Chemical Engineering Journal 400 (2020) 125469

Contents lists available at ScienceDirect

Chemical Engineering Journal

journal homepage: www.elsevier.com/locate/cej

Training set design for machine learning techniques applied to the approximation of computationally intensive first-principles kinetic models

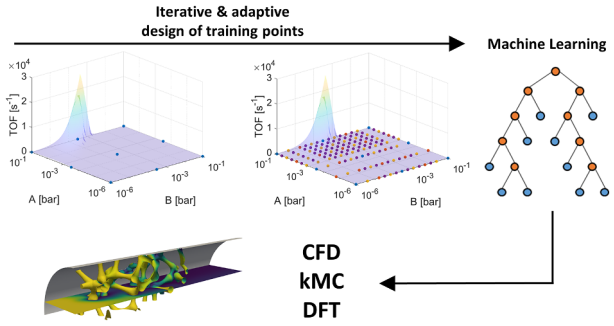
Mauro Bracconi[1], Matteo Maestri[*,2]

Laboratory of Catalysis and Catalytic Processes, Dipartimento di Energia, Politecnico di Milano, via la Masa 34, 20156 Milano, Italy

HIGHLIGHTS

- A design procedure of the training data for Machine Learning algorithms is proposed.
- Accurate prediction with up to 80% less datapoints than evenly distributed grid.
- Proof-of-concept of multiscale simulations based on kMC and microkinetic modelling.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Machine learning
Multiscale modelling
Microkinetic
Kinetic Monte Carlo
Computational fluid dynamics

ABSTRACT

We propose a design procedure for the generation of the training set for Machine Learning algorithms with a specific focus on the approximation of computationally-intensive first-principles kinetic models in catalysis. The procedure is based on the function topology and behavior, by means of the calculation of the discrete gradient, and on the relative importance of the independent variables. We apply the proposed methodology to the tabulation and regression of mean-field and kinetic Monte Carlo models aiming at their coupling with reactor simulations. Our tests – in the context both of mean-field kinetics and kinetic Monte Carlo simulations – show that the procedure is able to design a dataset that requires between 60 and 80% fewer data points to achieve the same approximation accuracy than the one obtained with an evenly distributed grid. This strong reduction in the number of points results in a significant computational gain and a concomitant boost of the approximation efficiency. The Machine Learning algorithms trained with the results of the procedure are then included in both macroscopic reactor models and computational fluid dynamics (CFD) simulations. First, a Plug Flow Reactor is employed to carry out a direct comparison with the solution of the full first-principles kinetic model. The results show an excellent agreement within 0.2% between the models. Then, the CFD simulation of complex tridimensional geometry is carried out by using a tabulated kMC model for CO oxidation on Ruthenium oxide, thus providing a showcase of the capability of the approach in making possible the multiscale simulation of complex chemical reactors.

* Corresponding author.
E-mail address: matteo.maestri@polimi.it (M. Maestri).
[1] ORCID – Mauro Bracconi: 0000-0001-7643-3214.
[2] ORCID – Matteo Maestri: 0000-0002-8925-3869.

https://doi.org/10.1016/j.cej.2020.125469
Received 10 April 2020; Received in revised form 10 May 2020; Accepted 11 May 2020
Available online 14 May 2020

1. Introduction

In recent years, a large effort has been devoted to the integration of mean-field microkinetic models and kinetic Monte Carlo (kMC) simulations in continuum modeling (e.g., Computational Fluid Dynamics – CFD) of chemical reactors [1–3]. This task is widely acknowledged to be crucial in view of leveraging the fundamental insights of first-principles computational catalysis to realistic and technologically relevant applications [4,5]. The coupling between reactors models and the heterogeneous chemistry intrinsically requires bridging phenomena spanning from the atomistic to the reactor level. This covers orders of magnitude of difference in both the time and length scales, thus making the direct coupling impractical from a numerical point of view. The analytical expression for the rates of the elementary steps and the closed form of the differential equation governing the evolution of the system makes possible their incorporation in reactor models. The strong non-linearity and stiffness, however, require specific treatments of the operator related to the chemistry source terms in the governing equations at the macroscale. For instance, as proposed by our group [6,7], the operator splitting algorithm can be used to incorporate detailed mean-field microkinetic models in CFD simulations. The solution of the heterogeneous chemistry increases the computational costs associated with the fluid dynamic accounting for 70–90% of the overall simulation time [8], thus hindering the application to large scale systems. Several methods have been proposed to reduce the computational burden, such as on-the-fly tabulation techniques, e.g. in situ adaptive tabulation (ISAT) [8] and cell agglomeration algorithms [9]. However, when the computational cost of the calculation of the rates is very computationally demanding (as for the case of kMC simulations), this approach becomes impractical, thus hindering the coupling of first-principles kinetic models with continuum simulations. In this situation, a possible strategy for the solution of the multiscale problem has been proposed [10] through the adoption of pre-computed production rates properly tabulated. This is based on the assumption that the catalyst surface instantaneously adapts and relaxes to a new steady-state catalytic activity when new local fluid phase conditions are experienced. This is a consequence of the much shorter catalyst time scale relative to the corresponding one at which the transport of momentum, energy, and mass occurs. In this view, it is sufficient to compute beforehand the steady-state turnover frequencies (or reaction rates) for the current gas-phase conditions and to employ them as source terms into the governing equations at the macroscale. The approach has been demonstrated in the context of the integration of kMC calculations into reactor models [10,11] which are extremely challenging due to their stochastic nature and to the significant computational cost of the calculations preventing the direct coupling which is limited to specific reactor geometries and operative conditions [12–14]. In principle, the approach can be extended also to mean-field microkinetics of large dimensions and, thus, it has a strong potential in increasing the efficiency of the coupling between chemical kinetics and reactor models. However, two aspects have to be considered and addressed. On the one hand, the pre-computed data need to be properly interpreted and tabulated to move from discrete information to a continuous representation that is required for the coupling with the reactor models. The tabulation algorithm is the enabling factor and it assumes a crucial role for the overall computational efficiency of the approach. On the other hand, the records in the pre-computed dataset can easily become extremely numerous when dealing with large (e.g., number of independent variables) kinetic schemes. Hence, the design, generation and management of the pre-computed data is a crucial task for the computational efficiency and the accuracy of the algorithm predictions.

In terms of tabulation and interpretation of the precomputed data, several methodologies have been introduced in the literature. For instance, splines have been proposed in the context of mean-field approximation due to the capability in providing accurate predictions through simple polynomial-like expressions initially by Votsmeier and co-workers [15–18] and later by Dixon and co-workers [19,20]. The application of spline-like methods to the tabulation of kMC data is, however, challenging mainly due to the abrupt activity changes over a narrow range of gas-phase conditions as shown by Matera et al. [10]. In this view, a modified Sheppard interpolation method [10,11,21] has been employed to improve prediction accuracy. All these methods suffer from a loss in both accuracy of the approximation and computational efficiency as the number of variables increases. Votsmeier and co-workers [15] showed that the number of spline coefficients reaches such values which require a large amount of memory for the storage, thus hampering the applicability even to simple systems. Hence, their application is currently limited to schemes of reduced dimensions.

As an alternative, Machine Learning (ML) techniques can overcome such limitations because they are specifically conceived to work with highly dimensional datasets in terms of both numbers of records and variables [22–31]. Among the several available methods, ensemble learning methods and artificial neural networks (ANN) [32] have been proposed for the effective tabulation and regression of reaction rates and turnover frequencies (TOFs). In particular, ensemble learning methods such as Random Forest (RF) [33] have been recently employed by Dixon and co-workers [34] as an effective solution for the tabulation of mean-field microkinetic models able to overcome the limitations imposed the interpolation techniques.

Regardless of the specific method employed for the tabulation and interpretation of the pre-computed rates, a crucial problem is related to the generation of the dataset. A conventional approach in the context of mean-field models is to consider evenly distributed points in each direction of the space [16,34]. In doing so, the training points are positioned following a geometrical criterium without any information on the actual behavior of the function. However, large datasets (i.e., hundreds of thousands of records) are required to achieve a sufficient quality of the predictions of the models [35] with the consequential non-trivial management of the data in terms of memory and storage [15,19]. Moreover, the generation of such a dataset can be extremely demanding, especially in the case of kMC calculations, thus resulting in a very high computational burden for the build-up of the training set. Another possibility for the generation of the multidimensional grid is through the adaptive sparse grid method, widely used in the context of the solution of partial differential equations. However, even this method is affordable in moderately high-dimensional parameter spaces [36,37]. As the number of independent variables increases, this aspect can become a strong limitation to the applicability of the envisioned approach.

In this work, we propose a methodology to generate the training dataset by an iterative procedure able to minimize the number of data points required to achieve a target level of accuracy. The procedure combines the capability of the RF algorithm to quantify the importance of each variable with an iterative addition of new training points based on the discrete gradient of the TOF function. As such, only the relevant variables are selected and employed for the generation of the training set. Moreover, the composition space is sampled following the TOF function trend. Consequently, the positions of the training points are designed to gather where the approximation of the function is more difficult (e.g., regions of sharp transitions of the activity). On the one hand, the number of training points required to accurately approximate the TOF function is minimized resulting in a consistent saving in the computational cost associated with the RF training without any penalty in the regression accuracy. On the other hand, the position of the training points is defined by the function trend increasing the effectiveness of the approximation and, in turn, of the accuracy of the method. We exemplify the potential of the procedure by considering two distinct kinetic models characterized by, on one side, a large number of variables and, on the other side, a significant computational cost. Then, the obtained RFs are assessed by comparing their performances in reactor models against the solution of the full kinetic scheme. Finally, an example of application to unsteady CFD simulation in

complex geometry is reported to show the capabilities of the methodology in the context of three-dimensional reactor modeling. As a whole, the success of the ML approach to industrially relevant problems relies on the capability of generating of training sets able to allow for the learning of the complex features of the real function. Here, we have presented and developed a procedure in the context of the tabulation of first-principles kMC models. On a more general perspective, this problem of optimizing the generation of the training data of a Machine Learning algorithm is very relevant for catalysis at large. Therefore, this concept can be extended, for example, to the reduction of the computational cost connected to first-principles calculations for the screening of new materials [27–29,38,39] or the generation of learning potentials and forces for molecular dynamics [40,41].

2. Methods

2.1. Random forest and ExtraTrees

Machine learning methods have been conceived to perform efficient and accurate inductive learning. This means that the algorithm seeks to infer general trends based on a selected number of training data. Random Forest (RF) [33] is an ensemble machine learning method employed mainly for classification or regression problems. As a tabulation/interpolation method, RF presents the capability to efficiently handle large data sets with high numbers of input dimensions (i.e., independent variables), named descriptors. Here, we tabulate simulation data from complex kinetic systems, e.g., mean-field or first-principles kMC, which depend on several descriptors, such as the partial pressure and the temperature. RF shows good tabulation and regression performances even by dealing with relatively small training datasets [42]. Furthermore, RF offers some unique features that make it suitable for the tabulation of kinetic data. These include built-in estimation of prediction accuracy and measures of descriptor importance, allowing for ranking the input features according to their relevance for the output. In principle, this enables the reduction of the size and of the computational cost of the training set by excluding the dependence from irrelevant variables. In this work, an improved version proposed by Geurts et al. [42] of the Random Forest algorithm (named ExtraTrees - acronym for ExTremely Randomised Trees) is employed to improve both prediction accuracy and computational performances. ExtraTrees has been shown to be effective in dealing with large multidimensional complex problems providing less variance than conventional RF [42]. Moreover, it outperforms conventional decision trees when irrelevant descriptors (i.e., ineffective on the prediction) are present [42]. These two features help in dealing with the tabulation of first-principles kinetic schemes. They, in principle, depend on many variables, e.g. partial pressure, adsorbed species, temperature; however, only few of them are often important for the evaluation of the rates.

