MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **78** (2017) 511-526

ISSN 0340 - 6253

Sampling Methodologies for High Throughput Materials Discovery

Laurent A. Baumes*
Instituto de Tecnologia Quimica, (UPV-CSIC)
Av. de Los Naranjos 0, E-46022 Valencia, Spain
baumesl@gmail.com

Jesus Vicente de Julián-Ortiz*
Departament de Química Física
Facultat de Farmàcia
Universitat de València – Estudi General
Av. V. Andrés Estellés, 0, E-46100, Burjassot, Valencia, Spain
jejuor@uv.es

(Received May 19, 2017)

Abstract

A mini-review of developments in sampling algorithms applied to the design of new catalysts and materials is presented. The data mining technology is increasingly being used in new industrial processes, which require automatic analysis of data and related results to proceed quickly to conclusions. However, for some applications, absolute automation may not be appropriate. Unlike traditional data mining contexts, processing of large amounts of data, some domains are characterized by the scarcity of data, due to the cost and time involved in the realization of simulations or the setting up of experimental apparatuses for the collection of data. In such domains, therefore, it is prudent to balance the speed through the automation and the utility of the generated data.

1 Introduction

The discovery of new materials is a key factor for the improvement of the industrial competitiveness. This has been evidenced by the emergence of initiatives such as the

Materials Genome Initiative [1-4] in 2011, *Metallurgy Europe* [5] in 2012, or the *Materials Research Consortium* [6] that have boosted the data science in this area.

In recent years, data mining has invoked great attention in academia but also in industry where this technology is more and more employed into new industrial processes, which usually require automatic analysis of data in order to proceed quickly to conclusion/decision. However, for some applications, an absolute automation may not be appropriate. In those cases when there exists the possibility of evaluate a whole dataset, retrospective analysis can yield algorithms to reproduce the obtained results, to gain deep insight in the key features involved in the problem.

Few articles can be found where a library of potential catalysts is synthesized and tested, and the results analyzed to get algorithms able to select reasonably the best catalysts [7-10]. Unlike traditional data mining contexts working with voluminous amounts of data, some domains are actually characterized by a scarcity of data, owing to the cost and time involved in conducting simulations or setting up experimental apparatus for data collection. In such domains, it is hence prudent to balance speed through automation and the utility of the generated data. Therefore, the human interaction and guidance must rule to attain better quality output.

In many natural learning tasks, knowledge is gained iteratively, by making action, queries, or experiments. Active learning (AL) is concerned with the integration of data collection, design of experiments and data mining, for making better data exploitation. The learner is not treated as a classical passive recipient of data to be processed. AL can be used in two extreme cases. i) The number of data available is very large, thus a mining algorithm uses a selected data subset rather than the whole available data. ii) The researcher has the control of data acquisition, and he has to pay attention on the iterative selection of samples for extracting the greatest benefit from future data treatments. The second situation becomes especially crucial when each data point is costly, the domain knowledge is imperfect, and theory-driven approaches are inadequate such as for materials science fields and, especially, heterogeneous catalysis [11].

2 Specificities of the domain of application: Towards an adequate strategy

A catalytic reaction is a chemical reaction in which transformations are accelerated thanks to a substance called *catalyst*. Starting molecules and intermediates, as soon as are formed, can

interact with the catalyst in a specific/discriminating manner. This infers that some transformation steps can be accelerated. Catalytic processes constitute the fundamentals of modern chemical industries. Over 90% of the newly introduced chemical processes are catalytic [12]. In the highly developed industrial countries, catalytic processes create about 20% of the Gross Domestic Product. Catalysis is responsible in the manufacture of over $3 trillion in goods and services. We will focus on heterogeneous catalysis which involves the use of catalysts acting in a different phase from the reactants, typically a solid catalyst with liquid or gaseous reactants. For further details, the reader is referred to [13].

During the whole catalytic development, a very large number of features and parameters have to be screened and therefore any detailed and relevant catalyst description remains a challenge. The selection of adequate input variables is a problem of the utmost importance [14-16]. All these parameters generate an extremely large degree of complexity. As a consequence, the entire catalyst development is long (~15 years) and costly. The conventional catalyst development relies on fundamental knowledge and know-how. The main drawback of this approach is to be a very time-consuming process, making and testing one material at a time. Another drawback comes from the relative importance of intuition for the initial choices of the development strategy [17]. To overcome these major drawbacks, attempts to shorten this process by using high throughput (HT) technology have been reported since about 20 years [18-31]. The HT approach is more pragmatic-oriented. It implies the screening of collections of samples.

