Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Streamlining Catalyst Development through Machine Learning: Insights from Heterogeneous Catalysis and Photocatalysis

Parisa Shafiee[a], Mitra Jafari[a], Julia Schowarte[a], Bogdan Dorneanu[a], Harvey Arellano-Garcia[a,*]

[a] FG Prozess- und Anlagentechnik, Brandenburgische Technische Universität Cottbus-Senftenberg, Cottbus, Germany
* Corresponding Author: arellano@b-tu.de.

ABSTRACT

Catalysis design and reaction condition optimization are considered the heart of many chemical and petrochemical processes and industries; however, there are still significant challenges in these fields. Advances in machine learning (ML) have provided researchers with new tools to address some of these obstacles, offering the ability to predict catalyst behaviour, optimal reaction conditions, and product distributions without the need for extensive laboratory experimentation. In this contribution, the potential applications of ML in heterogeneous catalysis and photocatalysis are explored by analysing datasets from different reactions, including Fischer-Tropsch synthesis and photocatalytic pollutant degradation. First, datasets were collected from literature. After cleaning and preparing the datasets, they were employed to train and test several models. The best model for each dataset was selected and applied for optimization.

Keywords: Catalysis, Machine Learning, Modelling, Optimization, Alternative Fuels, Environment, Fischer-Tropsch Synthesis, Photocatalysis

INTRODUCTION

Developing catalysts (heterogeneous and photocatalysts) and selecting the optimum reaction conditions are considerable challenges. In this process, complex variables need to be addressed, including active material selection, catalyst structure design, or operating parameter optimization [1]. All these tasks require considerable effort in lab and numerous reaction tests, which are not only energy- and time-intensive but also costly. Recent advances in Machine Learning (ML) provide scientists with powerful tools to model and predict the behavior of catalysts, optimize reaction conditions, and forecast product distributions [2]. These methods leverage large datasets and employ sophisticated algorithms, significantly reducing the energy, time, and money spent in laboratory research. By using ML to model and optimize the processes, researchers can now identify patterns, simulate complex reactions, and generate previously unattainable insights. ML enables correlation analysis to uncover relationships between various parameters and catalyst performance. Predictive models, trained on existing data, can forecast the effectiveness of new materials. Additionally, ML provides data-driven insights that offer valuable guidance for catalyst design and optimization [3].

This work introduces a comprehensive framework unifying various aspects of catalyst development and optimization, addressing a significant gap in existing research. The framework's novelty lies in its holistic approach, integrating multiple aspects of catalytic processes, including catalyst formulation, structure, preparation, activation, and operating conditions. Unlike previous studies focusing on individual aspects, this integrated approach provides a more complete understanding of catalyst behavior across diverse catalytic processes: environmental pollutant degradation using photocatalysts, and Fischer-Tropsch Synthesis (FTS) for liquid fuel and jet fuel production, respectively, using heterogeneous catalysts.

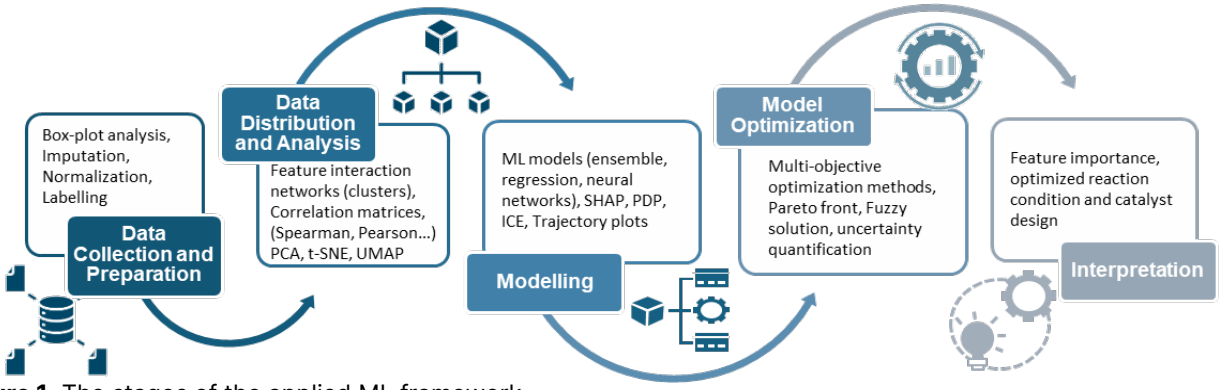

Figure 1. The stages of the applied ML framework.