ExtraTrees consists of a multitude (n_{dt}) of decision trees (T) generated at training time $\{T_1(\boldsymbol{x}), \cdots, T_{n_{dt}}(\boldsymbol{x})\}$where $\boldsymbol{x} = \{x_1, \cdots, x_q\}$ is a q-dimensional vector of descriptors (i.e., variables or features) needed to compute the output (y) of the model. Given a training set of n_p data points $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_{n_p}, y_{n_p})\}$ that each span over $q + 1$ dimensions where $\boldsymbol{x}$ represents the descriptors, e.g., partial pressures of the species, and y the output, i.e., the associated reaction rate or TOF, the generation of each decision tree proceeds as follows [43]. A subset is extracted from the n_p data points by bootstrapping, i.e., randomly sampling with replacement, and used to grow a tree. In doing so, the predictor space is divided into N distinct regions through a recursive binary splitting approach. The data is initially split into two regions R_1 and R_2 at the split point s. In contrast with conventional decision trees, split points are drawn fully at random for each of the q predictors in this approach. Then, the cut-point s is defined by choosing among the possible split points the one which minimizes the residual sum of square evaluated between the system output and the average value computed in each region. Then, each resulting region undergoes the same recursive splitting procedure which is iterated until a termination rule is satisfied. In regression problems, the termination rule corresponds to a certain number of records in each terminal node. Once the ExtraTrees is grown, it is possible to evaluate the output given by a certain query. A query is a single data with q dimensions that is fed to each decision tree in the forest. By starting at the tree root node, the query is compared with the split criterion of each node descending the tree until a terminal node is reached. The predictions obtained from each decision tree are then aggregated using the arithmetic average, thus yielding to a prediction value for the entire forest. In this work, we employ a fast and efficient implementation of the algorithm provided within the *scikit-learn* Python library [44]. Despite the original ExtraTrees algorithm does not use bootstrapping, we employ a modified version of the ExtraTrees which allows for it. The generation of the ensemble of decision tree by using bootstrap aggregation reduces the correlation between the trees as they are each built with different training sets and reduces the overall variance of the final prediction improving the accuracy of the method [43].

The assessment of the accuracy of the predictions generated by the algorithm is required to understand the performances of the tabulation. Ideally, an independent and large dataset not used for the training should be used to quantify the quality of the predictions. In practice, the available amount of data is usually limited, and some type of cross-validation has to be employed. RF performs such an assessment through the Out-of-Bag (OOB) sample. Each tree is grown by using a bootstrap sample where some of the data points are not used to generate the trees and these left-out datapoints create the OOB sample. Since OOB data are not used to grow the tree, they can be employed to estimate the prediction performances as the mean square error (MSE):

$$OOB_{err} = \frac{1}{n_{dt}} \sum_{k}^{n_{dt}} \sum_{i}^{n_{OOB}} \frac{1}{n_{OOB}} (\hat{y}_{i,k}(x) - y_i(x))^2 \quad (1)$$

where n_{dt} is the number of tree in the forest, n_{OOB} represents the number of OOB samples. We devote ~30% of the entire training dataset to the evaluation of the OOB [33].

Random Forest and ExtraTrees also show the capability of quantifying the importance of the descriptors returning a measure of how each variable contributes to the accuracy of the predictions. In other words, a metric called variable importance is computed by the algorithm providing the effect of each of the descriptors on the predictions. Several methods have been proposed in the literature to describe the variable importance such as decrease of node impurity [45], corrected node impurity [46,47], and permutation importance [33]. Here, the non-scaled permutation is employed as a measure of the effect of the descriptors on the prediction since it is more robust in the presence of correlated variables [48]. In particular, the decrease of node impurity tends to be biased toward variables characterized by different scales or number of categories [49]. Moreover, Strobl et al. [50] showed that the permutation importance can clearly identify the irrelevant predictors which is a very welcome feature for the selection and identification of the important variables. In the permutation importance, the OOB error is computed twice. First, the OOB error is computed according to Eq. (1). Then, a modified OOB set is considered. Each descriptor in the OOB data is randomly permuted one at a time leaving all other descriptors unchanged and the modified set is predicted by the tree. The increment of the OOB error between the permuted and original dataset is a measure of the importance.

2.2. Kinetic Monte Carlo

We have performed kMC simulations within the graph-theoretical (GT) kMC framework of Stamatakis and Vlachos [51], as implemented in Zacros [52]. In this framework, the catalytic surface is represented as a two-dimensional lattice graph where the vertices are surface sites and the edges define the neighboring connectivity. The simulation input consists of the reactor conditions (temperature and partial pressure of

the species), the lattice structure that reproduces the catalytic surface, an energetic model which accounts for both the binding energies of the species and the adsorbate-adsorbate interactions and the reaction mechanism which describes the elementary events that might occur on the catalyst.

The simulations are carried out by initially considering an empty lattice to evaluate the turnover frequency of a generic molecule, *i*, which is defined as the number of molecules *i* produced per active site, per unit of time. To determine the TOF, the number of molecules *i* – produced during the simulation and filtered out by the initial transients – is fitted using linear regression. Then, the slope is divided by the number of lattice sites to eventually evaluate the turnover frequency (TOF). A simulation is deemed to be convergent when at least $5{\cdot}10^4$ molecules are produced and the slope is constant over time. The surface coverages (θ) are obtained by time-averaging the site occupancies once a steady-state is reached.

2.3. Reactor models

Two different reactor models are employed in this work to simulate the reactive systems. On the one hand, a 1D heterogeneous model and, on the other hand, a CFD model able to deal with arbitrary complex three-dimensional geometries.

2.3.1. 1D heterogeneous model

The 1D heterogeneous model is derived under steady-state conditions. The mass balance for the generic *i*-species can be expressed as follows:

$$-u\frac{d\rho_i}{dz} = k_{MAT,i}S_v(\rho_i - \rho_i^S) \tag{2}$$

where u is the superficial velocity, ρ_i is the partial density of component i in the bulk of the gas phase while ρ_i^S is on the catalytic surface, $k_{MAT,i}$ is the mass transfer coefficient and S_v is the specific surface area of the system.

The species mass balance on the catalytic surface is given by the following expression.

$$k_{MAT,i}S_v(\rho_i - \rho_i^S) = R_i(\rho_i^S) \tag{3}$$

where $R_i(\rho_i^S)$ is the net consumption or production rate of component *i*.

In this work, we consider a honeycomb monolith with square channels as an example of reactor geometry. Thus, the mass transfer coefficient is computed through the asymptotic Sherwood numbers ($Sh = 3.087$) [53].

2.3.2. Computational fluid dynamics

The description of the momentum, mass, and energy transport at the reactor scale are evaluated by numerically solving the conservation equation for momentum, mass, and energy for a multicomponent and compressible gas phase.

The motion of the fluids is described by the Navier-Stokes equations. In this regard, the continuity equation (conservation of mass) and the momentum balance may be written as:

$$\frac{d\rho}{dt} + \nabla{\cdot}\rho u = 0 \tag{4}$$

$$\frac{d(\rho u)}{dt} + \nabla{\cdot}\rho uu = -\nabla\ \mathrm{p} - \nabla{\cdot}\underline{\tau} + \rho g \tag{5}$$

where the stress tensor ($\underline{\tau}$) for Newtonian fluids is defined as:

$$\underline{\tau} = -\mu(\nabla u + (\nabla u)^T) + \frac{\bar{2}}{3}\mu(\nabla{\cdot}u)I \tag{6}$$

In the equations above, t is the time, p is the pressure a, ρ is the density, μ is the dynamic viscosity, $\boldsymbol{u}$ is the velocity vector and $\boldsymbol{g}$ the acceleration vector due to gravity.

The equation of conservation for the species is written as:

$$\frac{d(\rho\omega_i)}{dt} + \nabla{\cdot}\rho\omega_i u = \nabla{\cdot}(\rho\omega_i V_i) + R_i^{\mathrm{hom}} \tag{7}$$

where the subscript i refers to the individual gas-phase component i, and ω_i is the mass fraction. R_i^{hom} is the formation rate of species i in the gas-phase, and $\boldsymbol{V_i}$ is the diffusion velocity defined as:

$$V_i = -\frac{\Gamma_i}{\omega_i}\nabla x_i \tag{8}$$

To enforce mass conservation, the approach proposed by Coffee and Heimer [54] is applied. This method is based on a correction diffusional velocity, which replaces $\boldsymbol{V_i}$ in Eqs. (7) and (8) with $\boldsymbol{V_i^C}$, which is defined as:

$$V_i^C = V_i + V_C \tag{9}$$

where $\boldsymbol{V_C}$ is a constant correction factor introduced to satisfy the mass conservation and is evaluated as:

$$V_C = -\sum_{i=1}^{NG}\omega_i V_i \tag{10}$$

The conservation of energy reads as follows:

$$\rho c_p\frac{dT}{dt} + \rho c_p u\nabla\ \mathrm{T} = -\sum_i^{N_s}\rho c_{p_i}\omega_i V_i\nabla T - \sum_i^{N_s}H_i R_i^{hom} + \nabla{\cdot}(\lambda\nabla T) \tag{11}$$

where T is the temperature, λ and c_p are the thermal conductivity and the specific heat at a constant temperature of the gas phase mixture, respectively; c_{p_i} is the specific heat at constant pressure for species *i*, and H_i is the mass-specific enthalpy in the gas phase of species *i*. The density of the mixture is calculated according to the approximation of ideal gases.

The solution of the conservation equations requires to specify proper boundary conditions for all the dependent fields, i.e. pressure, velocity, temperature, and mass fractions. For chemical reactors, the usual boundary conditions for the pressure is a fixed value at the outlet and a zero-gradient at the reactor inlet and walls. The velocity profile is assigned at the inlet whereas at the outlet a zero normal gradient is imposed under the assumption of the fully developed flow field. The boundary conditions for the gas-phase mass fraction and temperature reproduce the feed at the conditions that are simulated. In doing so, the temperature and the mass fractions are imposed at the inlet by a Dirichlet condition (i.e., fixed value). The outlet section of the rector is usually placed where the temperature and concentration gradients have vanished allowing for a zero gradient boundary condition. At the inert reactor wall, the mass flux is zero since no formation or consumption of the specie is allowed corresponding to a zero-gradient condition. When the inert walls are assumed adiabatic, a zero gradient condition is imposed. Otherwise, a fixed temperature or a fixed heat flux is imposed.

The surface chemistry as a boundary condition at the catalytic surfaces where the mass flux of species *i* compensate for the formation/consumption rate R_i^{het} due to the heterogeneous reactions:

$$\rho\omega_i V_i{\cdot}n = R_i^{het} = \sum_k^{N_R}\nu_{i,k}\Omega_k^{het}MWM_i \tag{12}$$

where $\boldsymbol{n}$ is the inward-pointing normal vector, Ω_k^{het} are the net reaction rates (measured as turnover-frequency per unit time and surface area), $\nu_{i,k}$ are the stoichiometric coefficients of species i in reaction k and MM_i is the molecular mass.

In analogy, the heat flux on the catalytic surface has to compensate the heat released Q^{het} by heterogeneous reactions.

$$\lambda\nabla T{\cdot}n = Q^{het} \tag{13}$$

The solution of these equations is carried out through the finite volume method [55], implemented in *OpenFOAM* [56], through the *catalyticFOAM* framework, proposed by Maestri and Cuoci [6], which

can solve the Navier-Stokes equations for reacting flows at surfaces. In particular, the compressible form of the Navier-Stokes equations is coupled with the governing equations for the species which account for the heterogeneous chemistry. In this work, a second-order upwind scheme (linear upwind) is adopted for the discretization of the convective terms, whereas a second-order scheme is employed for the diffusion terms.