It may be stressed out that the relevant parameters are usually unknown and some of them cannot be directly and individually controlled. In addition, it is in general a combination of factors that provides outstanding properties which are required to meet challenging targets. Principal component analysis (PCA) has been used in the data analysis and dimensional reduction in the hydrothermal synthesis of zeolites [32] and also to identify relevant key features for catalytic activities in a dataset of pentanary-mixed metal oxides [33]. This last approach allowed restricting the initial chemical diversity and focuses the sampling techniques in the most promising candidates [34]. Other techniques used with the purpose of reducing the parameters involved in the search are neural networks (NN) developments, such as the NN analysis of factors controlling catalytic activity [35] or the numerical partial differentiation of a trained NN pattern [36]. Another interesting feature of combinatorial materials search is the dependence between the composition and the reproducibility of the material properties [37,38].

The tools necessary for the combinatorial approach can be classified into two main categories *i)* HT equipments for fast and parallel synthesis and testing of catalysts and *ii)* computational methods. HT experimentation has become an accepted and important strategy in the development of catalysts and materials [39-44]. However, such an approach has more success in the optimization than in the discovery [11,45]. Despite of the fast syntheses and the testing robots, each catalytic experiment still requires a few hours. Here, the learner's most powerful tool is its ability to act gathering the data.

The general problem considered here is the efficacy of experimental data selection in HT of heterogeneous catalysis within a discovery program. Considering such domain, only very fast screening tools should be employed aiming at finding the various "groups" of catalyst outputs. This pre-screening of the search space shall extract information or knowledge from the restricted selected sampling in order to provide guidelines and well defined boundaries for further screenings and optimization. Here, the output performance is related to the identification of classes. Ranking (if exists) is not taken into account since the objective is not the optimization of the catalytic activity or selectivity. The chemist knowledge should permit to define *a priori* broad and "poorly-previously-explored" parameter space, letting opportunities to surprising or unexpected catalytic results. The typical distribution of catalytic outputs usually exhibits unbalanced datasets for which an efficient learning can be hardly carried out. Even if the overall recognition rate may be satisfactory, catalysts belonging to rare but usually interesting classes can be misclassified.

The sampling strategy in HT material science, and especially in heterogeneous catalysis, typically embodies an assessment of where might be a good location to collect data or to plan experiments in iterative optimization, in the chemical space, to get information on the selectivity or conversion [46]. This can be performed by an evolutionary algorithm [47-60] of specific design criteria usually, homogeneous covering [61-65] or traditional design of experiments [66-77] (DoE), the later is usually neglected due to the specificity of the different methods and the restrictions imposed by the domain. Other techniques comprise the optimal design of experiments, based on Levenberg-Marquard optimization scheme [78], the multiobjective design of experiments [79,80], and the random forest regression [81]. For most of studies, Simple Random Sampling (SRS) rules the domain. However, SRS should not be underestimated, see [82] for a detailed explanation of the SRS robustness.

Diversity monitoring is not a sampling method but merits to be mentioned here. It allows enhancing the reliability of the sampling by selecting samples that maximize a parameter based on distances [83].

In data science, Machine Learning (ML) can be defined as biomimetic modeling; in contrast to Statistical modeling that relies on probability theory. The two most important paradigms in ML are the supervised and the unsupervised learning techniques. Supervised learning generates functions that relate inputs to known outputs. By contrast, unsupervised learning models series of inputs. Applications of ML in materials science have been recently reviewed [84].

Supervised learning could be interesting in sampling, but few papers deal with such strategies in this domain. One of these methods, called *mapping*, has been used to guide *discovery* studies [32,85-88]. The Mapping method develops relationships among properties such as composition and synthesis conditions while these interactions may be obtained without searching for hits or lead materials. Then, the results of mapping studies can be used as inputs to guide subsequent screening or optimization experiments. The purpose of screening experiments is said to identify iteratively, by accumulation of knowledge, hits or small space regions of materials with promising properties. The last manner to guide the chemist, called optimization, is when experiments are designed to refine material properties. Mapping receives relatively little attention, being too often subsumed under screening. New methodologies are presented that aims at generating successively new samples in order to reach an improved final estimate of the entire search space investigated according to the knowledge accumulated iteratively through samples selection and corresponding obtained results.

A new iterative algorithm was proposed by the corresponding author of this review for the characterization of the space structure [89]. This algorithm, called MAP for *MAPping*, is able to: i) Increase the quality of the ML performed at the end of the first exploratory stage. ii) Work independently from the choice of the supervised learning system. iii) Decrease the misclassification rates of catalysts belonging to small frequency classes of performance. iv) Handle both quantitative and qualitative features. v) Proceed iteratively while capturing information contained into all the previous experiments. vi) Integrate inherent constraints such as *a priori* fixed reactor capacity, *i.e.* the amount of the iteratively selected samples to be labelled, and a maximum number of experiments to be conducted, so-called deadline.

