybrid model can then be taken to predict reactor behavior over time as a function of experimental conditions. How to choose these experimental conditions in order to achieve informative data sets for the subsequent identification step will be treated next.

3. Coverage design

A number of approaches have been proposed in literature to achieve sequential design for identification of unstructured models. Yet, when transferring these strategies to the identification of unstructured reaction kinetics $\mathbf{r} = \mathbf{m}(\mathbf{c})$ from transient concentration data $\mathbf{c}(t)$, two crucial problems are encountered:

(1) Due to the dynamics observed in the measurements, not an individual data point but rather a data set $\mathbf{c} = \{\mathbf{c}(t_q), q = 1, \ldots, n_Q\}$ is obtained from a single experiment. The data points are not independent, but behave according to the system inherent reaction dynamics. While the known sequential design criteria can be extended to predict multiple data points, the prediction of dependent data is not covered in literature.
(2) The reaction rates $\mathbf{r}$ are functions of the concentration data $\mathbf{c}$. In the functional relation to be established between the two quantities, the concentration data represent the independent variables. The set of concentration data points $\mathbf{c}$ however cannot be designed directly, as presumed in the active learning theories, but results from the set $\boldsymbol{\varphi}$ of experimental degrees of freedom, i.e. $\mathbf{c} = \mathbf{c}(\boldsymbol{\varphi})$. Here, a suitable new design $\boldsymbol{\varphi}^\star$ such as initial concentrations or feed rates is to be found which results in informative sets of concentrations and rates.

Motivated by the restrictions in existing theories, a new design criterion is presented below. Similar to space-filling designs, the approach strives to cover the multi-dimensional space of inputs to the unstructured model part by appropriately selecting the experimental degrees of freedom while taking into account the dynamic behavior of the reaction system. In the derivation of a suitable criterion for selecting the new set $\boldsymbol{\varphi}$ of design variables, data-driven modeling of a single functional relation is considered first. The criterion is then extended to systems with multiple functions to be modeled.

3.1. Single function input space coverage

Consider a dynamic system, where the quantity $y(t)$ is supposed to be described by the unstructured model

$$y(t) = f(\mathbf{x}(t)) \tag{4}$$

as a function of the n_I independent variables $\mathbf{x}(t) = [x_1(t), \ldots, x_{n_I}(t)]$.[2] Both $\mathbf{x}(t) \in \mathbb{X} \subset \mathbb{R}^{n_I}$ and $y(t) \in \mathbb{R}$ depend on the set $\boldsymbol{\varphi}$ of experimental degrees of freedom.

Assume that n_E experiments have already been conducted. For each of the n_E experiments, data sets $\mathrm{D}_i = (\mathbf{x}_i, \mathbf{y}_i)$, $\mathbf{x}_i = [\mathbf{x}_i^{\mathrm{T}}(t_1), \ldots, \mathbf{x}_i^{\mathrm{T}}(t_{n_{Q,i}})]^{\mathrm{T}}$, $\mathbf{y}_i = [y_i(t_1), \ldots, y_i(t_{n_{Q,i}})]^{\mathrm{T}}$, $i = 1, \ldots, n_E$, taken at discrete sampling times t_q, $q = 1, \ldots, n_{Q,i}$, are available, where each input set $\mathbf{x}_i$ contains the data of the n_I inputs taken at the $n_{Q,i}$ time instances sampled in the i-th experiment. A single vector of input data obtained at sampling time t_q for the n_I input variables is denoted by $(\mathbf{x}_i)_q$ and the total set of input data available is defined as

$$\mathbf{x}_{n_E}^{\mathrm{tot}} = \bigcup_{i=1}^{n_E} \mathbf{x}_i. \tag{5}$$

With the unknown function f identified from available data sets D_i, $i = 1, \ldots, n_E$, a dynamic process model can be constructed for predicting inputs $\mathbf{x}$ and outputs $\mathbf{y}$ as a function of experiment design $\boldsymbol{\varphi}$. The new set $\boldsymbol{\varphi}_{n_E+1}$ of design variables is chosen such that the new input data $\mathbf{x}_{n_E+1} = \mathbf{x}_{n_E+1}(\boldsymbol{\varphi}_{n_E+1})$ obtained from the $(n_E + 1)$ th experiment in a certain way fills out the n_I-dimensional space with additional data points.

A criterion for selecting $\boldsymbol{\varphi}_{n_E+1}$ needs to cover two requirements:

(1) The new set $\mathbf{x}_{n_E+1}$ should exhibit maximum distance to existing data $\mathbf{x}_{n_E}^{\mathrm{tot}}$. However, the exclusive application of such a criterion may lead to very compact new data sets with possible degeneration to a single data point (i.e. steady state behavior). Thus, a second criterion is desirable.
(2) The distances of data points within set $\mathbf{x}_{n_E+1}$ should be as large as possible in order to cover a preferably large input range.

The sum of minimum distances of the elements of the set $\mathbf{x}_{n_E+1}$ to those of the set $\mathbf{x}_{n_E}^{\mathrm{tot}}$ is

$$\delta_{n_E+1}^{\mathrm{tot}}(\boldsymbol{\varphi}_{n_E+1}) = \sum_{q=1}^{n_{Q,n_E+1}} \delta_{n_E+1,q}^{\mathrm{tot}}(\boldsymbol{\varphi}_{n_E+1}), \tag{6}$$

where $\delta_{n_E+1,q}^{\mathrm{tot}}$ denotes the distance of a new data point $(\mathbf{x}_{n_E+1})_q \in \mathbf{x}_{n_E+1}$ to the closest point within set $\mathbf{x}_{n_E}^{\mathrm{tot}}$ of already available data:

$$\delta_{n_E+1,q}^{\mathrm{tot}}(\boldsymbol{\varphi}_{n_E+1}) = \min_p \|(\mathbf{x}_{n_E+1})_q - (\mathbf{x}_{n_E}^{\mathrm{tot}})_p\|_2, \quad p \in [1, n_{Q,n_E}^{\mathrm{tot}}], \tag{7}$$

where $n_{Q,n_E}^{\mathrm{tot}} = \sum_{i=1}^{n_E} n_{Q,i}$ is the number of data points in the set $\mathbf{x}_{n_E}^{\mathrm{tot}}$.

Maximization of (6) fulfills the first requirement. The determination of $\delta_{n_E+1,q}^{\mathrm{tot}}$ constitutes a nearest-neighbor search which is conveniently performed using Delaunay tesselation [31]. For meeting the second requirement, the analogous criterion

$$\delta_{n_E+1}^{\mathrm{sst}}(\boldsymbol{\varphi}_{n_E+1}) = \sum_{q=1}^{n_{Q,n_E+1}} \delta_{n_E+1,q}^{\mathrm{sst}}(\boldsymbol{\varphi}_{n_E+1}), \tag{8}$$

is maximized, where $\delta_{n_E+1,q}^{\mathrm{sst}}$ expresses the distance of $(\mathbf{x}_{n_E+1})_q \in \mathbf{x}_{n_E+1}$ to the closest point in the same set:

$$\delta_{n_E+1,q}^{\mathrm{sst}}(\boldsymbol{\varphi}_{n_E+1}) = \min_p \|(\mathbf{x}_{n_E+1})_q - (\mathbf{x}_{n_E+1})_p\|_2,$$
$$p \in [1, n_{Q,n_E+1}], \quad p \neq q. \tag{9}$$

[2] As the approach is universally applicable to numerous problem settings including transient data, the generic notation x and y is used for the model inputs and outputs, respectively. In case of reaction kinetic identification, $x_i(t)$ are the concentration data and $y(t)$ corresponds to some kinetic quantity such as a flux or a rate.