3. Adaptive design of training points

The quality of the predictions of every machine learning algorithm strongly depends on the original dataset employed for the training procedure. The design of the training points is a crucial task for a highly accurate approximation of the real unknown function. A simplistic approach is to evenly distribute the points in each direction of the multi-dimensional space [34]. In doing this, the points are assigned to every direction without accounting for the importance of the variable on the predictions and the different shape and slope of the function. Hence, the same number of points (n_p) are employed in each direction for the description of the real model. When dealing with kMC simulations, the evaluation of a single model (i.e., TOF) is computationally expensive resulting in a large burden that can become impractical especially when the dimensionality of the system is high. Therefore, an effective generation of the dataset is the key to overcome the curse of dimensionality. To solve this crucial issue, we propose an adaptive procedure able to selectively add data points within the multi-dimensional space in the regions where the approximation of the function is more demanding.

The algorithm works as follows. A flowchart of the procedure is reported in Fig. 1. A limited number of evenly distributed points, i.e., three or four for each space direction, are defined at the first iteration. The corresponding function value at each point is computed and an ExtraTrees is trained using this dataset. Then, the variable importances (v_i^S) are computed by using the non-scaled permutation method [33]. Hence, a quantitative measure of the effect of each predictor on the real function is evaluated. The directions of higher importance are then considered for the subsequent steps of further point additions to the original training set. Such directions are defined as the ones which have an importance higher than a user-defined threshold (v_i^{th}). In doing so, only the most important directions are considered with the advantage of improving the quality of the ML response without the addition of redundant points. Once the directions that need refinement are identified, it is necessary to define the positions of the new points. This is done by analyzing the slope of the function. Given one of the important directions, the intervals between the existent training points are identified. Then, the partial derivative ($d_{i,j}^S$) in each of the intervals with respect to the selected direction is computed keeping the other directions fixed. The middle point of each interval is a potential new point. To decide whether the point has to be added or not to the training set, the maximum absolute value of the derivative in the intervals is stored as a measure of the maximum rate of variation of the system in the current intervals. As such, each interval is classified according to the rate of variation of the function. The intervals where the derivative is higher than a user-defined threshold (d^{th}) are the ones where additional training data are required to properly follow the nature of the system. The model values in the new points are computed and added to the training set which is used to compute a new ExtraTrees based on the new dataset. The algorithm iteratively proceeds by adding points to the dataset until the required level of accuracy is reached.

The most appropriate way to quantify the accuracy of the ExtraTrees would be to compute many times the original model with random input in the variable space to obtain a benchmark dataset. In doing this, a set of data which has not been used during the learning phase could be employed to assess the ML predictions. This approach would involve the evaluation of the model many times with a significant computational cost devoted just to the validation of the algorithm with no additional benefits to the training of the ML. To overcome such a limitation, we exploit two different strategies. On the one hand, the OOB error is evaluated since it represents the accuracy of the ExtraTrees with respect to the data points not used for the training. The main concern related to the OOB is the low statistical significance for very small datasets, which are the ones we employ at the beginning of the procedure. In this view, the ExtraTrees is trained several times at each iteration providing several OOB errors. The average and the maximum among the OOBs are evaluated to quantify the accuracy of the predictions. On the other hand, the second level of assessment is achieved by comparing the predictions of the ExtraTrees trained with the dataset obtained in the present and past iteration. In particular, the new points added in this iteration are used as a query for both the ExtraTrees trained at the past iteration and at the current one. The mean relative error computed between the two results is defined as relative approximation difference (RAD) and reveals the effect of the new points computed at the current iteration on the algorithm predictions. Hence, the smaller is the RAD, the lower is the effect and importance of the newly added points because they have marginally

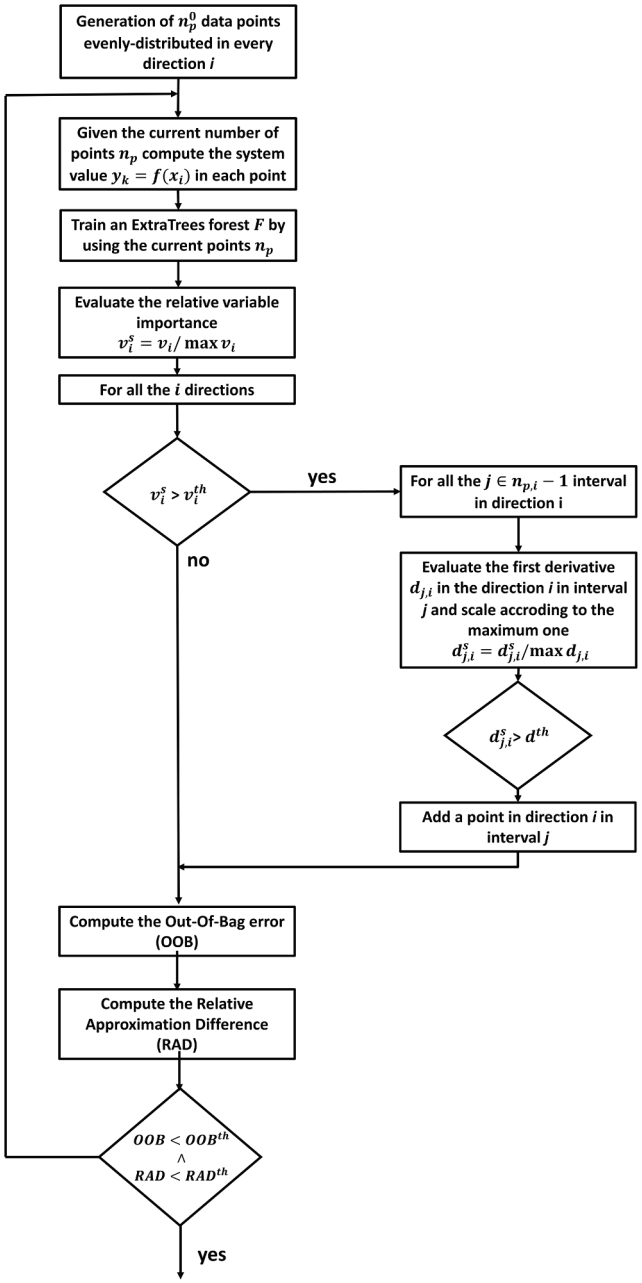


Fig. 1. Flowchart of the algorithm of the adaptive design procedure.

improved the ExtraTrees predictions and indicates that the current table is enough to describe the real model.

The OOB and the RAD are both used to end the procedure. In particular, the algorithm stops when the OOB normalized with respect to the one evaluated, and the first iteration ($OOB^{(1)}$) and the RAD are below a certain threshold. The normalized OOB and RAD are employed to control the performances of the procedure in terms of accuracy. Both the OOB and the RAD are related to the quality of the approximation since they represent an indicator of the deviations between the predictions from the ExtraTrees and the real function behavior. Moreover, we observed that the OOB and RAD are strongly related and numerically close to the benchmark error evaluated on a cross-validation dataset, as shown in Section 4. In this work, different termination criteria are employed in the examples of the application of the procedure to demonstrate the control of the output accuracy provided by their different values.

The definition of their values is crucial for the accuracy of the approximation but also for the computational burden related to the generation of the dataset. More stringent termination criteria would inevitably result in a larger number of training points and hence in higher computational cost spent in the construction of the ExtraTrees. As a result of the procedure, an ExtraTrees table representing the behavior of the real model is obtained by employing a dataset aimed at minimizing the number of training points. Hence, an accurate description of the multidimensional model is achieved minimizing the computational costs related to the evaluation of the real model.

The definition of the parameters of both the algorithm has been selected after a parametrical analysis to represent the best compromise between accuracy and dataset size as reported in the Supplementary Material (Section 1). Table 1 lists the algorithm parameters employed in this work. In particular, the terminal leaf size is defined to be 1 because the data employed in this work is not noisy, whereas the number of splitting variables is assumed to be equal to the number of variables as suggested by Geurts et al. [42]. The effect of the number of trees in the forest has been tested and we obtained that a number $\geq$ 200 is sufficient to achieve good results. The employed parameters are listed in Table 1.

4. Results and discussion

In this section, we illustrate the capabilities of the methodology for the generation of the training dataset for the ExtraForest algorithm through a simple showcase. Then, the adaptive design procedure described in Section 3 is employed for the tabulation of two different systems. The first-principles kMC CO oxidation on RuO_2(110) is considered to show the capabilities of the approach with respect to a system characterized by sharp transitions in reactivity. CFD simulations will be shown to demonstrate the possibility of introducing detailed kMC simulation in 3D reactor models. Then, a Water Gas Shift model on Rh computed through mean-field microkinetic model is considered to show the capabilities of the procedure in dealing with a reaction system characterized by high dimensionality. A PFR is employed to carry out a direct comparison between the direct solution of the kinetic model and the ExtraTrees predictions.

Table 1
Adaptive design procedure and ExtraTree parameters.

Algorithm parameters		ExtraTree parameters	
Param	Value	Param	Value
v_i^{th}	0.15	Number of trees	200
		Training fraction	0.7
d^{th}	0.5	Terminal leaf size	1
		Splitting variables	Number of variables

4.1. Showcase of the procedure

A mono-dimensional testing function is employed to elucidate the capability of the adaptive design procedure. The function is conceived to represents the trend of a typical kMC chemical model usually characterized by abrupt changes in catalytic activity [10]. In this view, the function reported in Eq. (14) is considered.

$$y = \frac{1}{x \cdot (1 + \exp(-150 \cdot (x - 0.5)))} \tag{14}$$

The function is able to show a sudden increase of the function value around $x = 0.5$, followed by a reduction of the function magnitude as shown in Fig. 2. The interval $x = [0.001, 1]$ is considered for the showcase. Moreover, a benchmark dataset consisting of 1000 points randomly distributed is evaluated to quantitatively assess the performances of the procedure. In this case, the termination criteria are the following $OOB/OOB^{(1)} = 0.015$ and $RAD = 15\%$.

The adaptive design procedure starts with 4 points equally distributed in the considered interval. An ExtraTrees is trained using these initial points, as shown in Fig. 2(a), and the OOB is computed. In the next iterations, the variable importance is evaluated, and the important directions are considered for the refinement. By evaluating the local derivative based on the actual function values, the intervals between consecutive points that require to be refined are selected. All the intervals are refined at the second iteration since the description of the function is poor by using the initial dataset. On the contrary, the intervals selected for the refinement in the next iterations are the one characterized by the rapid change of the slope of the function. The OOB and RAD termination criteria are reached after 7 iterations resulting in a training set of 24 points. The data points are amassed in the region of the composition space where the function shows the abrupt change in slope, as shown in Fig. 2(b).

The assessment of the accuracy of the predictions is carried out by direct comparison with the analytical function values (Eq. (14)). The approximation error is computed as the relative error between the real function value and the predictions of the ExtraTrees evaluated for each record in the benchmark dataset. The average and maximum errors are evaluated as reported in Eqs. (15) and (16).

$$\langle \varepsilon_{bench} \rangle = \frac{1}{N_Q} \sum_{i=0}^{N_Q} \left| \frac{y(x_{q,i}) - y_{ET}(x_{q,i})}{y(x_{q,i})} \right| \tag{15}$$

$$\varepsilon_{bench}^{MAX} = \max \left| \frac{y(x_{q,i}) - y_{ET}(x_{q,i})}{y(x_{q,i})} \right| \tag{16}$$

where $x_{q,i}$ is the query point from the benchmark dataset, $y(x_{q,i})$ is the actual function in the query point, $y_{ET}(x_{q,i})$ is the function value predicted by the ExtraTrees in the query point and N_Q is the size of the benchmark set.