3 Active learning

Active learning (AL) assumes that the data is provided by a source which is controlled by the researcher. Such control is used for different aims and in different ways. The various fields for which one may wish to use AL are numerous such as *optimization*, where the learner experiments to find a set of inputs that maximize some response variable, for example the response surface methodology [90,91] which guides hill-climbing through the input space; *adaptive control*, where one must learn a control policy by taking actions, one may face the complication that the result of a specific action remains unknown during a time; *model selection problem* for driving data collection in order to refine and discriminate a given set of models. For all types of application, the principal AL task is to determine an "optimal" sample selection strategy. Such optimization is defined through a criterion, called selection scheme, depending on the user aim. Therefore, considering the model selection problem, the approach can either be motivated by the need to disambiguate among models or to gain the most prediction accuracy from a ML algorithm, while requiring the fewest number of labels.

Before inspecting the different selection schemes proposed in the bibliography, it has to be noted that new samples can either be created by the system or selected from an unlabeled set. The first approach is not investigated here, and considering the domain of application, it remains difficult to generate samples without lack of coherence. A system could produce non existing materials to be labelled. This methodology has been explored and often describes "impossible" catalysts [92,93]. For example: to prepare by a precipitation process a solid consisting of 30% Ba, 50% Na, and 20% V (oxygen is excluded), using inorganic, non-halide precursors from aqueous solution, using suitable precursors, finding a precipitation agent which would precipitate all three metals at the same time is virtually impossible.

The second approach is the most common and corresponds to the one we are concerned. Two kinds of selection from an unlabeled set can be distinguished. The pool-based approach allows the selection among *a priori* restricted set of unlabelled samples while the other one allows picking up any sample to be labelled from an entire pre-defined search space.

Another criterion that should be taken into account when specifying an AL algorithm is the exact role of the ML system. AL usually starts from a very small number of labelled samples, and then iteratively asks for new samples. The following cases are discriminated on the basis of the frequency of learning system update. The selection of new samples may be done in order to update at each new round either the previously obtained model, increasing its performance and accuracy, or a given criterion which remains independent from the learning

system allowing a unique use of the ML when the whole selection is achieved. Using the semantic used in feature selection domain, the first protocol may be called a *wrapper* approach while the second one may be qualified as a *filter*. The advantage of using a wrapper technique is that the selection is optimized considering the learning algorithm that has been previously chosen. However, such choice is not always trivial, and depends on the complexity of the underlying system investigated (which is usually unknown or difficult to quantify) but also on the complexity of the ML system itself since considering complex algorithms it may be delicate to elaborate the selection scheme. Moreover, for many configurations, such methodology might be intractable.

3.1 Selection schemes

The primary question of AL is how to choose which points to try next. A simple strategy for sampling is to target locations to reduce our *uncertainty* in modelling, for example by selecting the location that minimizes the posterior generalized variance of a function. The distribution *P(Y/X)* being unknown, a classical approach consists in approximating *P* with many samples but then a great amount of hypothesis and simplifications have to be done to compute the estimated error reduction. For example, using a probabilistic classifier, uncertainty sampling would pick the observation for which the predicted class probabilities yield the greatest entropy. The query by committee utility [94,95] measures the classification disagreement of a committee of classifiers. By choosing an example with large disagreement, Cohn *et al.* [96] measured the expected reduction in the prediction of variance of NN and other models. Another closely-related solution is to select the most ambiguous sample. *Ambiguity-directed* sampling aims at clarifying the decision-making near the ambiguity. Making the assumption that close elements are similar, the knowledge of one sample should induce the knowledge of the neighboring. However, ambiguous points are likely to be neighbors. It is therefore important to select ambiguous points spread over the distribution of input variables. Other solutions for choosing these points are to look for "places" where there is no data [97], where it is expected to change the model [98]. Thus, reference [99] relies on measuring the variation in label assignments (of the unlabeled set) between the classifier trained on the training set only and the classifiers trained on the training set with a single unlabeled object added with all possible labels. Other closely related selection schemes are investigated which aims at choosing points where the system performs poorly [100], and where it was previously found data that resulted in learning [101]. Other solutions are directly

induced by the domain of application, for instance, robot navigation studies [102]. In such learning tasks, data-query is neither free nor of constant cost. Researchers try to integrate the fact that the cost of a query depends on the distance from the current location in state space to the desired query point. On the other hand, such notion of distance is not transferable to the synthesis of materials.

4 Final remarks

Although the previously reviewed literature is very valuable and gives theoretical justification of using AL, even without considering the specificities which make them unusable in our case, most of the relevant articles require a degree of statistical sophistication which is beyond the reach of most practitioners of the domain of high throughput materials science.