Taking both requirements into account, a multi-objective optimization problem results for designing $\boldsymbol{\varphi}_{n_\mathrm{E}+1}$ with respect to maximum coverage of $\mathbb{X}^{n_\mathrm{I}}$. An overall cost functional is constructed by formulation of a weighted sum of the individual objectives [32]. Here, equal weights are heuristically chosen. Hence,

$$\delta^{\mathrm{cov}}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \delta^{\mathrm{tot}}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) + \delta^{\mathrm{sst}}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}), \tag{10}$$

is used to calculate an appropriate design $\boldsymbol{\varphi}^{\star}_{n_\mathrm{E}+1}$ for the $(n_\mathrm{E}+1)$-th experiment from

$$\boldsymbol{\varphi}^{\star}_{n_\mathrm{E}+1} = \arg\max \delta^{\mathrm{cov}}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}). \tag{11}$$

3.2. Multiple function input space coverage

As a generalization of the MISO system assumed in the previous section, we consider next the time-varying MIMO system with n_O outputs $y^{(k)}(t)$, $k = 1, \ldots, n_\mathrm{O}$, described by the unstructured model

$$\begin{aligned} y^{(1)}(t) &= f^{(1)}(\mathbf{x}^{(1)}(t)) \\ &\vdots \\ y^{(n_\mathrm{O})}(t) &= f^{(n_\mathrm{O})}\left(\mathbf{x}^{(n_\mathrm{O})}(t)\right), \end{aligned} \tag{12}$$

with $\mathbf{x}^{(k)}(t) = \{x_j(t),\, j \in \mathrm{X}_k\}$, $k = 1, \ldots, n_\mathrm{O}$, denoting the respective input vector at time t for model $f^{(k)}$, where $\mathrm{X}_k \subseteq \mathrm{X}$ describes the relevant inputs for model k chosen from the set X associated with the available inputs. Depending on the physical background of inputs x_j, $j \in \mathrm{X}$, the magnitude of their values may differ substantially. To achieve comparable conditions in the calculation of distances, the inputs require appropriate scaling. This is accomplished by scaling $x_j(t)$, $j \in \mathrm{X}$, to dimensionless quantities $\xi_j(t) \in [0, 1]$, $j \in \mathrm{X}$, according to

$$\xi_j(t) = \frac{x_j(t) - x_j^{\min}}{x_j^{\max} - x_j^{\min}}, \tag{13}$$

where the expected (or known) minimum ($x_j^{\min}$) and maximum ($x_j^{\max}$) input values are used as lower and upper bounds. Vectors $\boldsymbol{\xi}^{(k)}(t)$ concatenating all $\xi_j(t)$, $j \in \mathrm{X}_k$, $k = 1, \ldots, n_\mathrm{O}$, represent the sets of scaled inputs at time t.

Data sets $(\boldsymbol{\xi}_i^{(k)}, \mathbf{y}_i^{(k)})$, $\boldsymbol{\xi}_i^{(k)} = [(\boldsymbol{\xi}_i^{(k)})^{\mathrm{T}}(t_1), \ldots, (\boldsymbol{\xi}_i^{(k)})^{\mathrm{T}}(t_{n_{\mathrm{Q},i}})]^{\mathrm{T}}$, $\mathbf{y}_i^{(k)} = [y_i^{(k)}(t_1), \ldots, y_i^{(k)}(t_{n_{\mathrm{Q},i}})]^{\mathrm{T}}$, $k = 1, \ldots, n_\mathrm{O}$, $i = 1, \ldots, n_\mathrm{E}$, taken at discrete sampling times t_q, $q = 1, \ldots, n_{\mathrm{Q},i}$, are available from the n_E experiments. The number $n_\mathrm{I}^{(k)}$ of relevant input variables for model $f^{(k)}$ is $n_\mathrm{I}^{(k)} = \dim(\mathrm{X}_k)$. Having identified the models $f^{(k)}$, $k = 1, \ldots, n_\mathrm{O}$, from the data sets $(\boldsymbol{\xi}_i^{(k)}, \mathbf{y}_i^{(k)})$, $k = 1, \ldots, n_\mathrm{O}$, $i = 1, \ldots, n_\mathrm{E}$, extracted from the previous n_E experiments, the new design vector $\boldsymbol{\varphi}_{n_\mathrm{E}+1}$ for the subsequent experiment needs to be calculated. In analogy to the MISO case, distance criteria $\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$ are constructed for each input space $\mathbb{X}_k = [0, 1]^{n_\mathrm{I}^{(k)}}$, where Eqs. (6)–(10) are replaced by

$$\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \delta^{\mathrm{tot},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) + \delta^{\mathrm{sst},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}), \tag{14}$$

$$\delta^{\mathrm{tot},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \sum_{q=1}^{n_{\mathrm{Q},n_\mathrm{E}+1}} \delta^{\mathrm{tot},(k)}_{n_\mathrm{E}+1,q}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}), \tag{15}$$

$$\delta^{\mathrm{tot},(k)}_{n_\mathrm{E}+1,q}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \min_p \|(\boldsymbol{\xi}^{(k)}_{n_\mathrm{E}+1})_q - (\boldsymbol{\xi}^{\mathrm{tot},(k)}_{n_\mathrm{E}})_p\|_2, \quad p \in [1, n^{\mathrm{tot}}_{\mathrm{Q},n_\mathrm{E}}], \tag{16}$$

$$\delta^{\mathrm{sst},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \sum_{q=1}^{n_{\mathrm{Q},n_\mathrm{E}+1}} \delta^{\mathrm{sst},(k)}_{n_\mathrm{E}+1,q}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}), \tag{17}$$

$$\delta^{\mathrm{sst},(k)}_{n_\mathrm{E}+1,q}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \min_p \|(\boldsymbol{\xi}^{(k)}_{n_\mathrm{E}+1})_q - (\boldsymbol{\xi}^{(k)}_{n_\mathrm{E}+1})_p\|_2, \quad p \in [1, n_{\mathrm{Q},n_\mathrm{E}+1}], \quad p \neq q, \tag{18}$$

where $\delta^{\mathrm{tot},(k)}_{n_\mathrm{E}+1}$ and $\delta^{\mathrm{sst},(k)}_{n_\mathrm{E}+1}$ are the distance criteria regarding $\boldsymbol{\xi}^{\mathrm{tot},(k)}_{n_\mathrm{E}}$ and $\boldsymbol{\xi}^{(k)}_{n_\mathrm{E}+1}$ with $k = 1, \ldots, n_\mathrm{O}$, respectively, and $\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}$ is the argument considering both.

Using (14), the distance criteria $\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$ can be derived for each new input set $\boldsymbol{\xi}^{(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$. However, in the construction of an overall coverage criterion $\breve{\delta}^{\mathrm{cov}}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$, attempting to achieve optimal coverage for all input spaces $\mathbb{X}_k$, $k = 1, \ldots, n_\mathrm{O}$, it turns out that the criteria $\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$ are not directly comparable. A weighting factor is required to relate the criteria $\delta^{\mathrm{cov},(k)}_{n_\mathrm{E}+1}$ as derived next.

Consider two random points $\boldsymbol{\omega}_1 \in \mathbb{X}$ and $\boldsymbol{\omega}_2 \in \mathbb{X}$ in $\mathbb{X} = [0, 1]^{\mathrm{d}}$. The average distance between $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ depends on the dimension d, as analyzed in the following. The two scalar random values ω_1 and ω_2 stem from uniform distribution over [0, 1]. The distance between ω_1 and ω_2 is defined as

$$\vartheta = |\omega_1 - \omega_2|. \tag{19}$$

Using the *total probability theorem* (see e.g. [33]), the probability distribution $p(\vartheta)$ of ϑ is

$$p(\vartheta) = \int_{\omega_1=0}^{1} p(\vartheta|\omega_1)\, p(\omega_1)\, \mathrm{d}\omega_1 \tag{20}$$

with

$$p(\vartheta|\omega_1) = H(\omega_1 - \vartheta) + H\left((1 - \omega_1) - \vartheta\right), \tag{21}$$

where $H(v)$ is the Heaviside step function defined as

$$H(v) = \begin{cases} 0, & v < 0 \\ \dfrac{1}{2}, & v = 0 \\ 1, & v > 0 \end{cases}. \tag{22}$$

Evaluation of (20) leads to the simple expression

$$p(\vartheta) = 2(1 - \vartheta). \tag{23}$$

For two vectors $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$, randomly distributed over $\mathbb{X} = [0, 1]^d$, the expected distance between the two data is calculated