A direct comparison with the benchmark dataset reveals the improvement of the error by introducing additional points according to the envisioned procedure, as shown in Fig. 2(a-b). The first set of points makes it possible to achieve a rough description of the real function due to the insufficient number of information provided to train the ExtraTrees. By the iterative addition of new points, the agreement improves since the ExtraTrees predictions move closer to the analytical function values. At the end of the procedure, the predictions of the ExtraTrees table are superimposed to the actual function values evaluated in the benchmark dataset, as shown in Fig. 2(b).

Fig. 3 shows the evolution of the benchmark error $\langle \varepsilon_{bench} \rangle$ with the iterations. The error decreases with the iterations due to the higher number of training points. At the final iteration, the average deviation is below 1% while the maximum one is < 10% and located in the region of the abrupt change in slope.

It is worth comparing the accuracy achieved through the design procedure against an evenly-distribute dataset. Fig. 3 reveals that the

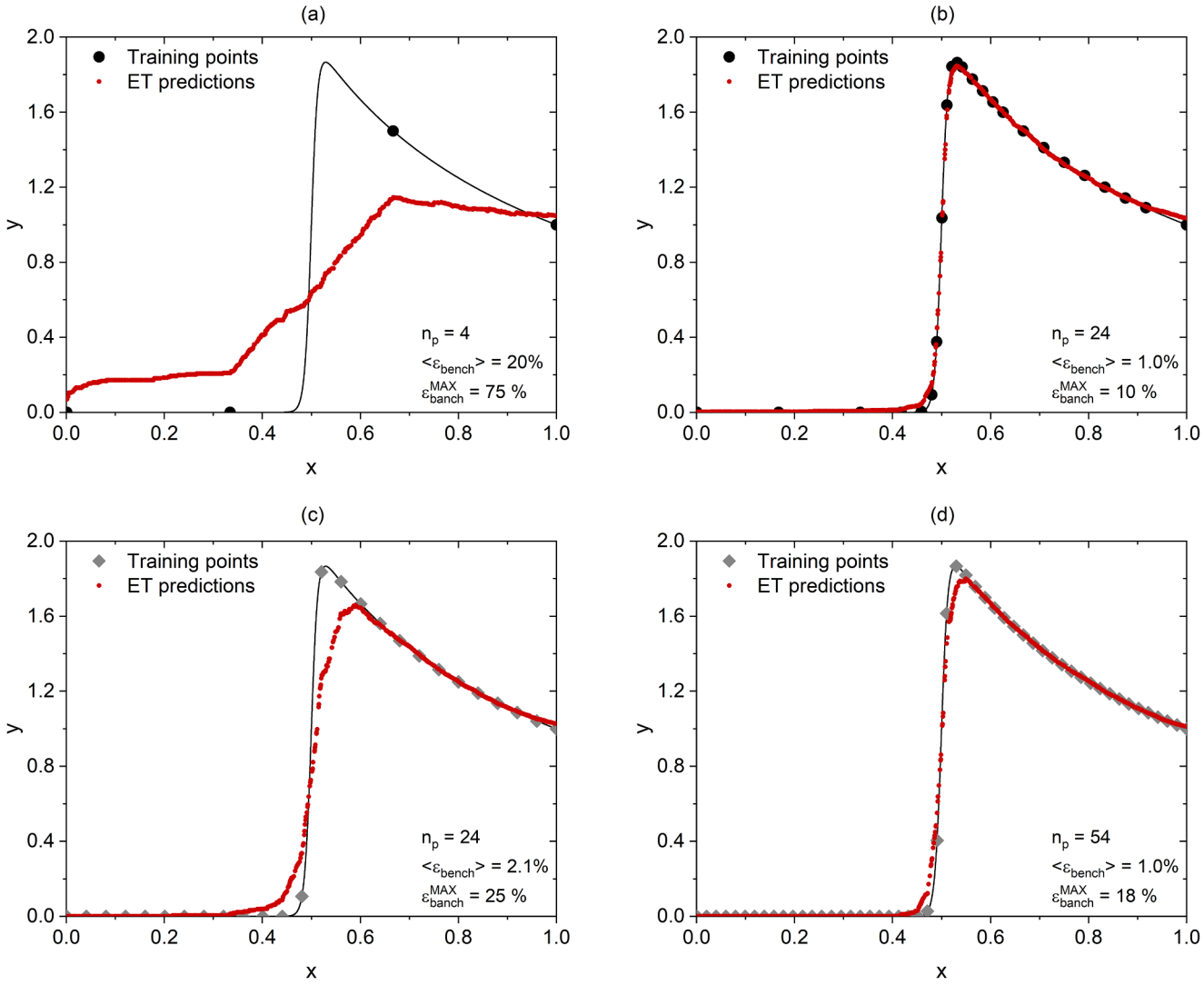


Fig. 2. ExtraTrees (ET) predictions (red dots) compared with the real function values (black continuous line) obtained at the first (a) and last (b) iteration of the advanced design procedure and for evenly distributed grids with the same number of datapoints obtained at the end of the adaptive procedure (c) and with the number of training points required to achieve the same approximation accuracy of the design procedure (d). The training points are also shown in each panel: black circles correspond to the adaptive procedure whereas grey diamonds to equally spaced grids. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performances of the dataset obtained by the adaptive design procedure always outperform the ones obtained with an evenly distributed grid as shown by the higher average benchmark error. This proves the effectiveness of the procedure in the choice of the position of the training points. Indeed, the addition of new data points in the proper positions selected by the algorithm enables a rapid reduction of the benchmark error with a significantly higher rate with respect to the equally spaced cases. In particular, a uniform grid with 24 points (i.e. the same obtained by the iterative procedure) achieves a $\langle \varepsilon_{bench} \rangle$ equal to 2.1% and a maximum deviation around 25%. The average and maximum benchmark errors are 2 and 2.5 times, respectively, larger than the one obtained with the training set generated by the adaptive procedure. Fig. 2 (c) shows the ExtraTress predictions for the 24 evenly distributed points. The rapid variation of the function values around x = 0.5 is poorly described despite the significant number of training points. This is ascribed to the positions of the data points which are mainly used to describe the quasi-flat region before x = 0.5 and the smooth decrement after x = 0.6. The region of the sharp variation of the function value is described by only 4 points whereas the adaptive procedure employs 7 points in this region. Fig. 2(d) also shows that the same average accuracy is reached by employing more than the double the number of evenly distributed points (i.e., 54) but with still a higher maximum error (18 vs 10%), since the region around x = 0.5, characterized by a sharp change in the function slope, is still poorly described.

4.2. Application of the procedure to the tabulation of first-principles kMC data

The procedure for the generation of the training points has been assessed by using a simple first-principles kMC model to showcase the capability of the approach in dealing with such systems. The CO oxidation on RuO_2(110) proposed by Reuter and Scheffler [57] has been employed. The model employs a lattice representation of the active surface considering different site types, i.e. bridge and cus. In these simulations, a lattice consisting of 20x20 surface sites (200 bridge and 200 cus sites) and periodic boundary conditions are employed as proposed by Reuter and Scheffler [57]. The kMC model accounts for all the elementary events that can occur on the lattice: dissociative O adsorption, associative O desorption, unimolecular CO adsorption, and desorption along with Langmuir-Hinshelwood CO + O surface

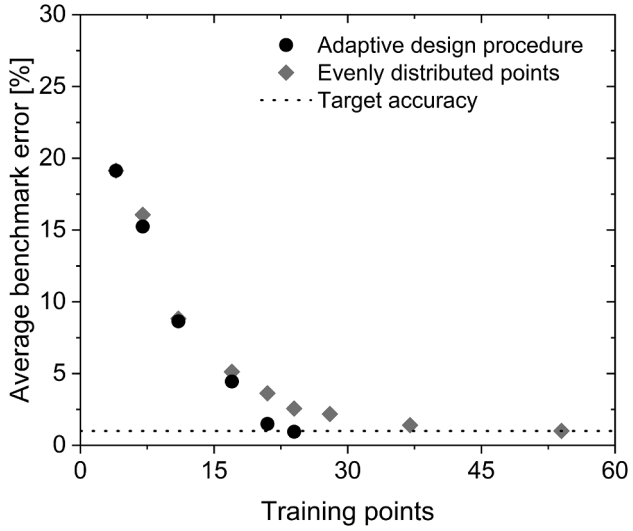


Fig. 3. Average benchmark error for the showcase system as a function of the training points for each iteration of the procedure (black circle) along with the error achieved by means of evenly distributed sets of points (grey diamond). The dotted line highlights the target accuracy of the procedure.

Table 2
Range of partial pressures and temperature employed for the tabulation of the CO_2 net production rate in the first-principles kMC CO oxidation on RuO_2(110).

Variable	Range
CO	$1{\cdot}10^{-6}-1{\cdot}10^{-1}$ bar
O_2	$1{\cdot}10^{-6}-1{\cdot}10^{-1}$ bar

reactions. The CO_2 re-adsorption is neglected along with the superficial diffusion of CO and O due to the high partial pressure of the species considered [57]. The first-principles kMC model has been implemented in the Zacros package. For a given set of operating conditions (i.e., partial pressures of CO, O_2 and temperature), the outcome of a first-principles kMC simulation is the steady-state catalytic TOF and the corresponding coverage distribution. The range of operating conditions investigated is reported in Table 2, while the temperature is assumed to be constant and equal to 600 K. In the case of the first-principles kMC data, the TOF and the corresponding coverages are tabulated through the ExtraTrees. In doing this, the coupling between the chemistry and the numerical simulations at the reactor scale is possible without losing the relevant information of the status of the catalytic surface provided by the Monte-Carlo simulations. The termination criteria employed in this case are the following: $OOB/OOB^{(1)} = 0.1$ and $RAD = 10\%$.

First, the procedure is employed for the tabulation of the TOF resulting from the kMC simulations. Fig. 4 depicts the three-dimensional representation of the actual TOF obtained from the 250 benchmark simulations. The net production rate shows a low value of the TOF ($< 1\ s^{-1}$) in a large region of the composition space. This is due to the poisoning of the surface, which is fully covered by a single species, resulting in an inhibition of the catalytic activity. However, the system has an abrupt change in reactivity in the region characterized by a ratio between the partial pressure of CO and O_2 between 10 and 100 where the TOF reaches values higher than $10^4\ s^{-1}$ [10,57]. A good approximation of the function requires an effective distribution of the training points rather than an evenly distributed grid. The initial training set consists of 3 equally spaced points for each direction sampled on a logarithmic scale for the partial pressures as shown by the blue circle in Fig. 4(a). It is worth emphasizing that the actual TOF trend is generally not known during the generation of the training dataset. Despite the poor description of the function, the initial data points are sufficient to evaluate the discrete gradient. In doing this, it is possible to discover the regions of the composition space where the steeper variations of the function are present. Fig. 4(b) shows the distribution at the second iteration, where the red circles represent the additional points, which are still insufficient to provide an accurate approximation of the function. In the next iterations, additional points are added based on the gradient of the function. Hence, they are positioned around the peak of the TOF, thus providing an additional improvement with respect to a blind addition. By proceeding with the iterations, the procedure expands the dataset, thus gathering the points in the region of the composition space corresponding to the steeper variation of the TOF. Fig. 4(d) represents the situation at the final iteration when the termination criteria are satisfied. The final dataset corresponds to 140 points distributed in 10 and 14 for CO and O_2, respectively.