The empirical results of the method MAP have demonstrated the effectiveness of the active mining strategy on synthetic datasets. The strategy has been tested against simple random sampling (SRS) on numerous benchmarks with different levels of complexity [89]. The method is a stochastic group sequential biased sampling which iteratively proposes a sample of the search space. The approach does not need to sample the entire combinatorial space, but only enough to be able to identify the structure of classes without forgetting classes obtained only with few experiments [89].

At the moment, such approach is used in a research program for the discovery of new zeolites.

References

[1] National Science and Technology Council. Materials genome initiative for global competitiveness, 2011, June. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf. Accessed on 2017, March.

[2] C. Ward, Materials genome initiative for global competitiveness, in: *23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition*, American Society for Metals, 2012.

[3] https://mgi.nist.gov/. Accessed on 2017, March.

[4] https://www.mgi.gov/. Accessed on 2017, March.

[5] Materials Science and Engineering Expert Committee (MatSEEC), *Metallurgy Europe – a renaissance programme for 2012–2022*, European Science Foundation, 2012.

[6] S. Curtaroloa, W. Setyawana, G. L. W. Hartc, M. Jahnateka, R. V. Chepulskiia, R. H. Taylora, S. Wanga, J. Xuea, K. Yanga, O. Levyd, M. J. Mehl, AFLOW: an automatic framework for high-throughput materials discovery, *Comput. Mater. Sci.* **58** (2012) 218-226.

[7] S. A. Schunk, A. Sundermann, H. Hibst, Retrospective hit-deconvolution of mixed metal oxides: spotting structure-property-relationships in gas phase oxidation catalysis through high throughput experimentation, *Comb. Chem. High Throughput Screen.* **10** (2007) 51-57.

[8] N. Vriamont, B. Govaerts, P. Grenouillet, C. de Bellefon, O. Riant, Design of a genetic algorithm for the simulated evolution of a library of asymmetric transfer hydrogenation catalysts, *Chem. Eur. J.* **15** (2009) 6267-6278.

[9] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, H. Lam, Combinatorial and high-throughput screening of materials libraries: Review of state of the art, *ACS Comb. Sci.* **13** (2011) 579-633.

[10] N. P. Lavery, S. Mehraban, C. Pleydell-Pearce, S. G. R. Brown, D. J. Jarvis, W. Voice, M. Brunnock, Combinatorial development and high throughput materials characterisation of steels, *Ironmaking & Steelmaking* **42** (2015) 727-733.

[11] C. Klanner, D. Farrusseng, L. A. Baumes, M. Lengliz, C. Mirodatos, F. Schüth, The development of descriptors for solids: teaching "catalytic intuition" to a computer, *Angew. Chem. Int. Ed.* **43** (2004) 5347-5349.

[12] G.W. Zhan, H.C. Zeng, Integrated nanocatalysts with mesoporous silica/silicate and microporous MOF materials. *Coord. Chem. Rev.* **320** (2016) 181-192.

[13] G. Ertl, H. Knözinger, J. Weitkamp, *Handbook of Heterogeneous Catalysis*, Wiley-VCH, 1997.

[14] J. Procelewska, J. L. Galilea, F. Clerc, D. Farrusseng, F. Schueth, Computational methods in the development of a knowledge-based system for the prediction of solid catalyst performance, *Comb. Chem. High Throughput Screen.* **10** (2007) 37-50.

[15] D. Farrusseng, High-throughput heterogeneous catalysis, *Surf. Sci. Rep.* **63** (2008) 487-513.

[16] E. Jeroro, K. V. Bussche, The role of characterization and modeling techniques in fostering the era of computer-based catalyst and reactor design, *Curr. Opin. Chem. Eng.* **13** (2016) 186-192.

[17] B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, W. H. Weinberg, Combinatorial materials science and catalysis, *Angew. Chem. Int. Ed.* **38** (1999) 2494-2532.

[18] A. M. Porte, J. Reibenspies, K. Burgess, Design and optimization of new phosphine oxazoline ligands via high-throughput catalyst screening, *J. Am. Chem. Soc.* **120** (1998) 9180-9187.

[19] M .T. Reetz, Combinatorial and evolution-based methods in the creation of enantioselective catalysts, *Angew. Chem. Int. Ed.* **40** (2001) 284-310.

[20] P. Chen, Electrospray ionization tandem mass spectrometry in high-throughput screening of homogeneous catalysts, *Angew. Chern. Int. Ed.* **42** (2003) 2832-2847.