Table 1
Expected distance E_d and dimensional weighting factor τ_d for $d \in [1, 10]$

d	1	2	3	4	5	6	7	8	9	10
E_d	0.333	0.521	0.662	0.778	0.879	0.969	1.052	1.128	1.200	1.268
τ_d	1.000	0.639	0.504	0.429	0.379	0.344	0.317	0.295	0.278	0.263

as

$$\begin{aligned} E_d &= E(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2) \\ &= \int_{\vartheta_1=0}^{1} \dots \int_{\vartheta_d=0}^{1} \|\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\|_2 p(\vartheta_1) \dots p(\vartheta_d)\, \mathrm{d}\vartheta_1 \dots \mathrm{d}\vartheta_d \\ &= \int_{\vartheta_1=0}^{1} \dots \int_{\vartheta_d=0}^{1} \left[\sum_{i=1}^{d} \vartheta_i^2\right]^{1/2} 2^d \prod_{i=1}^{d} (1-\vartheta_i)\, \mathrm{d}\vartheta_1 \dots \mathrm{d}\vartheta_d, \end{aligned} \tag{24}$$

with $\vartheta_i = |(\boldsymbol{\omega}_1)_i - (\boldsymbol{\omega}_2)_i|$, where $(\boldsymbol{\omega}_j)_i$, $j = 1, 2$ is the i-th element of $\boldsymbol{\omega}_j$.

Numerical values for E_d are listed in Table 1. Expected distances increase considerably with dimensionality d. Consequently, distance criteria $\delta_{n_\mathrm{E}+1}^{\mathrm{cov},(k)}$ are influenced analogously by the input space dimensionality $n_\mathrm{I}^{(k)}$. To reduce the dependency on the number of inputs, the scaling factor

$$\tau_d = \frac{E_1}{E_d} \tag{25}$$

is introduced. Numerical values of τ_d are included in Table 1 for $d \in [1, 10]$.

Finally, an overall coverage criterion $\breve{\delta}_{n_\mathrm{E}+1}^{\mathrm{cov}}(\boldsymbol{\varphi}_{n_\mathrm{E}+1})$ is defined as

$$\breve{\delta}_{n_\mathrm{E}+1}^{\mathrm{cov}}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}) = \sum_{k=1}^{n_\mathrm{O}} \tau_{n_\mathrm{I}^{(k)}} \delta_{n_\mathrm{E}+1}^{\mathrm{cov},(k)}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}). \tag{26}$$

Analogously to (11), the appropriate design for the new experiment is calculated from

$$\boldsymbol{\varphi}_{n_\mathrm{E}+1}^{\star} = \arg\max \breve{\delta}_{n_\mathrm{E}+1}^{\mathrm{cov}}(\boldsymbol{\varphi}_{n_\mathrm{E}+1}). \tag{27}$$

The approach strives to maximize the coverage of the function input spaces. As no demands on a specific identification process are made, it can be used with any identification technique.

4. Application example: acetoacetylation system

The sequential design for reaction model identification using the incremental approach [22] is illustrated for the acetoacetylation of pyrrole with diketene [34]. To validate the proposed experimental design procedure, simulated data are used. This way, the results of the identification process can easily be compared to the model assumptions made during creation of the data. The simulation is based on the experimental work of Ruppen et al. [35], who developed a kinetic model of the reaction system.

In Section 4.1, the reaction system and the experimental conditions are introduced. The iterative identification of the reaction system is carried out in Section 4.2 using unstructured models for the reaction kinetic laws.

4.1. Reaction system and experimental conditions

The reaction system comprises the reactions

$$\mathrm{P} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{PAA}, \tag{28a}$$

$$\mathrm{D} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{DHA}, \tag{28b}$$

$$\mathrm{D} \rightarrow \mathrm{OL}, \tag{28c}$$

$$\mathrm{PAA} + \mathrm{D} \xrightarrow{\mathrm{K}} \mathrm{G}. \tag{28d}$$

In addition to the desired reaction (28a) of diketene (D) and pyrrole (P) to 2-acetoacetyl pyrrole (PAA), there are three undesired side reactions (28b)–(28d). These include the dimerization and oligomerization of diketene to dehydroacetic acid (DHA) and oligomers (OL) as well as a consecutive reaction to the by-product G.

The reactions take place in an isothermal laboratory-scale semi-batch reactor, to which a diluted solution of diketene is added continuously. Reactions (28a), (28b) and (28d) are catalyzed by pyridine (K), the concentration of which continuously decreases during the run due to addition of the diluted diketene feed. Reaction (28c), which is assumed to be promoted by other intermediate products, is not catalyzed. Hence, in the simulation model the reaction rates are described by the constitutive equations

$$r_a(t) = k_a c_\mathrm{P}(t) c_\mathrm{D}(t) c_\mathrm{K}(t), \tag{29a}$$

$$r_b(t) = k_b c_\mathrm{D}^2(t) c_\mathrm{K}(t), \tag{29b}$$

$$r_c(t) = k_c c_\mathrm{D}(t), \tag{29c}$$

$$r_d(t) = k_d c_\mathrm{PAA}(t) c_\mathrm{D}(t) c_\mathrm{K}(t), \tag{29d}$$

where k_a, k_b, k_c and k_d represent the rate constants.

The reaction fluxes $\mathbf{F}^\mathrm{r} = \left[\mathbf{f}_\mathrm{D}^\mathrm{r}, \mathbf{f}_\mathrm{P}^\mathrm{r}, \mathbf{f}_\mathrm{PAA}^\mathrm{r}, \mathbf{f}_\mathrm{DHA}^\mathrm{r}, \mathbf{f}_\mathrm{OL}^\mathrm{r}, \mathbf{f}_\mathrm{G}^\mathrm{r}\right]$ can be related to the reaction rates $\mathbf{R} = [\mathbf{r}_a, \mathbf{r}_b, \mathbf{r}_c, \mathbf{r}_d]$ by

$$\mathbf{F}^\mathrm{r} = \mathbf{V}\mathbf{R}\mathbf{N},$$

with the stoichiometric matrix

$$\mathbf{N} = \begin{bmatrix} -1 & -1 & +1 & 0 & 0 & 0 \\ -2 & 0 & 0 & +1 & 0 & 0 \\ -1 & 0 & 0 & 0 & +1 & 0 \\ -1 & 0 & -1 & 0 & 0 & +1 \end{bmatrix} \tag{30}$$

Table 2
Values of rate constants

	k_a (l^2/mol^2 min)	k_b (l^2/mol^2 min)	k_c (1/min)	k_d (l^2/mol^2 min)
Value	0.053	0.128	0.028	0.001

for the set of species S = {D, P, PAA, DHA, OL, G} and $\mathbf{V} = \mathrm{diag}\{\mathbf{v}\}$ representing the volume measurements $\mathbf{v} = [v(t_0), \ldots, v(t_{n_Q-1})]$.

The catalyst is not affected by the chemical reactions occurring. Its dilution during the run of the experiment can be modeled as

$$c_K(t) = \frac{v_0}{v(t)} c_{K,0}, \tag{31}$$

where $c_{K,0}$ is the initial concentration of catalyst in the reactor. Under the assumption that no volume change is induced by the reactions occurring, the reactor volume is modeled as

$$\frac{dv(t)}{dt} = q^{in}, \; v(t_0) = v_0, \tag{32}$$

with constant volumetric feed flow rate q^{in}.

The material balance for species $i \in S$ reads as

$$\frac{dc_i(t)}{dt} = \frac{q^{in}}{v(t)}[c_i^{in} - c_i(t)] + \frac{f_i^r(t)}{v(t)}, \quad c_i(t_0) = c_{i,0}. \tag{33a}$$

c_D^{in} is the constant concentration of diketene in the feed. For all other species, $c_i^{in} = 0$, $i \neq D$. The initial conditions $c_{i,0}$ are known.

To assess the performance of the identification approach and to allow a comparison of the modeled and the true kinetics, concentration trajectories are generated using the model described above with rate constants given in Table 2.