The accuracy of the resulting ExtraTrees is assessed by considering a benchmark dataset generated by considering 250 randomly generated points in the range of partial pressure of interest. The average benchmark error ($\langle\varepsilon_{bench}\rangle$) computed in the case of the adaptive procedure grid is equal to 12%. Then, we evaluated the accuracy of two evenly distributed grids. First, we considered 144 (i.e., 12 for CO and 12 for O_2) equally spaced points that correspond to the same amount of data obtained by the adaptive procedure. In this case, we observed an average benchmark error of 14%. Hence, the adaptive procedure can provide a superior approximation with respect to evenly distributed grid even in a simple system characterized by two variables. The same level of accuracy is obtained by considering a uniform grid consisting of 20 points in each direction (400 training data) which reaches a $\langle\varepsilon_{bench}\rangle = 11\%$, analogous at that of the design procedure. The same level of accuracy is achieved by the design procedure by considering 65% points less than the evenly distributed grid, resulting in a consistent saving in computational time. Moreover, the computational gain is expected to be more significant by moving to systems either with a higher number of species and/or more complex chemical reactivity.

We have also carried out a more detailed comparison of the performances of the two training sets on the model predictions by considering the outcome of every single query. Fig. 5 shows a parity plot where the TOFs evaluated through the ExtraTrees are compared to the actual values resulting from kMC simulations. The TOF values are in good agreement with deviations generally below 15% in a broad range of operating conditions. However, Fig. 5 reveals that the predictions obtained using the ExtraTrees corresponding to the 144 evenly distributed dataset are generally more scattered. It can be noticed that the largest deviations are observed for low values of the TOF ($< 10^{-2}\ s^{-1}$). Despite the error in this region can be relevant ($> 70\%$), the error on the TOF results in an insignificant error in the reactor output concentrations. Fig. 5 (b) reports a zoom in the region of high TOF ($> 10^3\ s^{-1}$)). Hereby, the predictions of the adaptive design procedure are close to the ones of an evenly distributed grid with 400 points and generally fall between the $\pm$ 15% range. Conversely, the grid with 144 points is characterized by larger deviations from the parity line due to the worse description in particular of the region of high reactivity. Most of the points are outside the 15% of the kMC values and the predictions are generally more distant from the parity line than the ones of the forest trained with the same number of points positioned by the adaptive design procedure.

It can be noticed that the ExtraTrees suffers from the problem of generally underestimating the peak of the TOF. This is ascribed by the fact the ExtraTrees is able to predict at best an average of the data employed in the training. The peak values of the TOF are not guaranteed to be included in the training data, thus those sharp function peaks require extrapolation to be properly described since their value is outside the range withing the ExtraTrees is trained. ExtraTrees is not able to extrapolate outside the training set neither in the independent variable domain nor in the dependent variable domain resulting in an inaccurate description in this region. As such, the design of the training points reveals to be even more crucial since an accurate definition of

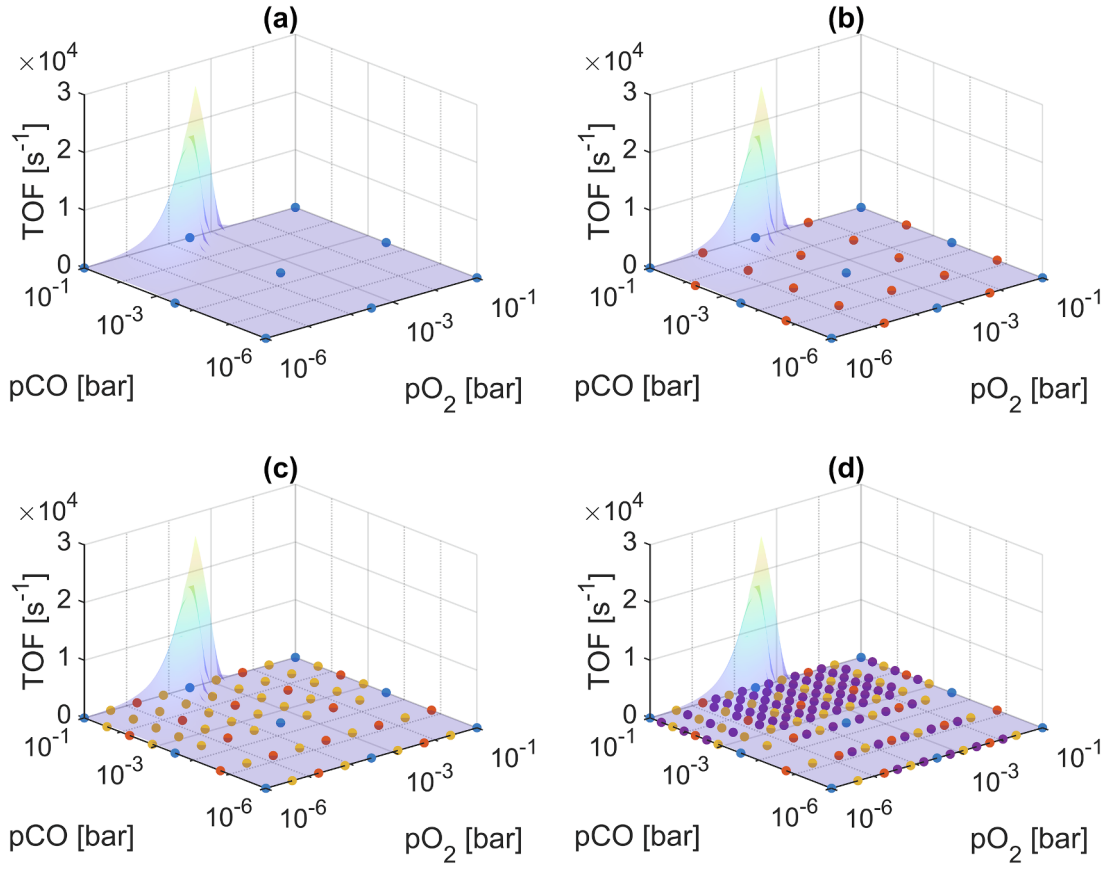


Fig. 4. Evolution of the dataset employed for the training of the ExtraTrees in the case of the CO oxidation over RuO_2 along with the iterations necessary to reach the target accuracy from the first iteration (a) to the final iteration (d).

the training set positions is also able to boost accuracy.

In principle, ExtraTrees can learn and regress whatsoever function. In this view, it is possible to tabulate also other outputs of the kMC simulations, such as the coverages (θ) or even the coverage speciation on the different active sites. To show the potential of the approach, we have also evaluated the performances of the ExtraTrees in tabulating the site coverages coming from the kMC simulations. To do this, we post-processed the kMC simulations and we computed the time-averaged site occupancies once steady-state conditions are reached. According to the kinetic scheme hereby considered, the adsorbed species are the CO, named CO*, and the atomic oxygen, named O*.

Fig. 6 shows the parity plot of the coverages (θ) between the ExtraTrees predictions and the kinetic Monte Carlo simulations. The vast majority of the coverages computed with the advanced design procedure are within 15% of the values evaluated with kMC and comparable results are achieved through the evenly distributed grid with 400 points. By considering the ExtraTrees trained with the dataset from the adaptive design procedure, a benchmark error around 15% and 1.5%

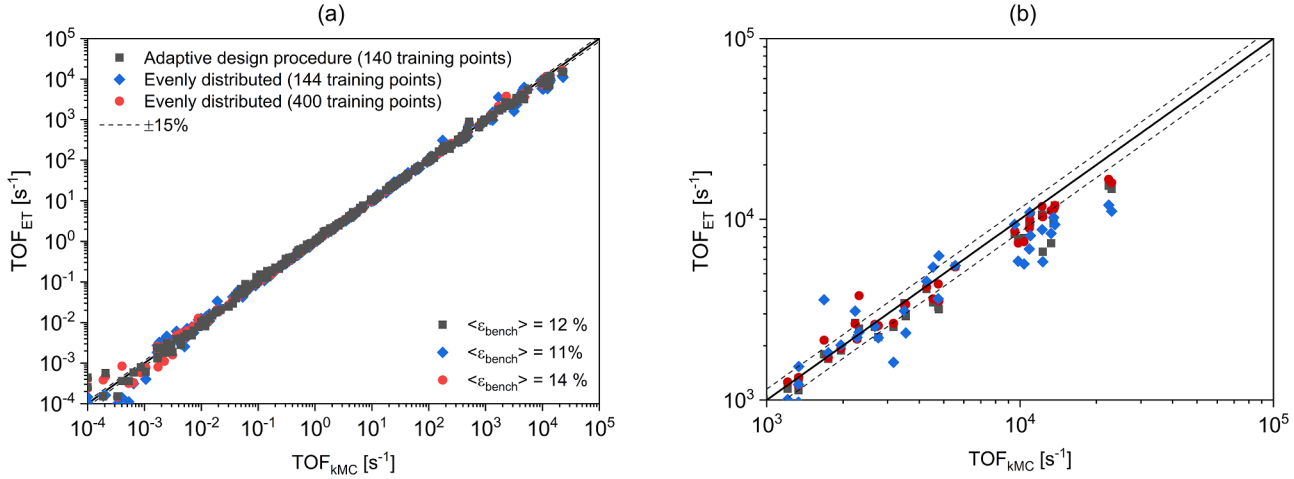


Fig. 5. Parity plot comparing the TOF from kMC simulations with the predictions of the ExtraTrees (ET) trained with the dataset generated by the adaptive design procedure (grey squares) and by evenly distributed grids with 144 (blue diamonds) and 400 training points (red circles) (a) along with a zoom in the region of high TOF (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

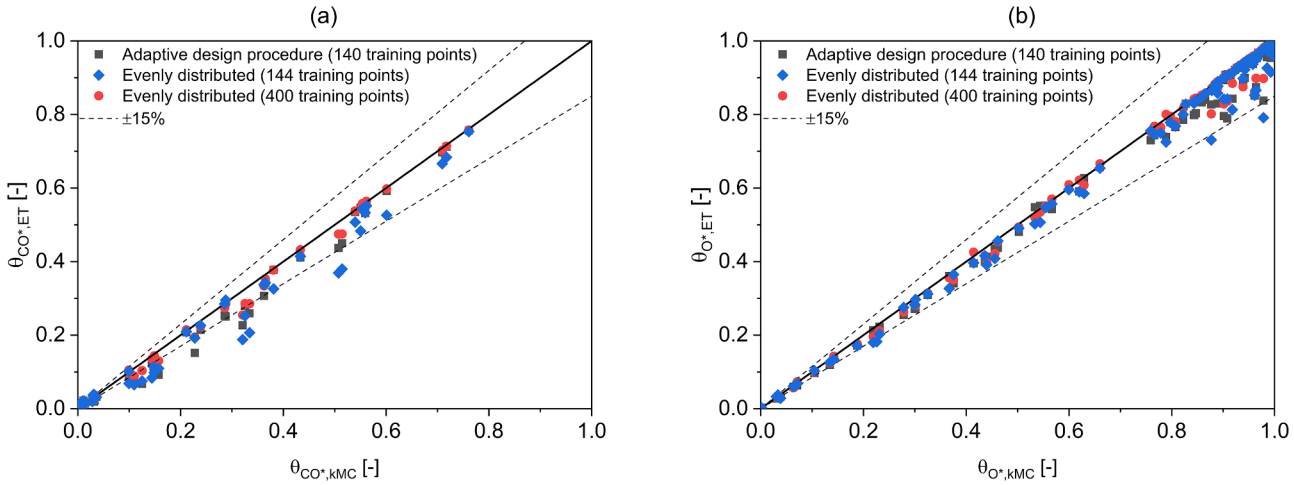


Fig. 6. Parity plot comparing the coverages (θ) from kMC simulations with the predictions of the ExtraTrees (ET) trained with the dataset generated by the adaptive design procedure (grey squares) and by evenly distributed grids with 144 (blue diamonds) and 400 training points (red circles) for CO* (a) and O* (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are evaluated for CO* and O*, respectively. At the same time, the evenly distributed grid with 144 points is characterized by $\langle \varepsilon_{bench} \rangle$ equal to 23% and 1.7% for CO* and O*, respectively. This reveals the superior accuracy achieved employing the advanced design procedure with respect to an equally spaced grid even in this simple system. A similar level accuracy is achieved by considering the 400 equally spaced points ($\langle \varepsilon_{bench} \rangle$ equal to 10% and 1.4% for CO* and O*, respectively) corresponding to a significant increment of the computational cost. Thus, both CO* and O* are well predicted by the ExtraTrees trained with the adaptive procedure elucidating the capability of the ExtraTrees of tabulating not just the TOF but also other relevant chemical important properties such as the coverages.