[21] S. Garbacia, C. Hillairet, R. Touzani, O. Lavastre, New nitrogen-rich tripodal molecules based on bis (pyrazol-1-ylmethyl) amines with substituents modulating steric hindrances and electron density of donor sites, *Collect. Czech. Chem. Commun.* **70** (2005) 34-40.

[22] W. F. Maier, K. Stöwe, S. Sieg, Combinatorial and high-throughput materials science, *Angew. Chem. Int. Ed.* **46** (2007) 6016-6067.

[23] M. Baerns, M. Holeňa, *Combinatorial Development of Solid Catalytic Materials: Design of High-Throughput Experiments, Data Analysis, Data Mining*, Imperial College Press, London, 2009.

[24] Q. Y. Zhang, D. D. Zhang, J. Y. Li, Y. M. Zhou, L. Xu, Virtual screening of a combinatorial library of enantioselective catalysts with chirality codes and counterpropagation neural networks, *Chemom. Intell. Lab. Systems* **109** (2011) 113-119.

[25] W. B. Park, N. Shin, K. P. Hong, M. Pyo, K. S. Sohn, A new paradigm for materials discovery: heuristics-assisted combinatorial chemistry involving parameterization of material novelty, *Adv. Funct. Mater.* **22** (2012) 2258-2266.

[26] M. L. Green, I. Takeuchi, J. R. Hattrick-Simpers, Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials, *J. Appl. Phys.* 113 (2013) 9_1.

[27] J. Bennett, New classes of piezoelectrics, ferroelectrics, and anti ferroelectrics by first-principles high-throughput materials design, *Bull. Am. Phys. Soc.* **58** (2013) A21.00008.

[28] J. W. Lee, S. Timilsina, G. W. Kim, J. S Kim, A new strategy for novel binder discovery in nano and μ powder injection molding: A metaheuristics-assisted virtual combinatorial materials search, *Powder Techn.* **302** (2016) 187-195.

[29] K. Rajan (Ed.), *Informatics for Materials Science and Engineering: Data-Driven Discovery for Accelerated Experimentation and Application*, Butterworth-Heinemann, 2013.

[30] B. Colin, O, Lavastre, S. Fouquay, G. Michaud, F. Simon, O. Laferte, J. M. Brusson, Development of new high-throughput screening method to compare and to detect efficient catalysts for adhesive materials, *Int. J. Adhes. Adhes.* **68** (2016) 47-53.

[31] T. Lookman, F. J. Alexander, K. Rajan, *Information Science for Materials Discovery and Design*, Springer, Switzerland, 2016.

[32] A. Corma, M. Moliner, J. M. Serra, P. Serna, M. Díaz-Cabañas, L. A. Baumes, A new mapping/exploration approach for HT synthesis of zeolites, *Chem. Mat.* **18** (2006) 3287-3296.

[33] S. C. Sieg, C. Suh, T. Schmidt, M. Stukowski, K. Rajan, W. F. Maier, Principal component analysis of catalytic functions in the composition space of heterogeneous catalysts, *QSAR Comb. Sci.* **26** (2007) 528-535.

[34] C. Suh, A. Rajagopalan, X. Li, K. Rajan, The application of principal component analysis to materials science data, *Data Sci. J.* **1** (2002) 19-26.

[35] T. Hattori, S. Kito, Analysis of factors controlling catalytic activity by neural network, *Catal. Today* **111** (2006) 328-332.

[36] S. Kito, T. Hattori, Analysis of catalytic performance by partial differentiation of neural network pattern, *Chem. Eng. Sci.* **62** (2007) 5575-5578.

[37] A. K. Sharma, C. Kulshreshtha, K. Sohn, K. S. Sohn, Systematic control of experimental inconsistency in combinatorial materials science, *J. Comb. Chem.* **11** (2009) 131-137.

[38] A. K. Sharma, C. Kulshreshtha, K. S. Sohn, Discovery of new green phosphors and minimization of experimental inconsistency using a multi-objective genetic algorithm-assisted combinatorial method, *Adv. Functional Mat.* **19** (2009) 1705-1712.

[39] S. Senkan, Combinatorial heterogeneous catalysis — a new path in an old field, *Angew. Chem. Int. Ed.* **40** (2001) 312-329.

[40] S. I. Woo, K. W. Kim, H. Y. Cho, K. S. Oh, M. K. Jeon, N. H. Tarte, T. S. Kim, A. Mahmood, Current status of combinatorial and high-throughput methods for discovering new materials and catalysts, *QSAR Comb. Sci.* **24** (2005) 138-154.

[41] J. Cui, Y. S. Chu, O. O. Famodu, Y. Furuya, J. Hattrick-Simpers, R. D. James, A. Ludwig, S. Thienhaus, M. Wuttig, Z. Zhang, I. Takeuchi, Combinatorial search of thermoelastic shape-memory alloys with extremely small hysteresis width, *Nature Mater.* **5** (2006) 286–290.