Concentration data are assumed to be available for the set of species S ={D, P, PAA, DHA, OL, G}. The measured concentrations are assumed to stem from a high-resolution in situ measurement technique such as IR or Raman spectroscopy, taken at a sampling interval $t_s = 20$ s over the batch time $t_f = 60$ min. The data are corrupted with normally distributed white noise. The standard deviation σ_i differs for each species i, depending on its calibration range. The calibration ranges of the species can be taken from Table 3, where concentration data are expected in the range $0 \leq c_i \leq c_i^{max}$, $i \in S$. The same relative, normally distributed error $\alpha_\sigma = 1.0\%$ within the component specific calibration range $[0, c_i^{max}]$ is assumed for each species. The standard error on the data thus is assumed to follow the relation

$$\sigma_i = \alpha_\sigma c_i^{max}, \quad i \in S. \tag{34}$$

The time-varying reactor volume $v(t)$ is measured with negligible error. In addition, error-free data on q^{in} and c_D^{in} exist. The concentration of catalyst K can be calculated from the volume and the initial concentration of catalyst according to Eq. (31).

To achieve reliable approximations of the multivariate reaction rates, experiments are to be designed to obtain concentration data over a large domain. Seven design variables (cf. Table 4) can be chosen independently between the limits specified in Table 4, to vary the experimental conditions. These are the four initial concentrations $c_{D,0}, c_{P,0}, c_{PAA,0}$ and $c_{DHA,0}$, the volumetric feed rate q^{in}, constant during a single run, the concentration of diketene in the feed c_D^{in} and the initial reactor volume v_0. The initial values of the design variables $\boldsymbol{\varphi}_1 = [c_{D,0}, c_{P,0}, c_{PAA,0}, c_{DHA,0}, q^{in}, c_D^{in}, v_0]$ in the first experiment can also be taken from Table 4. Negligible amounts of both oligomers (OL) and the by-product G are supposed to be present in the reactor at $t_0 = 0$, i.e. $c_{OL,0} = 0.01$ mol/l and $c_{G,0} = 0.01$ mol/l.

4.2. Sequential identification of unstructured reaction kinetics

The iterative work process for the identification of unstructured reaction kinetic laws is illustrated by means of the acetoacetylation example. Concentration data are generated using the model described above, which are then taken to identify the reaction system. The stoichiometric matrix **N** of the potential reactions has been given in Eq. (30). However, the number and type of actually occurring reactions are assumed unknown and need to be identified from the data. In addition, structured model candidates such as those in Eq. (29) are assumed unavailable for the description of the unknown rate laws and data-driven approaches are applied to construct the reaction kinetics. Using the coverage design criterion introduced in Section 3, experiments are planned iteratively, based on the hybrid model predictions identified from previous runs. Neural networks with 3 nodes in the hidden layer and Bayesian regularization [23,36] as a training algorithm are used to predict appropriate kinetic models from available (simulated) experimental data.

With the noise-corrupted concentration measurements obtained from the initial experiment, a primary model of the reaction kinetics needs to be identified first: reaction fluxes are estimated for the set of measured species S using mole balances [22]. Using the estimated fluxes, the network stoichiometric model is then identified using recursive target factor analysis [28]. In our example, all reactions (28a) to (28d) are accepted, which is in accordance with the assumption made in data generation. The reaction rates $\hat{r}_j(t)$, $j \in \{a, b, c, d\}$ of the occurring reactions are then calculated as a function of time from the available reaction fluxes and the network stoichiometries identified.

Table 3
Concentration ranges for calibration

	c_D (mol/l)	c_P (mol/l)	c_{PAA} (mol/l)	c_{DHA} (mol/l)	c_{OL} (mol/l)	c_G (mol/l)
Min	0.00	0.00	0.00	0.00	0.00	0.00
Max	0.38	0.80	0.45	0.63	0.52	0.05

Table 4
Initial values and admissible range of independent variables

	$c_{D,0}$ (mol/l)	$c_{P,0}$ (mol/l)	$c_{PAA,0}$ (mol/l)	$c_{DHA,0}$ (mol/l)	q^{in} (ml/min)	c_D^{in} (mol/l)	v_0 (l)
Initial	0.14	0.30	0.08	0.01	5.0	6.0	0.5
Min	0.02	0.30	0.08	0.01	5.0	3.0	0.5
Max	0.14	0.80	0.20	0.05	10.0	6.0	1.0

As a next step, smooth concentration estimates $\hat{c}_i(t), i \in S$, are obtained from the noisy measurements.

At this point, data sets $D_1^{(k)} = (\mathbf{x}_1^{(k)}, \mathbf{y}_1^{(k)})$, $\mathbf{x}_1^{(k)} = [(\mathbf{x}_1^{(k)})^T(t_0), \ldots, (\mathbf{x}_1^{(k)})^T(t_{n_{Q,1}-1})]^T$, $\mathbf{y}_1^{(k)} = [y_1^{(k)}(t_0), \ldots, y_1^{(k)}(t_{n_{Q,1}-1})]^T$, $k = 1, \ldots, n_O$, have been obtained from the first experiment. Inputs $\mathbf{x}$ are the estimated concentration data $\hat{\mathbf{c}}$ and outputs $\mathbf{y}$ correspond to the estimated rates $\hat{\mathbf{r}}$. The number of outputs is $n_O = 4$ and the number of sampling points in the first experiment amounts to $n_{Q,1} = 181$. Input sets X_k for the kinetic models correspond to the sets of concentrations for the species S_k influencing the individual reactions, which are assumed known.[3] Here, $S_1 = \{D,P,K\}$, $S_2 = \{D,K\}$, $S_3 = \{D\}$ and $S_4 = \{D,PAA,K\}$. The outputs are $y_1^{(1)}(t) = \hat{r}_a(t)$, $y_1^{(2)}(t) = \hat{r}_b(t)$, $y_1^{(3)}(t) = \hat{r}_c(t)$ and $y_1^{(4)}(t) = \hat{r}_d(t)$. Using data sets $D_1^{(k)}$, $k = 1, \ldots, n_O$, models $f^{(k)}$ are derived for the description of the individual reaction kinetics based on neural network approximation. The resulting hybrid model, consisting of the dynamic mole balances, the reaction stoichiometries and the reaction kinetic laws identified, can now be taken to predict concentration trajectories as a function of the experiment design $\boldsymbol{\varphi}$.

The design for the subsequent experiment is calculated according to Section 3. For scaling the concentration data, serving as inputs to the approximation, to unity domain according to (13), the concentration calibration range (Table 3) specifies lower and upper bounds x_j^{min} and x_j^{max}, respectively. Based on the predictions of the identified dynamic model, the design vector $\boldsymbol{\varphi}_2^\star$ for the following experimental run is calculated as the solution of optimization problem (27).

With the augmented data set, stemming from the first and the second experiment, the reaction kinetic models are updated according to the procedure sketched above. A new design vector $\boldsymbol{\varphi}_3^\star$ is then calculated using the resulting, more precise dynamic model of the reaction system. Within the iterative experiment planning and model identification procedure, the model accuracy and thus the quality of the experiment design using the coverage approach gradually improve. The iterative process is continued up to the pre-specified number of experiments, $n_E = 16$. Alternatively, stopping criteria based on approximation accuracy may be chosen to limit the number of experiments.

To assess the quality of estimates achieved with the proposed design criterion, identification results are compared to those obtained with *a priori* factorial design. To keep the number of experiments at a moderate level, a 2^{7-3} fractional factorial design [13] with $n_E = 16$ experiments has been chosen, where initial concentrations $c_{D,0}$, $c_{P,0}$, $c_{PAA,0}$, and the concentration of diketene in the feed, c_D^{in}, are considered as the main influencing (dominant) factors. The resulting design can be taken from Table 5. In comparison, Table 6 summarizes the resulting design vectors for the first $n_E = 16$ experiments using coverage design.

Two representative concentration spaces,[4] (c_D, c_P) and (c_D, c_K), are depicted in Figs. 3 and 4 for the fractional factorial design and the coverage design, respectively. For the latter, the designs are numbered in the order of their experimental realization. Obviously, the coverage approach features a more complete filling of the concentration spaces compared to the fractional factorial design. Exemplarily, the resulting prediction for reaction rate $r_b(c_D, c_K)$ after $n_E = 16$ experiments using coverage design is depicted in Fig. 5, together with the reaction rates estimated for each of the experimental runs.