In doing this, the relevant information provided by kMC is not lost in the coupling with reactor scale simulations since they are properly provided by a dedicated tabulation procedure. Moreover, the adaptive design procedure could be also employed leveraging from the current grid obtained for the description of the TOF to be eventually refined to improve the prediction of the other quantities of interest, such as the coverages.

Finally, the analysis of the computational cost related to the evaluation of the TOF through the ExtraTrees is compared to the computational cost required by the solution of the kMC simulations. The effort for a kMC simulation strongly depends on the operating conditions. On average, the computational cost required by a kMC simulation of CO oxidation over RuO_2(110) is around 900 s on an Intel(R) Xeon(R) Gold 6148 2.40 GHz. On the contrary, ExtraTrees is able to provide the estimation of the TOF in about 10 ms on the same machine resulting in a significant speed-up. As a result, the adaptive training procedure is able to reduce the computational cost related to the generation of the training set, whereas the ExtraTrees is capable of dramatically reducing the burden related to the prediction of the TOF, thus enabling the coupling of kMC simulation in macroscopic reactor models.

4.3. CFD simulation of kMC CO oxidation on RuO_2(110) in a complex geometry.

We integrate first-principles kMC based heterogeneous chemistry models in 3D real reactor models by coupling CFD and ML methods. In this view, complex and tri-dimensional geometry of an open-cell foam has been considered to elucidate the capability of the envisioned approach to enable the accounting for the the transport phenomena along with the accurate description of the surface reactivity. Moreover, we emphasize how the intrinsic catalytic activity can be captured through first-principles kMC simulations carried resulting in non-intuitive effects at the reactor level. In doing this, the RuO_2 model catalyst surface previously precomputed and tabulated in an ExtraTrees (trained with the proposed adaptive procedure) is employed.

The geometry considered for this showcase is a 9 mm tubular reactor filled with a catalytic open-cell foam. The structure is 1.0 cm length and it is placed 2 mm downstream of the inlet. The open-cell foam surface is assumed to be catalytically active. Due to the symmetry of the system, a quarter of the reactor tube is simulated. The open-cell foams geometry shows a void fraction of the 90% and a cell size of 3.2 mm and it has been generated according to the procedure reported by Bracconi et al. [58]. The mesh has been generated through the *snappyHexMesh* utility of the *OpenFOAM* framework [56]. A background mesh with a resolution of 0.16 mm has been employed and the region of the mesh close to the foam surface has been refined up to the three levels obtaining a cell size equal to 0.02 mm. In doing so, the computational domain is refined in the region of the mass and temperature gradients within the boundary layer according to the mesh convergence results already reported in literature [59]. The simulation is carried out in isothermal conditions. The feed conditions correspond to $x_{CO}/x_{O_2} = 5$, $p = 1\ bar$, $T = 600\ K$ and a flowrate of 0.95 $NL\ min^{-1}$. The simulation is carried out for 0.2 s (equal to 5 residence times) with a constant time step of $1.5 \cdot 10^{-6}$ s resulting in a Courant number below 0.08.

The precomputed ExtraTrees table enables us to provide the TOF required by the CFD simulations to accurately describe the heterogeneous chemistry in a reasonable computational time. The solution of the surface reactivity in each computational cell requires the evaluation of the TOF a large number of times (i.e., 10^2-10^3). The fast and accurate estimates provided by the ExtraTrees strongly reduces the computational costs associated with the evaluation of the reaction rates by orders of magnitude if compared to actual kMC simulations. Hence, an effective coupling between CFD and kMC is achieved with a reasonable computational burden.

As an example of the performances of the coupling, the steady-state simulations results are reported in Fig. 7. The reacting mixture enters into the reactor and once approaches the catalyst the CO and O_2 are consumed to produce CO_2, as shown in Fig. 7(a). The CFD simulations can catch the spatial variations of the gas and adsorbed species. Moreover, the complex interplay between the transport properties and the chemistry results in reaction rates which might be different even by an order of magnitude for adjacent foam surface regions, as shown in Fig. 7(b).

The CO* and O* site fractions at the catalytic surface are reported in Fig. 8. The envisioned approach is also able to provide an insight into

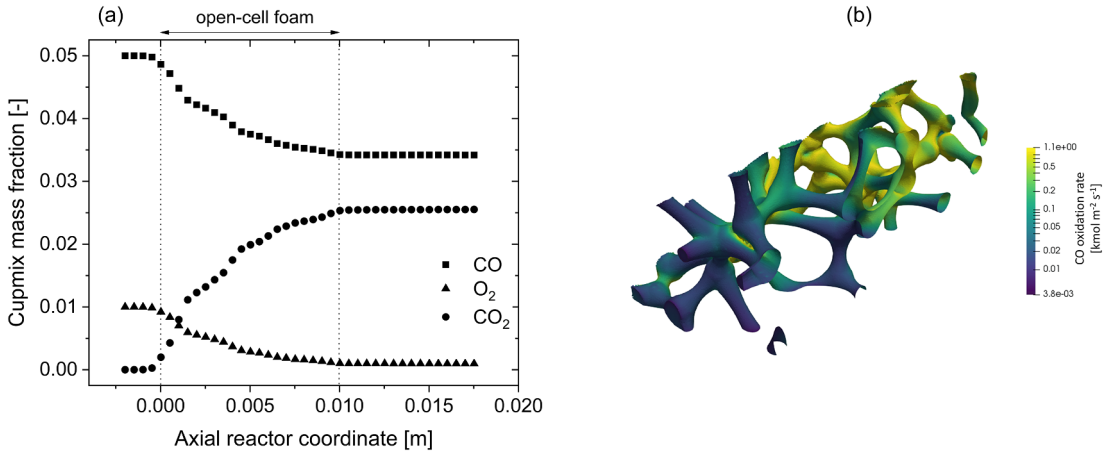


Fig. 7. Cupmix mass fraction along of the reactor axis for CO (square), O_2 (triangle), CO_2 (circle) (a) and local reaction rates on the foam surface (b).

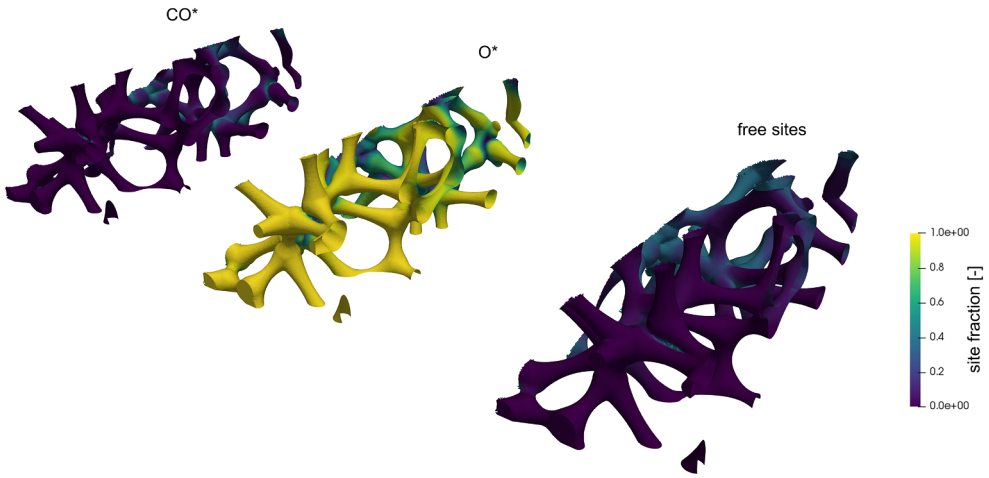


Fig. 8. Site fractions maps on the catalytic surface for CFD-kMC simulation of CO oxidation on $RuO_2(110)$ in isothermal conditions (T = 600 K) on an open-cell foams.

the status of the catalytic surface during the simulation revealing the different adsorbed species and their amount. In particular, O* is the most abundant adsorbed species at the beginning of the reactor where molecular oxygen is present in the gas phase consistently with the results by Reuter and Scheffler [57] for this operating condition. Where oxygen tends to be highly depleted, the catalytic surface begins to be partially covered by CO* as shown on the left-hand side of Fig. 8. As a whole, the adaptive design procedure and ExtraTrees are capable of enabling the coupling between first-principles kinetic schemes and complex tridimensional CFD simulations paving the way for the detailed description of the mutual interaction between the catalytic chemistry and the transport.

4.4. Application of the procedure to a large dimension system

Machine Learning methods and the adaptive design procedure allow for the tabulation of the computationally expensive first-principles kinetic models shown in Section 4.2. Furthermore, they enable the facile management of a system characterized by a high number of variables. To this scope, here we test the adaptive procedure by using a mean-field water gas shift (WGS) microkinetic model on Rh [60], as an example of a reactive system of high dimensionality. This test makes it also possible to assess a direct comparison between the microkinetic model and the trained Machine Learning algorithm (ExtraTrees) in reactor simulations.

The WGS system is characterized by four species, i.e. CO, H_2O, CO_2, H_2, and temperature, resulting in five descriptors and it is represented by the mean-field microkinetic model proposed by Maestri et al. [60]. The net production rate of CO_2 is the quantity considered for the tabulation with the ExtraTrees techniques. A set of partial pressures of the gas species which span the range reported in Table 3 is employed for the generation of the training data. In particular, the microkinetic model is solved for any given set of partial pressures and temperatures to evaluate the distribution of the site fractions and the net production rate of the target molecule [34]. In this case, the termination criteria are the following $OOB/OOB^{(1)} = 0.05$ and RAD = 2.5%.

The initial training set consists of 3 evenly distributed points for each direction sampled on a logarithmic scale for the partial pressures and a linear scale for the temperature. The procedure is deemed to provide an accurate result when the OOB in terms of mean relative

Table 3
Range of partial pressures and temperature employed for the tabulation of the CO_2 net production rate in the WGS system.