[43] L. M. Kustov, Catalysis à la combi, *Russ. J. Gen. Chem.* **80** (2010) 2527-2540.

[44] Z. Qun Zheng, X. Ping Zhou, High speed screening technologies in heterogeneous catalysis, *Comb. Chem. High Throughput Screen.* **14** (2011) 147-159.

[45] D. Nicolaides, Robust material design: a new workflow for high-throughput experimentation and analysis, *QSAR Comb. Sci.* **24** (2005) 15-21.

[46] J. N. Cawse (Ed.), *Experimental Design for Combinatorial and High Throughput Materials Development*, New York: Wiley-Interscience, 2003.

[47] M. Holena, M. Baems, Feedforward neural networks in catalysis: A tool for the approximation of the dependency of yield on catalyst composition, and for knowledge extraction, *Catal. Today* **81** (2003) 485-494.

[48] K. Omata, T. Umegaki, Y. Watanabe, M. Yamada, Simple GA program developed for optimization of methanol and dimethyl ether synthesis, *Stud. Surf. Sci. Catal.* **145** (2003) 291-294.

[49] L. A. Baumes, P. Jouve, D. Farrusseng, M. Lengliz, N. Nicoloyannis, C. Mirodatos, Dynamic Control of the Browsing-Exploitation Ratio for Iterative Optimisations, in: V. Palade, R. J. Howlett, L. Jain (Eds.), *Knowledge-Based Intelligent Information and Engineering Systems. KES 2003*, Springer, Berlin, Heidelberg, 2003, pp. 265-270.

[50] U. Rodemerck, M. Baerns, M. Holena, D. Wolf, Application of a genetic algorithm and a neural network for the discovery and optimization of new solid catalytic materials, *Appl. Surf. Sci.* **223** (2004) 168-174.

[51] Y. Watanabe, T. Umegaki, M. Hashimoto, K. Omata, M. Yamada, Optimization of Cu oxide catalysts for methanol synthesis by combinatorial tools using 96 well microplates, artificial neural network and genetic algorithm, *Catal. Today.* **89** (2004) 455-464.

[52] J. N. Cawse, M. Baems, M. Holena, Efficient discovery of nonlinear dependencies in a combinatorial catalyst data set, *J. Chem. Inf. Comput. Sci.* **44** (2004) 143-146.

[53] J. S. Paul, R. Janssens, J. F. M. Denayer, G. V. Baron, P. A. Jacobs, Optimization of MoVSb oxide catalyst for partial oxidation of isobutane by combinatorial approaches, *J. Comb. Chem.* **7** (2005) 407-413.

[54] A. Corma, J. M. Serra, P. Serna, S. Valero, E. Argente, V. Botti, Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques), *J. Catal.* **229** (2005) 513-524.

[55] O. C. Gobin, F. Schüth, On the suitability of different representations of solid catalysts for combinatorial library design by genetic algorithms, *J. Comb. Chem.* **10** (2008) 835-846.

[56] J. Beckers, F. Clerc, J. H. Blank, G. Rothenberg, Selective hydrogen oxidation catalysts via genetic algorithms, *Adv. Synth. Catal.* **350** (2008) 2237-2249.

[57] J. E. Kreutz, A. Shukhaev, W. Du, S. Druskin, O. Daugulis, R. F. Ismagilov, Evolution of catalysts directed by genetic algorithms in a plug-based microfluidic device tested with oxidation of methane by oxygen, *J. Am. Chem. Soc.* **132** (2010) 3128-3132.

[58] D. P. Fenning, J. Hofstetter, A. E. Morishige, D. M. Powell, A. Zuschlag, G. Hahn, T. Buonassisi, Darwin at high temperature: advancing solar cell material design using defect kinetics simulations and evolutionary optimization, *Adv. Energy Mater.* **4** (2014) 1400459.

[59] S. K. Suram, M. Z. Pesenson, J. M.Gregoire, High throughput combinatorial experimentation+ informatics= combinatorial science, in: *Information Science for Materials Discovery and Design*, Springer, 2016, pp. 271-300.

[60] T. C. Le, D. A. Winkler, Discovery and optimization of materials using evolutionary approaches, *Chem. Rev.* **116** (2016) 6107-6132.

[61] D. S. Bem, E. J. Erlandson, R. D. Gillespie, L. A. Harmon, S. G. Schlosser, A. J. Vayda, Combinatorial experimental design using the optimal coverage approach, in: J. N. Cawse (Ed.), *Experimental Design for Combinatorial and High Throughput Materials Development*, Wiley, Hoboken, New Jersey, 2003, pp. 89-107.