To be able to compare the results achieved for the fractional factorial (FFC) and the coverage (COV) design, experimental runs are simulated for $n_C = 50$ random designs based on the models identified using neural networks after $n_E = 16$ experiments. The average ($\bar{\epsilon}_j^r$) and maximum ($\epsilon_j^{r,max}$) relative prediction errors between predicted and true reaction rates for reaction j are calculated as

$$\bar{\epsilon}_j^r = \frac{1}{n_C n_Q} \sum_{\ell=1}^{n_C} \sum_{q=0}^{n_Q-1} \epsilon_{j,q,\ell}^r, \tag{35a}$$

$$\epsilon_j^{r,max} = \max_{q,\ell} \epsilon_{j,q,\ell}^r, \tag{35b}$$

$$\epsilon_{j,q,\ell}^r = \left| \frac{r_j^{pred,\ell}(t_q) - r_j^{true,\ell}(t_q)}{r_j^{true,\ell}(t_q)} \right|, \tag{35c}$$

where $r_j^{pred,\ell}(t_q)$ denotes the model prediction for reaction $j \in \{a, b, c, d\}$, at time t_q for the ℓ th random design and $r_j^{true,\ell}(t_q)$ is the true reaction rate obtained from the data generation model. The number of sampling points is $n_Q = 181$ for each run. The average and maximum relative prediction errors $\bar{\epsilon}_i^c$ and $\epsilon_i^{c,max}$, respectively, between the hybrid model predictions $c_i^{pred,\ell}(t_q)$

[3] In the case of unknown input sets, identification techniques with inherent input selection based on neural networks [37], kernels [38] or sparse grids [39] may be applied. These techniques alternatively also allow the inputs to be determined from already available data prior to the iterative identification process.

[4] Space (c_D, c_K) corresponds to the input set $S_2 = \{D,K\}$ for the kinetic model of r_b. Due to the three-dimensional nature of set $S_1 = \{D, P, K\}$, subset $\{D, P\}$ has been representatively chosen in Figs. 3 (left) and 4 (left) to visualize the concentration space covered.

Table 5
Fractional factorial design

Run	$c_{D,0}$ (mol/l)	$c_{P,0}$ (mol/l)	$c_{PAA,0}$ (mol/l)	$c_{DHA,0}$ (mol/l)	q^{in} (ml/min)	c_D^{in} (mol/l)	v_0 (l)
01	0.0200	0.3000	0.0800	0.0100	5.0000	3.0000	0.5000
02	0.0200	0.3000	0.0800	0.0500	5.0000	6.0000	1.0000
03	0.0200	0.3000	0.2000	0.0500	10.000	3.0000	1.0000
04	0.0200	0.3000	0.2000	0.0100	10.000	6.0000	0.5000
05	0.0200	0.8000	0.2000	0.0500	5.0000	3.0000	0.5000
06	0.0200	0.8000	0.2000	0.0100	5.0000	6.0000	1.0000
07	0.0200	0.8000	0.0800	0.0100	10.000	3.0000	1.0000
08	0.0200	0.8000	0.0800	0.0500	10.000	6.0000	0.5000
09	0.1400	0.3000	0.2000	0.0100	5.0000	3.0000	1.0000
10	0.1400	0.3000	0.2000	0.0500	5.0000	6.0000	0.5000
11	0.1400	0.3000	0.0800	0.0500	10.000	3.0000	0.5000
12	0.1400	0.3000	0.0800	0.0100	10.000	6.0000	1.0000
13	0.1400	0.8000	0.0800	0.0500	5.0000	3.0000	1.0000
14	0.1400	0.8000	0.0800	0.0100	5.0000	6.0000	0.5000
15	0.1400	0.8000	0.2000	0.0100	10.000	3.0000	0.5000
16	0.1400	0.8000	0.2000	0.0500	10.000	6.0000	1.0000

Table 6
Coverage-based experimental design for neural networks

Run	$c_{D,0}$ (mol/l)	$c_{P,0}$ (mol/l)	$c_{PAA,0}$ (mol/l)	$c_{DHA,0}$ (mol/l)	q^{in} (ml/min)	c_D^{in} (mol/l)	v_0 (l)
01	0.1400	0.3000	0.0800	0.0100	5.0000	6.0000	0.5000
02	0.0200	0.8000	0.0889	0.0335	5.0000	3.0000	1.0000
03	0.1400	0.3000	0.0898	0.0301	10.000	6.0000	0.5000
04	0.1400	0.8000	0.1694	0.0164	10.000	4.8692	1.0000
05	0.0200	0.8000	0.1694	0.0164	5.0000	3.0000	0.5237
06	0.0200	0.8000	0.1694	0.0164	6.8362	3.9801	0.5306
07	0.0200	0.8000	0.1694	0.0164	10.000	3.0386	1.0000
08	0.0200	0.3625	0.1693	0.0164	9.3750	6.0000	0.9677
09	0.0200	0.3000	0.1694	0.0164	10.000	3.3322	0.8947
10	0.0569	0.7655	0.1765	0.0355	9.9019	5.9103	0.7761
11	0.0200	0.3291	0.1765	0.0355	5.0000	3.0000	0.5000
12	0.0200	0.7218	0.1765	0.0355	6.9526	3.0000	0.5000
13	0.0734	0.7887	0.1765	0.0355	7.8018	3.9550	0.8273
14	0.0200	0.3000	0.1765	0.0355	6.3914	3.0000	1.0000
15	0.1400	0.3000	0.1765	0.0355	10.000	4.7675	1.0000
16	0.0202	0.7990	0.1765	0.0355	10.000	4.2775	0.9990

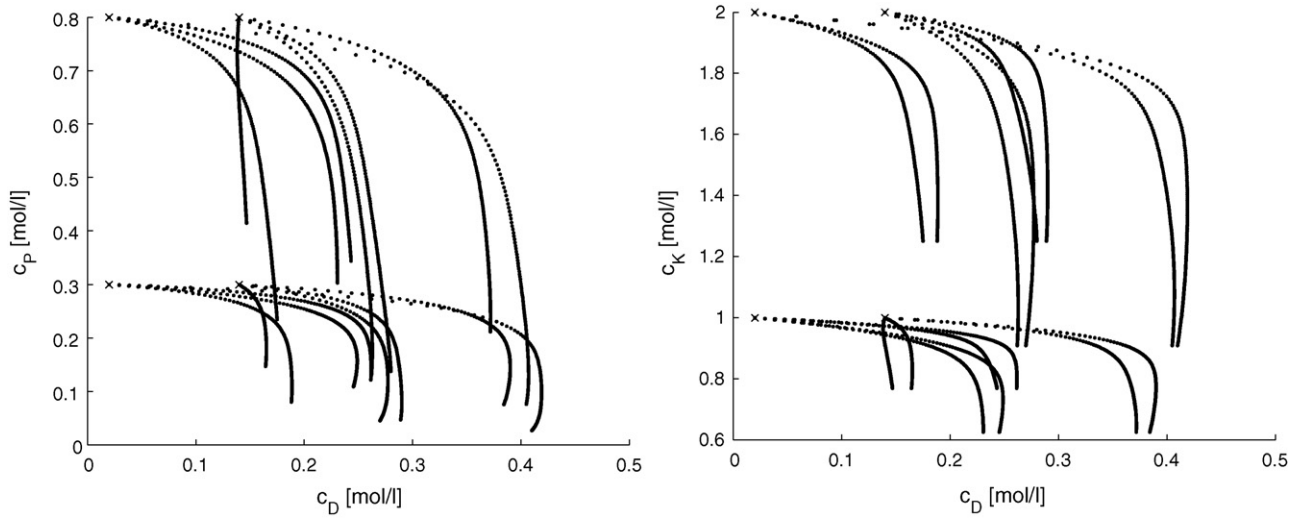

Fig. 3. Concentration trajectories for (c_D, c_P) (left) and (c_D, c_K) (right) using fractional factorial design.