Variable	Range
CO	$1{\cdot}10^{-2}-1{\cdot}10^{-1}$ bar
H_2O	$1{\cdot}10^{-2}-1{\cdot}10^{-1}$ bar
H_2	$1{\cdot}10^{-6}-1{\cdot}10^{-3}$ bar
CO_2	$1{\cdot}10^{-6}-1{\cdot}10^{-3}$ bar
T	650–900 K

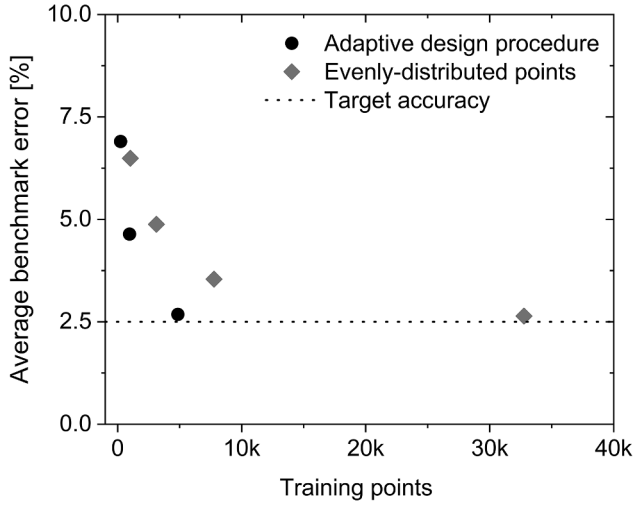


Fig. 9. Average benchmark error for the WGS system as a function of the training points for each iteration of the procedure (black circle) along with the error achieved by means of evenly distributed sets of points (grey diamond). The dotted line highlights the target accuracy of the procedure.

error is around 2.5% which is expected to provide the same average benchmark error. To assess the accuracy of the procedure, a benchmark dataset is also computed by generating 10^4 random sets of composition and temperature in the range of interest.

Fig. 9 shows the evolution of the average benchmark error as a function of the number of training points along with the target error represented by a dotted line. Each circle corresponds to an iteration of the adaptive design procedure of the training points. Based on the initial dataset, the procedure begins with the tabulation of 243 (3^5) data points in an ExtraTrees table which provides a rough estimation of the target production rate, with an average benchmark error of around 7%. Then, additional points are sequentially added according to the envisioned procedure in three further iterations. As expected, the average benchmark error (Eq. (15)) decreases by increasing the data points with a quasi-exponential fashion and reaches the grid reaches the termination criteria by considering 4860 points. More stringent termination criteria would lead to a different amount and distribution of the training points and a concomitant lower benchmark error.

The accuracy of the procedure is assessed by computing the average benchmark error which is equal to 2.7%. The final point distribution is uneven as listed in Table 4 along with the evolution of the number of training points at each iteration. Temperature shows the highest number of points, followed by water, hydrogen, and carbon monoxide, whereas a negligible effect is observed for carbon dioxide. These results reflect the intrinsic kinetics of the system which can be macroscopically described as a first-order reaction rate with respect to water with inhibition through the most abundant reaction intermediates which are carbon monoxide and hydrogen [61]. It is worth noticing that the procedure can select the most relevant descriptors avoiding the addition of points for irrelevant variables in the conditions of interest, such as CO_2. The negligible effect of the CO_2 can be explained by the operating conditions which are distant from thermodynamic equilibrium resulting in the presence of the sole forward water-gas-shift.

Table 4
Distribution of the number of training points at each iteration of the adaptive procedure for each of the descriptors involved in the WGS system along with the total number of points.

Iteration	CO	H_2O	H_2	CO_2	T	Total
#1	3	3	3	3	3	243
#2	4	4	4	3	5	960
#3	6	6	5	3	9	4860

The capability of the procedure in reducing the number of training points is assessed with a comparison with evenly distributed points in all the directions. First, we select a number of equally spaced points in each direction to make the grid dimension as similar as possible to size of the training set from the procedure. The final number of points falls in between an evenly distributed grid with 5 and 6 points, corresponding to 3125 and 7776 data points, respectively, as shown in Fig. 9. Hence, we compared the performance of the advance procedure with both the resulting sets. The data points are employed to train an ExtraTrees which provides predictions of the benchmark set with an average relative error of 4.9% and 3.5%, both larger than the error achieved by the adaptive design procedure.

We have also quantified the number of data points required by an evenly distributed grid to ensure the same accuracy on the predictions. Fig. 9 shows that an evenly distributed grid converges to the target accuracy much slower, requiring roughly 32,000 points to obtain an analogous accuracy. This means that the adaptive design procedure is able to provide the expected accuracy by employing 1/8 of the data points required by an evenly distributed set. The reduction in data points results in a relevant reduction of the computational cost spent in the generation of the ExtraTrees. As a whole, this result reveals that the adaptive design of the training points is the enabling factor to increase the overall dimensionality of the system being able to minimize the computational cost without hindering the accuracy.

4.5. Simulation of a PFR with WGS on Rh

ExtraTrees has been employed to properly tabulate and interpolate the results of the microkinetic model. Here, we employ the ExtraTrees net production rates obtained with the adaptive procedure to carry out the reactor simulation. This enables a direct assessment of the accuracy of the methodology within reactor simulations by a comparison with the results obtained by simultaneously solving also the microkinetic model.

We simulate WGS on Rh in a washcoated monolith with square channels. The channel is 12 mm length with an initial zone of 2 mm without coating. The channel diameter is 3.1 mm. The catalyst loading is $4.14{\cdot}10^5\ m_{cat}^2\ m^{-3}$. A mixture of CO and H_2O with molar fractions equal to 0.02 is fed to the system along with inert N_2 with a flowrate of 30 NL min^{-1}. The system is isothermal at the temperature of 850 K and it is kept at a total pressure of 1 bar.

Fig. 10 compares the profiles of reactants and products along the reactor axis. A perfect agreement (i.e., deviation below 0.2%) between the predictions obtained through ExtraTrees and complete solution of the microkinetic model is observed. Hence, the simulations carried out with net rates tabulated with the ExtraTrees method provides the same results of the complete solution of the microkinetic scheme in the model. As such, these results highlight the possibility of achieving an effective coupling between complex kinetic schemes and reactors models through the combined effect of Machine Learning and the adaptive procedure for the generation of the training points.

5. Conclusions

A procedure for the adaptive design of the training points for ML techniques has been proposed in this work. The procedure is conceived to generate a curated training set minimizing the number of records required for the accurate approximation of the real function. To achieve this, the capability of the RF and ExtraTrees to evaluate the variable importance is employed to selectively refine the training set over the directions of more impact on the ML predictions. Moreover, the positions of the training points are defined according to the gradient of the function. Hence, the position of the data points follows the function trend and most of the training data are placed in the region of sharp variations.

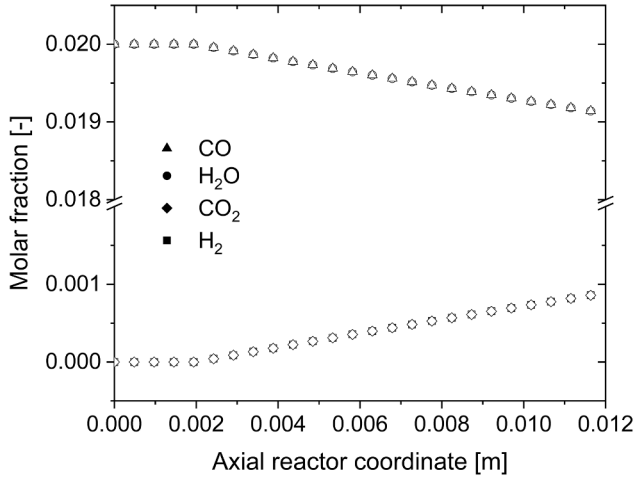


Fig. 10. Species molar fractions along the reactor length evaluated with the direct solution of the microkinetic model (full symbols) and by means of the ExtraTrees tabulated reaction rates (empty symbols).

We reveal the potential of the procedure employing the tabulation through ExtraTrees of first-principles base kinetic scheme aiming at their inclusion in reactor models on two levels. First, we assessed the efficiency of the design of the points by comparing the ExtraTrees predictions with respect to an evenly distributed rectangular grid. The same accuracy is achieved by considering significantly less training points ($< 60\%$), resulting in a relevant computational saving. Moreover, we demonstrated the accuracy of the tabulated kinetic models in the description of the reactor behavior through a direct assessment with the full kinetic scheme in the case of the mean-field model. The predictions of the two models are in perfect agreement with deviations below 0.2%. Finally, we included the first-principles kMC kinetic model in a CFD simulation through the ExtraTrees tabulated TOF enabling the coupling between the accurate description of the chemistry and the transport. The simulation of CO oxidation on a complex three-dimensional geometry of an open-cell foam revealed the capability of the methods in predicting the gas and surface species concentrations and trends. The tabulation of the site coverages also enabled the inclusion of specific information (i.e., site coverages) provided by the kMC simulation.

The envisioned procedure reveals to be an effective methodology in reducing and optimizing the size and shape of Machine Learning training set. On a broader perspective, this concept can also be extended to the reduction of the cost connected to the application of ML in the context of the approximation of computationally expensive functions important in catalysis at large, such as first-principles calculations or generation of potential and forces for molecular dynamics.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The project leading to this work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 677423 (ERC project SHAPE). Computational time at CINECA, Bologna (Italy) is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2020.125469. The latest version of the source code is available at the Github page: https://github.com/mbracconi/adaptiveDesignProcedure.