[62] J. N. Cawse, R. Wroczynski, Combinatorial materials development using gradient arrays: designs for efficient use of experimental resources, in: J. N. Cawse (Ed.), *Experimental Design for Combinatorial and High Throughput Materials Development*, Wiley, Hoboken, New Jersey, 2003, pp. 109-127.

[63] J. M. Serra, A. Corma, D. Farrusseng, L. A. Baumes, C. Mirodatos, C. Flego, C. Perego, Styrene from toluene by combinatorial catalysis, *Catal. Today* **81** (2003) 425-436.

[64] J. Sjöblom, D. Creaser, K. Papadakis, C. U. I. Odenbrand, Use of experimental design in development of a catalyst system, in: *11th Nordic Symposium on Catalysis*, Oulu, Finland, Elsevier, Amsterdam, 2004, pp. 243-248.

[65] L. A. Harmon, Experiment planning for combinatorial materials discovery, *J. Mat. Sci.* **38** (2003) 4479-4485.

[66] S. N. Deming, S. L. Morgan, *Experimental Design: A Chemometric Approach*, Elsevier, 1993.

[67] D. C. Montgomery, *Design and Analysis of Experiments*, Wiley, 1991.

[68] M. Tribus, G. Sconyi. An alternative view of the Taguehi approach, *Qual. Prog.* **22** (1989) 46-48.

[69] J. M. Serra, A. Corma, S. Valero, E. Argente, V. Botti, Soft computing techniques applied to combinatorial catalysis: a new approach for the discovery and optimization of catalytic materials, *QSAR Comb. Sci.* **26** (2007) 11-26.

[70] J. M. Serra, L. A. Baumes, M. Moliner, P. Serna, A. Corma. Zeolite synthesis modelling with support vector machines: a combinatorial approach, *Comb. Chem. High Throughput Screen.* **10** (2007) 13-24.

[71] R. Vijay, J. Lauterbach, Design of experiments combined with high-throughput experimentation for the optimization of DeNOx catalysts, *Stud. Surf. Sci. Catal.* **171** (2007) 325-359.

[72] K. Omata, M. Hashimoto, G. Ishiguro, Y. Watanabe, T. Umegaki, M. Yamada, Artificial neural network-aided catalyst research for low-pressure DME synthesis from syngas, *ACS Symposium Ser.* **959** (2007) 211-224.

[73] S. Valero, E. Argente, V. Botti, J. M. Serra, P. Serna, M. Moliner, A. Corma, DoE framework for catalyst development based on soft computing techniques, *Comput. Chem. Eng.* **33** (2009) 225-238.

[74] K. Omata, Y. Yamazaki, Y. Kobayashi, M. Yamada, Application of response surface method using rapid screening, support vector machine, and multiple regression on the acidity and activity of Si-Al-Zr ternary oxide, *J. Comb. Chem.* **12** (2010) 435-444.

[75] Q. Tang, Y. Lau, S. Hu, W. Yan, Y. Yang, T. Chen, Response surface methodology using Gaussian processes: towards optimizing the trans-stilbene epoxidation over Co2+–NaX catalysts, *Chem. Eng. J.* **156** (2010) 423–431.

[76] Y. Yang, T. Lin, X. L. Weng, J. A. Darr, X. Z. Wang, Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials, *Comput. Chem. Eng.* **35** (2011) 671-678.

[77] M. Protière, N. Nerambourg, O. Renard, P. Reiss, Rational design of the gram-scale synthesis of nearly monodisperse semiconductor nanocrystals, *Nanoscale Res. Lett.* **6** (2011) 472.

[78] T. A. Beltrán-Oviedo, I. Batyrshin, J. M. Domínguez, The optimal design of experiments (ODOE) as an alternative method for catalysts libraries optimization, *Catal. Today* **148** (2009) 28-35.

[79] O. C. Gobin, A. M. Joaristi, F. Schüth, Multi-objective optimization in combinatorial chemistry applied to the selective catalytic reduction of NO with C3H6, *J. Catal.* **252** (2007) 205-214.

[80] J. Llamas-Galilea, O. C. Gobin, F. Schüth, Comparison of single- and multiobjective design of experiment in combinatorial chemistry for the selective dehydrogenation of propane, *J. Comb. Chem.* **11** (2009) 907-913.

[81] J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling, *Phys. Rev. X* **4** (2014) 011019.

[82] C. Sammut, J. Cribb, Is Learning rate a good performance criterion for learning? in: *Proceedings of the 7th Int. Machine Learning Conf. Austin*, Morgan Kaufmann, 1990, pp. 170-170.

[83] D. Farrusseng, F. Clerc, Diversity management for efficient combinatorial optimization of materials, *Appl. Surf. Sci.* **254** (2007) 772-776.