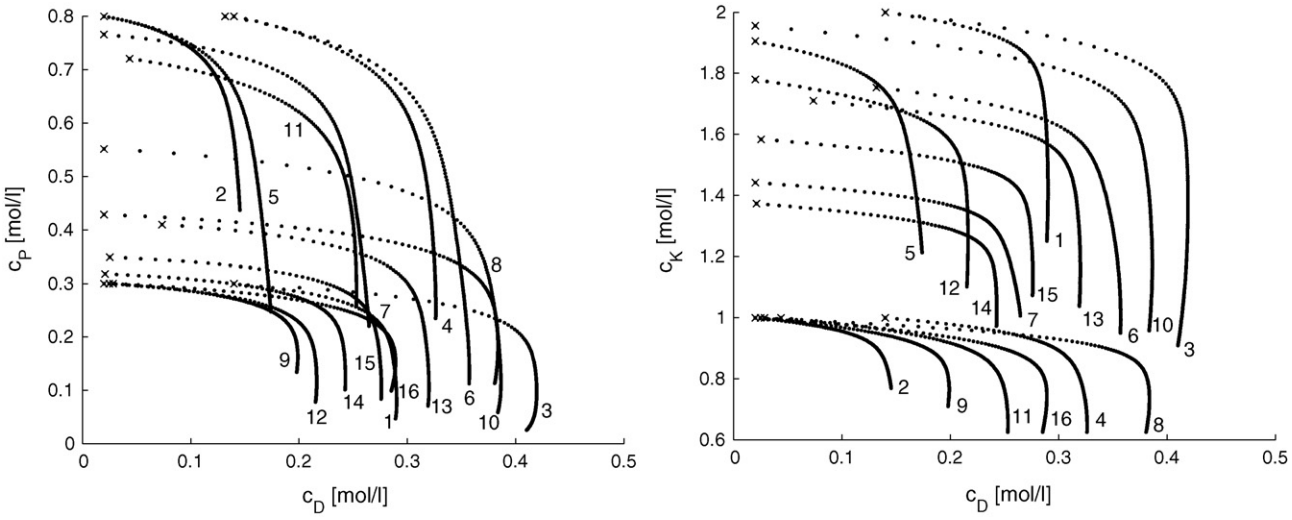

Fig. 4. Concentration trajectories for (c_D, c_P) (left) and (c_D, c_K) (right) using the coverage approach.

and the true concentrations $c_i^{\text{true},\ell}(t_q)$ for species $i \in S$ are calculated analogously.

The average and maximum relative reaction rate prediction errors for the n_C random designs are given in Table 7. Approximation accuracies achieved with the coverage design clearly exceed those obtained by factorial design. For the rates r_a through r_c, good results are achieved concerning average prediction errors $\bar{\epsilon}_j^{\text{r}}$. Unreliable estimates result for the small values of rate r_d, the rate is not identifiable in practice. For both types of design, high maximum errors result for all rates. These are caused by very small reaction rates, where already small absolute errors have a large impact on relative accuracy.

Table 8 lists the average and maximum prediction errors with respect to concentration data for the relevant species D, P, PAA and DHA. The more precise models for the individual reaction rates achieved by coverage design accordingly lead to better predictions of the corresponding hybrid models, which show excellent accuracies throughout the sets.

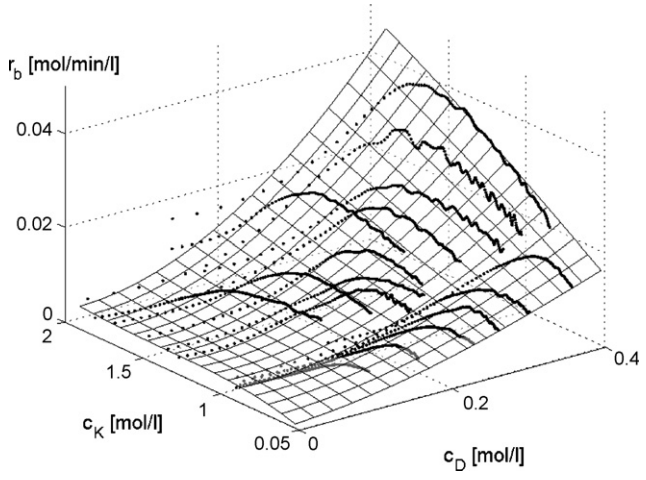

Fig. 5. Estimated reaction rates r_b (dotted lines) and reaction kinetic laws (surface plot) using the coverage approach.

As a next step, the prediction errors are analyzed for a varying number of experimental runs using coverage design, exemplarily evaluated on the predicted concentration transients. Naturally, the prediction error is expected to decrease for an increasing number of experimental runs. In Fig. 6 (left), the convergence of prediction ratio $\Lambda_i(n_E)$ for the individual concentrations $i \in S = \{D, P, PAA, DHA\}$ is depicted, where the prediction error $\bar{\epsilon}_i^{\text{c}}$ after a number of n_E experiments is scaled to the prediction after $n_E = 16$ experiments according to

$$\Lambda_i(n_E) = \ln\left(\frac{\bar{\epsilon}_i^{\text{c}}(n_E)}{\bar{\epsilon}_i^{\text{c}}(n_E = 16)}\right). \tag{36}$$

Table 7
Relative reaction rate prediction errors $\bar{\epsilon}_j^{\text{r}}$ and ϵ_j^{r},max obtained from fractional factorial (FFC) and coverage (COV) design

	r_a (%)	r_b (%)	r_c (%)	r_d (%)
FFC				
$\bar{\epsilon}_j^{\text{r}}$	6.624	21.28	7.174	349.9
ϵ_j^{r},max	379.3	3808	627.1	57685
COV				
$\bar{\epsilon}_j^{\text{r}}$	3.861	8.751	3.247	199.2
ϵ_j^{r},max	231.4	1321	263.8	18313

Table 8
Relative concentration prediction errors $\bar{\epsilon}_i^{\text{c}}$ and $\epsilon_i^{\text{c,max}}$ obtained from fractional factorial (FFC) and coverage (COV) design

	c_D (%)	c_P (%)	c_{PAA} (%)	c_{DHA} (%)
FFC				
$\bar{\epsilon}_j^{\text{c}}$	0.536	2.766	2.231	1.192
ϵ_j^{c},max	2.216	8.299	5.103	7.371
COV				
$\bar{\epsilon}_j^{\text{r}}$	0.384	0.489	0.977	0.944
ϵ_j^{c},max	3.067	3.051	2.352	10.87

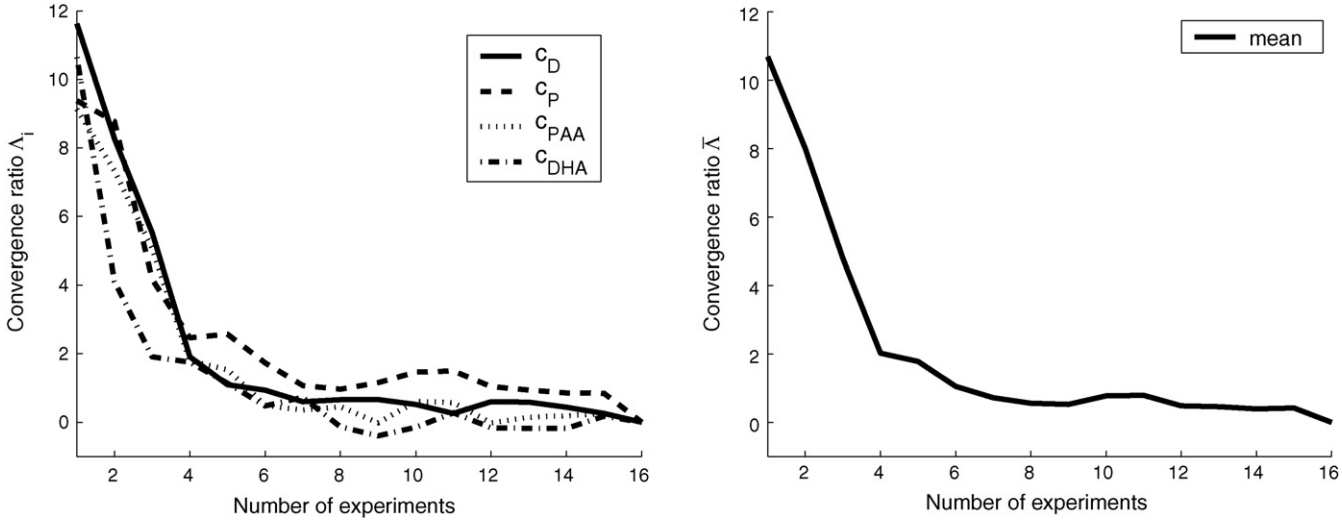

Fig. 6. Convergence of predictions for the coverage approach— $\Lambda_i(n_E)$ for the individual concentrations (left) and mean $\bar{\Lambda}(n_E)$ (right).