References

[1] M.K. Sabbe, M.F. Reyniers, K. Reuter, First-principles kinetic modeling in heterogeneous catalysis: an industrial perspective on best-practice, gaps and needs, Catal. Sci. Technol. 2 (2012) 2010–2024, https://doi.org/10.1039/c2cy20261a.
[2] A. Bruix, J.T. Margraf, M. Andersen, K. Reuter, First-principles-based multiscale modelling of heterogeneous catalysis, Nat. Catal. 2 (2019) 659–670, https://doi.org/10.1038/s41929-019-0298-3.
[3] S. Matera, W.F. Schneider, A. Heyden, A. Savara, Progress in accurate chemical kinetic modelling, simulations, and parameter estimation for heterogeneous catalysis, ACS Catal. 9 (2019) 6624–6647, https://doi.org/10.1021/acscatal.9b01234.
[4] M.P. Dudukovic, Frontiers in reactor engineering, Science 325 (5941) (2009) 698–701, https://doi.org/10.1126/science:1174274.
[5] M. Maestri, Escaping the trap of complication and complexity in multiscale microkinetic modelling of heterogeneous catalytic processes, Chem. Commun. 53 (2017) 10244–10254, https://doi.org/10.1039/c7cc05740g.
[6] M. Maestri, A. Cuoci, Coupling CFD with detailed microkinetic modeling in heterogeneous catalysis, Chem. Eng. Sci. 96 (2013) 106–117, https://doi.org/10.1016/j.ces.2013.03.048.
[7] T. Maffei, G. Gentile, S. Rebughini, M. Bracconi, F. Manelli, S. Lipp, A. Cuoci, M. Maestri, A multiregion operator-splitting CFD approach for coupling microkinetic modeling with internal porous transport in heterogeneous catalytic reactors, Chem. Eng. J. (2016), https://doi.org/10.1016/j.cej.2015.08.080.
[8] M. Bracconi, M. Maestri, A. Cuoci, In situ adaptive tabulation for the CFD simulation of heterogeneous reactors based on operator-splitting algorithm, AIChE J. 63 (2017) 95–104, https://doi.org/10.1002/aic.15441.
[9] S. Rebughini, A. Cuoci, A.G. Dixon, M. Maestri, Cell agglomeration algorithm for coupling microkinetic modeling and steady-state CFD simulations of catalytic reactors, Comput. Chem. Eng. 97 (2017) 175–182, https://doi.org/10.1016/j.compchemeng.2016.11.033.
[10] S. Matera, M. Maestri, A. Cuoci, K. Reuter, Predictive-quality surface reaction chemistry in real reactor models: Integrating first-principles kinetic monte carlo simulations into computational fluid dynamics, ACS Catal. 4 (2014) 4081–4092, https://doi.org/10.1021/cs501154e.
[11] J.E. Sutton, J.M. Lorenzi, J.T. Krogel, Q. Xiong, S. Pannala, S. Matera, A. Savara, Electrons to reactors multiscale modeling: catalytic CO oxidation over RuO2, ACS Catal. 8 (2018) 5002–5016, https://doi.org/10.1021/acscatal.8b00713.
[12] D.G. Vlachos, Multiscale integration hybrid algorithms for homogeneous-heterogeneous reactors, AIChE J. 43 (1997) 3031–3041, https://doi.org/10.1002/aic.690431115.
[13] D. Majumder, L.J. Broadbelt, A multiscale scheme for modeling catalytic flow reactors, AIChE J. 52 (2006) 4214–4228, https://doi.org/10.1002/aic.11030.
[14] C. Schaefer, A.P.J. Jansen, Coupling of kinetic Monte Carlo simulations of surface reactions to transport in a fluid for heterogeneous catalytic reactor modeling, J. Chem. Phys. 138 (2013), https://doi.org/10.1063/1.4789419.
[15] M. Klingenberger, O. Hirsch, M. Votsmeier, Efficient interpolation of precomputed kinetic data employing reduced multivariate Hermite Splines, Comput. Chem. Eng. 98 (2017) 21–30, https://doi.org/10.1016/j.compchemeng.2016.12.005.
[16] M. Votsmeier, Efficient implementation of detailed surface chemistry into reactor models using mapped rate data, Chem. Eng. Sci. (2009), https://doi.org/10.1016/j.ces.2008.12.006.
[17] M. Votsmeier, A. Scheuer, A. Drochner, H. Vogel, J. Gieshoff, Simulation of automotive NH3 oxidation catalysts based on pre-computed rate data from mechanistic surface kinetics, Catal. Today (2010) 271–277, https://doi.org/10.1016/j.cattod.2010.01.018.
[18] A. Scheuer, O. Hirsch, R. Hayes, H. Vogel, M. Votsmeier, Efficient simulation of an ammonia oxidation reactor using a solution mapping approach, Catal. Today (2011), https://doi.org/10.1016/j.cattod.2011.03.036.
[19] B. Partopour, A.G. Dixon, Computationally efficient incorporation of microkinetics into resolved-particle CFD simulations of fixed-bed reactors, Comput. Chem. Eng. 88 (2016) 126–134, https://doi.org/10.1016/j.compchemeng.2016.02.015.
[20] B. Partopour, A.G. Dixon, Resolved-particle fixed bed CFD with microkinetics for ethylene oxidation, AIChE J. 63 (2017) 87–94, https://doi.org/10.1002/aic.15422.
[21] J.M. Lorenzi, T. Stecher, K. Reuter, S. Matera, Local-metrics error-based Shepard interpolation as surrogate for highly non-linear material models in high dimensions, J. Chem. Phys. 147 (2017), https://doi.org/10.1063/1.4997286.
[22] B.R. Goldsmith, J. Esterhuizen, J.-X. Liu, C.J. Bartel, C. Sutton, Machine learning for heterogeneous catalyst design and discovery, AIChE J. 64 (2018) 2311–2323, https://doi.org/10.1002/aic.16198.
[23] P. Schlexer Lamoureux, K.T. Winther, J.A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, Machine learning for computational heterogeneous catalysis, ChemCatChem 11 (2019) 3581–3601, https://doi.org/10.1002/cctc.201900595.
[24] D.A.C. Beck, J.M. Carothers, V.R. Subramanian, J. Pfaendtner, Data science: accelerating innovation and discovery in chemical engineering, AIChE J. 62 (2016) 1402–1416, https://doi.org/10.1002/aic.15192.

[25] Y. Basdogan, M.C. Groenenboom, E. Henderson, S. De, S.B. Rempe, J.A. Keith, Machine learning-guided approach for studying solvation environments, J. Chem. Theory Comput. 16 (2020) 633–642, https://doi.org/10.1021/acs.jctc.9b00605.
[26] J.R. Kitchin, Machine learning in catalysis, Nat. Catal. 1 (2018) 230–232, https://doi.org/10.1038/s41929-018-0056-y.
[27] A.R. Singh, B.A. Rohr, J.A. Gauthier, J.K. Nørskov, Predicting chemical reaction barriers with a machine learning model, Catal. Letters. 149 (2019) 2347–2354, https://doi.org/10.1007/s10562-019-02705-x.
[28] Z.W. Ulissi, A.J. Medford, T. Bligaard, J.K. Nørskov, To address surface reaction network complexity using scaling relations machine learning and DFT calculations, Nat. Commun. 8 (2017), https://doi.org/10.1038/ncomms14621.
[29] S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran, Z.W. Ulissi, Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts, J. Phys. Chem. Lett. 10 (2019) 4401–4408, https://doi.org/10.1021/acs.jpclett.9b01428.
[30] M.R. Malik, B.J. Isaac, A. Coussement, P.J. Smith, A. Parente, Principal component analysis coupled with nonlinear regression for chemistry reduction, Combust. Flame 187 (2018) 30–41, https://doi.org/10.1016/j.combustflame.2017.08.012.
[31] G. Aversano, A. Bellemans, Z. Li, A. Coussement, O. Gicquel, A. Parente, Application of reduced-order models based on PCA & Kriging for the development of digital twins of reacting flow applications, Comput. Chem. Eng. 121 (2019) 422–441, https://doi.org/10.1016/j.compchemeng.2018.09.022.
[32] N. Shenvi, J.M. Geremia, H. Rabitz, Efficient chemical kinetic modeling through neural network maps, J. Chem. Phys. 120 (2004) 9942–9951, https://doi.org/10.1063/1.1718305.
[33] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[34] B. Partopour, R.C. Paffenroth, A.G. Dixon, Random Forests for mapping and analysis of microkinetics models, Comput. Chem. Eng. 115 (2018) 286–294, https://doi.org/10.1016/j.compchemeng.2018.04.019.
[35] B. Partopour, R.C. Paffenroth, A.G. Dixon, Random Forests for mapping and analysis of microkinetics models, Comput. Chem. Eng. (2018), https://doi.org/10.1016/j.compchemeng.2018.04.019.
[36] S. Döpking, C.P. Plaisance, D. Strobusch, K. Reuter, C. Scheurer, S. Matera, Addressing global uncertainty and sensitivity in first-principles based microkinetic models by an adaptive sparse grid approach, J. Chem. Phys. 148 (2018), https://doi.org/10.1063/1.5004770.
[37] S. Döpking, S. Matera, Error propagation in first-principles kinetic Monte Carlo simulation, Chem. Phys. Lett. 674 (2017) 28–32, https://doi.org/10.1016/j.cplett.2017.02.043.
[38] S. Back, K. Tran, Z.W. Ulissi, Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning, ACS Catal. 9 (2019) 7651–7659, https://doi.org/10.1021/acscatal.9b02416.
[39] C. Wang, A. Tharval, J.R. Kitchin, A density functional theory parameterised neural network model of zirconia, Mol. Simul. 44 (2018) 623–630, https://doi.org/10.1080/08927022.2017.1420185.
[40] O.-P. Koistinen, F.B. Dagbjartsdóttir, V. Ásgeirsson, A. Vehtari, H. Jónsson, Nudged elastic band calculations accelerated with Gaussian process regression, J. Chem. Phys. 147 (2017) 152720, , https://doi.org/10.1063/1.4986787.
[41] X. Chen, C.F. Goldsmith, Accelerating variational transition state theory via artificial neural networks, J. Phys. Chem. A 124 (2020) 1038–1046, https://doi.org/10.1021/acs.jpca.9b11507.
[42] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42, https://doi.org/10.1007/s10994-006-6226-1.
[43] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer New York Inc., New York, NY, USA, 2001.
[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[45] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), Adv. Neural Inf. Process. Syst. 26, Curran Associates, Inc., 2013, pp. 431–439 http://papers.nips.cc/paper/4928-understanding-variable-importances-in-forests-of-randomized-trees.pdf.
[46] M. Sandri, P. Zuccolotto, Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms, Stat. Comput. 20 (2010) 393–407, https://doi.org/10.1007/s11222-009-9132-0.
[47] S. Nembrini, I.R. König, M.N. Wright, The revival of the Gini importance? Bioinformatics 34 (2018) 3711–3718, https://doi.org/10.1093/bioinformatics/bty373.
[48] K.K. Nicodemus, Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures, Brief. Bioinform. 12 (2011) 369–373, https://doi.org/10.1093/bib/bbr016.
[49] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: illustrations, sources and a solution, BMC Bioinf. 8 (2007), https://doi.org/10.1186/1471-2105-8-25.
[50] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinf. 9 (2008) 1–11, https://doi.org/10.1186/1471-2105-9-307.
[51] M. Stamatakis, D.G. Vlachos, A graph-theoretical kinetic Monte Carlo framework for on-lattice chemical kinetics, J. Chem. Phys. 134 (2011), https://doi.org/10.1063/1.3596751.
[52] J. Nielsen, M. d'Avezac, J. Hetherington, M. Stamatakis, Parallel kinetic Monte Carlo simulation framework incorporating accurate models of adsorbate lateral interactions, J. Chem. Phys. 139 (2013) 224706, , https://doi.org/10.1063/1.4840395.
[53] R.K. Shah, A.L. London, Laminar Flow Forced Convection in Ducts: A Source Book for Compact Heat Exchanger Analytical Data, Academic Press, 1978.
[54] T.P. Coffee, J.M. Heimerl, Transport algorithms for premixed, laminar steady-state flames, Combust. Flame 43 (1981) 273–289, https://doi.org/10.1016/0010-2180(81)90027-4.
[55] J.H. Ferziger, M. Peric, Computational Methods for Fluid Dynamics, Springer Berlin Heidelberg, 2001 https://books.google.it/books?id=1D3EQgAACAAJ.
[56] H. Jasak, A. Jemcov, Z. Tukovic, OpenFOAM: A C + + library for complex physics simulations, Int. Work. Coupled Methods Numer. Dyn. (2007) 1–20.
[57] K. Reuter, M. Scheffler, First-principles kinetic Monte Carlo simulations for heterogeneous catalysis: application to the CO oxidation at Ru O2 (110), Phys. Rev. B – Condens. Matter Mater. Phys. 73 (2006) 1–17, https://doi.org/10.1103/PhysRevB.73.045433.
[58] M. Bracconi, M. Ambrosetti, M. Maestri, G. Groppi, E. Tronconi, A systematic procedure for the virtual reconstruction of open-cell foams, Chem. Eng. J. 315 (2017) 608–620, https://doi.org/10.1016/j.cej.2017.01.069.
[59] M. Bracconi, M. Ambrosetti, M. Maestri, G. Groppi, E. Tronconi, A fundamental investigation of gas/solid mass transfer in open-cell foams using a combined experimental and CFD approach, Chem. Eng. J. 352 (2018) 558–571, https://doi.org/10.1016/j.cej.2018.07.023.
[60] M. Maestri, D.G. Vlachos, A. Beretta, G. Groppi, E. Tronconi, A C1 microkinetic model for methane conversion to syngas on Rh/Al2O3, AIChE J. 55 (2009) 993–1008, https://doi.org/10.1002/aic.11767.
[61] M. Maestri, D.G. Vlachos, A. Beretta, G. Groppi, E. Tronconi, Steam and dry reforming of methane on Rh: microkinetic analysis and hierarchy of kinetic models, J. Catal. 259 (2008) 211–222, https://doi.org/10.1016/j.jcat.2008.08.008.