[84] T. Mueller, A. G. Kusne, R. Ramprasad, Machine learning in materials science: Recent progress and emerging applications, *Rev. Comput. Chem.* **29** (2016) 186-264.

[85] A. Tompos, J. L. Margitfalvi, E. Tfirst, L. Vegvari, M. A. Jaloull, H. A. Khalfalla, M. M. Elgarni, Development of catalyst libraries for total oxidation of methane. A case study for combined application of holographic research strategy and artificial neural networks in catalyst library design, *Appl. Catal. A* **285** (2005) 65-78.

[86] A. Tompos, J. L. Margitfalvi, E. Tfirst, L. Vegvari, Evaluation of catalyst library optimization algorithms: Comparison of the holographic research strategy and the genetic algorithm in virtual catalytic experiments, *Appl. Catal. A* **303** (2006) 72-80.

[87] A. Tompos, L. Vegvari, E. Tfirst, J. L. Margitfalvi, Assessment of predictive ability of artificial neural networks using holographic mapping, *Comb. Chem. High Throughput Screen.* **10** (2007) 121-134.

[88] A. Tompos, J. L. Margitfalvi, L. Vegvari, A. Hagemeyer, T. Volpe, C. J. Brooks, Visualization of large experimental space using holographic mapping and artificial neural networks. benchmark analysis of multicomponent catalysts for the water gas shift reaction, *Top. Catal.* **53** (2010) 100-107.

[89] L. A. Baumes, MAP: An iterative experimental design methodology for the optimization of catalytic search space structure modeling, *J. Comb. Chem.* **8** (2006) 304-314.

[90] G. Box, N. Draper, *Empirical Model-Building and Response Surfaces*, Wiley, 1987.

[91] L. V Candioti, M. M. De Zan, M. S. Cámara, H. C. Goicoechea, Experimental design and multiple response optimization. Using the desirability function in analytical methods development, *Talanta* **124** (2014) 123-138.

[92] F. Schülth, L. A. Baumes, F. Clerc, D. Demuth, D. Farrusseng, J. Llamas-Galilea, C. Klanner, J. Klein, A. Martinez-Joaristi, J. Procelewska, M. Saupe, S. Schunk, M. Schwickardi, W. Strehlau, T. Zech, High throughput experimentation in oxidation calalysis: Higher integration and "intelligent" software, *Catal. Today* **117** (2006) 284-290.

[93] L. A. Baumes, P. Collet, Examination of genetic programming paradigm for high-throughput experimentation and heterogeneous catalysis, *Comput. Mater. Sci.* **45** (2009) 27-40.

[94] H. S. Seung, M. Opper, H. Sompolinsky, Query by committee, in: D. Haussler (Ed.), *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, Pittsburgh, PA, ACM Press, New York, 1992, pp. 287-294.

[95] R. Gilad-Bachrach, A. Navot, N. Tishby, Query by committee made real, *Adv. Neural Inf. Process. Syst.* **5** (2005) 443-450.

[96] D. A. Cohn, Z. Ghahramani, M. I. Jordan, Active learning with statistical models, *J. Artif. Intell. Res.* **4** (1996) 129-145.

[97] S. D. Whitehead, D. H. Ballard, *A Study of Cooperative Mechanisms for Faster Reinforcement Learning*, Univ. Rochester, Dept. Comput. Sci., 1991.

[98] D. Cohn, L. E. Atlas, R. Ladner, M. A. El-Sharkawi, R. J. Marks II, M. E. Aggoune, D.C. Park, Training connectionist networks with queries and selective sampling, in: D. Touretzky (Ed.), *Advances in Neural Information Processing Systems* **2**, Morgan Kaufmann, 1990, pp. 566-573.

[99] P. Juszczak, R. P. W. Duin, On classifier domains of competence. *Proc. 17th Int. Conf. on Pattern Recognition*, IEEE Comp. Soc., Los Alamitos, CA, 2004.

[100] A. Linden, F. Weber, Implementing inner drive by competence reflection, in: H. Roitblat (Ed), *Proceedings 2d International Conference on Simulation of Adaptive Behavior*, MIT Press, Cambridge, 1993, pp. 321-326.

[101] J. Schmidhuber, J. Storck, *Reinforcement Driven Information Acquisition in Nondeterministic Environments*, Tech. Report, Fakultät für Informatik, Technische Universität München, 1993.

[102] S. Thrun, K. Möller, Active exploration in dynamic environments, in: J. E. Moody, S. J. Hanson, R. P. Lippmann (Eds.), *Advances in Neural information Processing Systems* **4**, Morgan Kaufmann, San Mateo CA, 1992, pp. 531-538.