Fig. 6 (right) shows the mean value for the set of concentrations,

$$\bar{\Lambda} = \frac{1}{\dim(S)} \sum_{i \in \mathcal{S}} \Lambda_i. \tag{37}$$

Note that the prediction error is not strictly monotonically decreasing, due to the fact that depending on the design chosen, the estimated fluxes may exhibit substantial errors and thus impair the overall prediction.

To further analyze the performance of the coverage approach, its predictions are compared to those achieved using random designs (RND) for a variable number of experiments. The expression

$$\Gamma_i(n_E) = \ln\left(\frac{\bar{\epsilon}_i^{c,cov}(n_E)}{\bar{\epsilon}_i^{c,rnd}(n_E)}\right) \tag{38}$$

describes the ratio of average, relative concentration prediction errors obtained from designing the experiments using the coverage approach ($\bar{\epsilon}_i^{c,cov}$) on the one hand, and the random approach ($\bar{\epsilon}_i^{c,rnd}$) on the other hand. The ratio Γ_i depends on the number n_E of experiments realized. Concerning random designs, the random selection of a single, particularly suitable or disadvantageous set of designs can lead to untypically good or poor values of $\bar{\epsilon}_i^{c,rnd}(n_E)$ and cause notable scatter in $\Gamma_i(n_E)$. To obtain a representative result, the term $\bar{\epsilon}_i^{c,rnd}(n_E)$ represents an average acquired from a number of $n_D = 40$ random experimental design sets, each set containing n_E individual designs of the experimental degrees of freedom.

In Fig. 7 (left), Γ_i is depicted for the individual concentration transients as a function of the experiments conducted. The average,

$$\bar{\Gamma} = \frac{1}{\dim(S)} \sum_{i \in \mathcal{S}} \Gamma_i, \tag{39}$$

referring to the whole concentration set, is shown in Fig. 7 (right). Clearly, the prediction accuracies using COV exceed those of random designs, in particular for moderate n_E. A minimum of $\bar{\Gamma}$ is found for $n_E = 3$, where the errors achieved with the coverage approach are only 5% of those obtained by random design. For $n_E = 8$, the ratio is 20% and for $n_E = 16$, we calculate 40%.

A comparison to the prediction convergence of (fractional) factorially designed experiments is more difficult to realize. For those, complete sets of experiments are selected in advance the number of which usually represents a power of 2 for the most commonly employed two-level fractional designs [13]. The design chosen also depends on the experimenter's often subjective choice of the dominant and subordinate parameters. In Fig. 7 (right), the prediction ratio for the 16-experiment fractional factorial design used hitherto is marked by a circle. Obviously, the design does not perform much better than an average random design for 16 experiments. Yet, as a single instance of the multitude of potential fractional factorial designs, the result may not be representative for the class of fractional factorial designs as such.

To allow a better comparison between coverage, fractional factorial and random designs, identification results were calculated for a range of fractional factorial designs, varying in the choice of dominant parameters and the dependencies chosen for the remaining (subordinate) design parameters. Similar to the random designs examined, identification results were computed for $n_F = 40$ different factorial designs and the prediction ratios were calculated in analogy to Eqs. (38) and (39). The results obtained for 4 (2^{7-5}), 8 (2^{7-4}) and 16 (2^{7-3}) experiments are plotted in Fig. 7 (right) likewise.

Apparently, the average fractional factorial design yields better results compared to the random choice of design parameters, but the accuracy is worse compared to the coverage approach. Compared to average fractional factorial designs, the residuals achieved with COV are 25% for 4 experiments and 50% for 8 and 16 experiments, respectively. Generally, the differ-

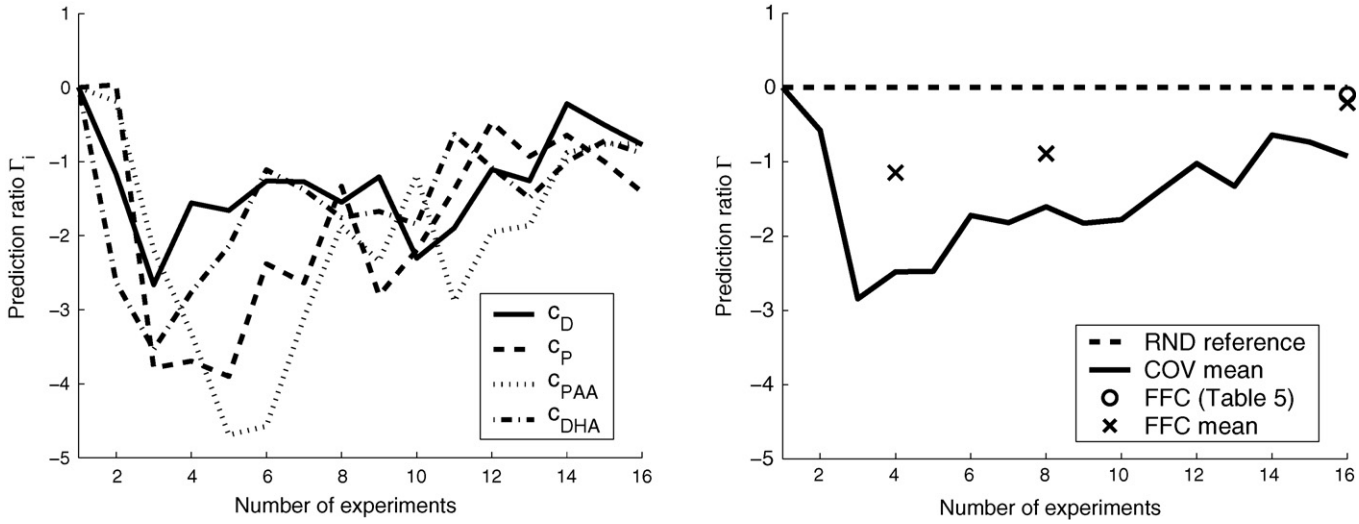

Fig. 7. Ratio of prediction errors for COV vs. RND— $\Gamma_i(n_E)$ for the individual concentrations (left) and mean $\bar{\Gamma}(n_E)$ (right).

ence between the design procedures becomes less distinct for a high number of experiments conducted ($\bar{\Gamma} \overset{n_E \to \infty}{\to} 0$). This is not an unexpected result: with an increasing coverage of the feature space, the quality of designs becomes less important. Hence, the coverage approach is particularly powerful for a limited number of experiments to be conducted.

5. Conclusions

A novel design criterion has been derived for the identification of data-driven model parts from dynamic data. The multi-dimensional spaces of inputs to the models to be identified are covered by appropriately selecting the experimental degrees of freedom. Therewith, the new experiment can simultaneously be designed for multiple models to be identified. With a dynamic process model – identified from data gathered in previous experiments – required to optimize the experimental degrees of freedom for the following experiment, an iterative design and identification procedure results.

When applied to the identification of reaction kinetics using neural networks in a hybrid process model structure, the approach performed significantly better than *a priori* chosen factorial or even random designs. Much smaller prediction errors were achieved for the unknown reaction kinetics and the corresponding dynamic hybrid process model, in particular for a limited number of experiments. The coverage design criterion proposed is not limited to the identification of reaction kinetics but universally applicable to model identification problems from (dynamic) data in which the independent model inputs cannot be designed directly.

The design criterion can be applied independently of the identification technique used. In this work, incremental identification has been used for its flexibility and computational efficiency. The modeler is free to choose any data-driven or hybrid models to construct the unknown model parts. High computational efficiency is achieved due to problem decoupling. Yet, it has been shown that additional experiments may also deteriorate the approximation result using incremental identification. Missing error estimates on the reaction fluxes may lead to erroneous reaction flux estimates, thus impairing the accuracy of identified models. Both model accuracy and consequently the design process are expected to perform even better once the errors on the fluxes can be calculated.

Incremental identification supports the use of both structured and unstructured models for the reaction kinetic laws to be identified, based on the available knowledge on the individual model parts [12]. While combined estimation of both model types from data is straightforward, the question arises on how to design the new experiment with respect to the identification of both structured and unstructured kinetic models from the same set of data. A possible solution approach is seen in the application of alternative cost functionals in the design for data-driven models. While the criterion proposed optimizes the coverage of the input space, maximization of the expected information gain (see [17]) may represent a further option. Such a strategy bears resemblance with that pursued in the design for structured approaches, i.e. maximization of the information content of the experiments. A combination of both approaches thus appears feasible. Yet, to specify the information content gained from an experiment, error estimates on the fluxes need further investigation.

Acknowledgements

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center (SFB 540) “Model-based experimental analysis of kinetic phenomena in fluid multiphase reactive systems”.

References

[1] R.J. Berger, E.H. Stitt, G.B. Marin, F. Kapteijn, J.A. Moulijn, EUROKIN. Chemical reaction kinetics in practice, CATTECH 5 (1) (2001) 30–60.

[2] A. Helbig, O. Abel, W. Marquardt, Model predictive control for on-line optimization of semi-batch reactors, in: Proceedings of the 1998 American Control Conference, vol. 1, Philadelphia, 1998, pp. 1695–1699.

[3] D.C. Psichogios, L.H. Ungar, A hybrid neural network – first principles approach to process modeling, AIChE J. 38 (10) (1992) 1499–1511.
[4] A. Tholudur, W.F. Ramirez, Neural-network modeling and optimization of induced foreign protein production, AIChE J. 45 (8) (1999) 1660–1670.
[5] P.F. Van Lith, B.H.L. Betlem, B. Roffel, A structured modeling approach for dynamic hybrid fuzzy first-principles models, J. Process Control 12 (2002) 605–615.
[6] M. Brendel, A. Mhamdi, D. Bonvin, W. Marquardt, An incremental approach for the identification of reaction kinetics, in: Proceedings of the 7th IFAC Symposium on Advanced Control of Chemical Processes, ADCHEM 2003, HongKong, 2003, pp. 177–182.
[7] O. Kahrs, W. Marquardt, The validity domain of hybrid models and its application in process optimization, Chem. Eng. Process. 46 (11) (2007) 1054–1066.
[8] E. Walter, L. Pronzato, Qualitative and quantitative experiment design for phenomenological models—a survey, Automatica 26 (2) (1990) 195–213.
[9] A.F. Emery, A.V. Nenarokomov, Optimal experiment design, Meas. Sci. Technol. 9 (1998) 864–876.
[10] S.P. Asprey, S. Macchietto, Statistical tools for optimal dynamic model building, Comput. Chem. Eng. 24 (2000) 1261–1267.
[11] W. Marquardt, Model-based experimental analysis of kinetic phenomena in multi-phase reactive systems, Trans. IChemE, Part A: Chem. Eng. Res. Des. 83 (A6) (2005) 561–573.
[12] M. Brendel, Incremental Identification of Complex Reaction Systems, VDI-Verlag, Düsseldorf, 2006.
[13] A.C. Atkinson, A.N. Donev, Optimum Experimental Designs, Oxford Statistical Science Series, Oxford, 1996.
[14] M.D. McKay, W.J. Conover, R.J. Beckman, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (1979) 239–245.
[15] J.P.C. Kleijnen, S.M. Sanchez, T.W. Lucas, T.M. Cioppa, A user's guide to the brave new world of designing simulation experiments, Tech. rep., Tilburg University, Center for Economic Research, 2003.
[16] W. Kinzel, P. Ruján, Improving a network generalization ability by selecting examples, Europhys. Lett. 13 (1990) 473–477.
[17] H.S. Seung, M. Opper, H. Sompolinsky, Query by committee, in: Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, ACM Press, New York, 1992, pp. 287–294.
[18] D. Cohn, Neural network exploration using optimal experiment design, in: J.D. Cowan, G. Tesauro, J. Alspector (Eds.), Advances in Neural Information Processing Systems, vol. 6, Morgan Kaufmann, San Francisco, 1994, pp. 679–686.
[19] D. Cohn, L. Atlas, R. Ladner, Improving generalization with active learning, Mach. Learn. 15 (2) (1994) 201–221.
[20] B.T. Zhang, Accelerated learning by active example selection, Int. J. Neural Syst. 5 (1) (1994) 67–75.
[21] M. Hasenjäger, H. Ritter, Active learning in neural networks, in: L.C. Jain, J. Kacprzyk (Eds.), New Learning Paradigms in Soft Computing, vol. 84 of Studies in Fuzziness and Soft Computing, Physica-Verlag GmbH, Heidelberg, Germany, 2002, pp. 137–169.
[22] M. Brendel, D. Bonvin, W. Marquardt, Incremental identification of kinetic models for homogeneous reaction systems, Chem. Eng. Sci. 61 (2006) 5404–5420.
[23] D.J.C. MacKay, A practical Bayesian framework for backpropagation networks, Neural Comput. 4 (3) (1992) 448–472.
[24] M. Agarwal, Combining neural and conventional paradigms for modelling, prediction and control, Int. J. Syst. Sci. 28 (1) (1997) 65–81.
[25] R. Oliveira, Combining first principles modelling and artificial neural networks: a general framework, Comput. Chem. Eng. 28 (2004) 755–766.
[26] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, Heidelberg, 1995.
[27] W. Marquardt, Towards a process modeling methodology, in: R. Berber (Ed.), Methods of Model-based Control, vol. 293 of NATO-ASI Ser. E, Applied Sciences, Kluwer, Dordrecht, 1995, pp. 3–41.
[28] D. Bonvin, D.W.T. Rippin, Target factor analysis for the identification of stoichiometric models, Chem. Eng. Sci. 45 (12) (1990) 3417–3426.
[29] Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974.
[30] A. Bardow, W. Marquardt, Incremental and simultaneous identification of reaction kinetics: methods and comparison, Chem. Eng. Sci. 59 (13) (2004) 2673–2684.
[31] A. Okabe, B. Boots, K. Sugihara, S.N. Chiu, Spatial Tessellations—Concepts and Applications of Voronoi Diagrams, John Wiley & Sons, New York, 2000.
[32] P.A. Clark, A.W. Westerberg, Optimization for design problems having more than one objective, Comput. Chem. Eng. 7 (1983) 259–278.
[33] I.N. Bronshtein, K.A. Semendyayev, Handbook of Mathematics, 3rd ed., Springer, London, 1985.
[34] D. Ruppen, A Contribution to the Implementation of Adaptive Optimal Operation for Discontinuous Chemical Reactors, Ph.D. Thesis, ETH Zürich, 1994.
[35] D. Ruppen, D. Bonvin, D.W.T. Rippin, Implementation of adaptive optimal operation for a semi-batch reaction system, Comput. Chem. Eng. 22 (1-2) (1998) 185–199.
[36] D.J.C. MacKay, Bayesian interpolation, Neural Comput. 4 (3) (1992) 415–447.
[37] M.A. Kramer, Nonlinear principal component analysis using autoassociative neural networks, AIChE J. 37 (2) (1991) 233–243.
[38] A.M. Jade, B. Srikanth, V.K. Jayaraman, B.D. Kulkarni, J.P. Jog, L. Priya, Feature extraction and denoising using kernel PCA, Chem. Eng. Sci. 58 (2003) 4441–4448.
[39] O. Kahrs, M. Brendel, W. Marquardt, Incremental identification of NARX models by sparse grid approximation, in: Proceedings of the 16th IFAC World Congress, July 3–8, 2005, Prague, Czech Republic, 2005.