METHODOLOGY

First the data is collected from literature. The datasets are categorized based on different features like catalyst formulation, catalyst characterization, activation, and reaction conditions with the goal of predicting catalyst performance. To prepare the datasets, the data is cleaned, labelled, imputed and normalized. Techniques such as Spearman correlation matrices, dendrograms, pair plots, and dimensionality reduction methods are applied to uncover relationships between descriptors and catalyst performance. The data is then employed to train and test several models. Performance metrics such as root mean square error (RMSE), mean absolute percentage error (MAPE), Akaike Information Criterion (AIC), and coefficient of determination (R^2) are used to assess model accuracy. Given the non-linear nature of the models, RMSE was selected as the primary evaluation metric for model selection, while R2 is reported as a complementary measure to provide additional context. After selecting the best model, it is used for optimization to find the optimum point with the highest catalysis performance. Figure 1 illustrates this framework, and the subsequent subsections provide a more detailed description of each step in the process.

Data Collection and Preparation

The three datasets used in this work are for perfluorooctanoic acid (PFOA) degradation with photocatalysis, FTS for liquid fuel production, and FTS for jet fuel production.

Table 1. Number of descriptors and target responses in the different datasets of three case studies.

Reaction	Descriptors	Target Responses
PFOA photocatalysis	9	2
FTS for liquid fuel	26	2
FTS for jet fuel	21	2

The data is manually extracted from literature. The derived data is classified as descriptors and target responses as mentioned in Table 1. These descriptors are different parameters describing the catalyst formulation, catalyst characterization, catalyst activation, and reaction conditions.

The data is prepared using standard methods. These steps include removing outliers with box-plot analysis, labeling through label encoding or one hot labeling, imputing missing data using methods such as Soft Impute, k-nearest neighbours (KNN), etc., and normalizing to improve data quality.

Data Distribution and Analysis

Cluster analysis and Spearman correlation analysis were also used to examine the relationship between features.

In the datasets with a large number of datapoints, principal component analysis (PCA) was used to reduce the dimension of the experimental datasets. PCA transforms high-dimensional data into a lower-dimensional space, preserving key patterns and relationships by identifying directions (principal components) that capture maximum variance in the data.

Modelling

Several ML algorithms (models) including Random Forest, CatBoost, Boosted Gradient, Artificial Neural Networks (ANN) etc. are used to predict the target responses. The performance of the models is evaluated and compared using the performance metrics mentioned above. Important features were identified using Shapley Additive Explanations (SHAP) and Partial Dependency Diagram (PDP) techniques.

Optimization

The Non-dominated Sorting Genetic Algorithm II (NSGA-II) was used for multi-objective optimization to find the optimum conversion and selectivity as target responses [4]. This method helps to identify the optimum points on the Pareto front. Subsequently, a Fuzzy-based

method selected a balanced solution using membership functions with parameters [0.2, 0.5, 0.8] from normalized data. Here, 0.2 sets the minimum acceptable value, 0.5 the ideal target, and 0.8 the upper limit. The solution maximizing the minimum satisfaction score ensured both objectives remained strong without trade-offs, aligning with multi-criteria decision-making [5]. The most important descriptors for the best solution are determined and compared with the experimental data.

RESULTS AND DISCUSSION

Photocatalysis of PFOA

The database for the case study of photocatalytical PFOA degradation consists of 127 datapoints, with two target responses and nine descriptors. One-hot labeling is used to convert categorical data to numerical data. For dealing with missing data, five different imputation methods were tested like Stochastic Within-Zone (SWZ), Robust Principal Component Analysis (RPCA), smallest squares, Automatic JMP imputing and mean Python imputing, as well as the removal of these points. Before running the model, the data is also normalized.

The data distribution shows mostly similar patterns, yet some unique distributions result from the different imputation methods. A prominent example (pH) is shown in Figure 2. Further differences include: The smallest squares method leads to additional peaks in higher regions for pH and degradation in comparison to SWZ and RPCA. Automatic JMP and mean Python distributions are highly similar, with a prominent brooding of pH- and defluorination-distributions. The mean Python method shows a distribution for light type and degradation as similar to the method, where all missing values were removed. Yet, this last method shows the broadest distribution for all descriptors and target responses. Thus, it can be concluded that the data distribution is biased.

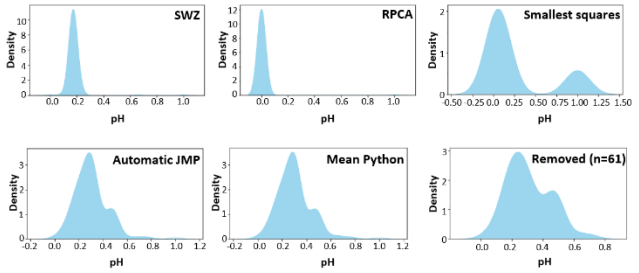

Figure 2. Data distribution with different imputation methods.

Pearson correlation analysis reveals that the imputation method plays an important role in the linearity, as this correlation specifically measures linear relationships between two variables. It can be noticed that there is very little linear correlation in this photocatalytic dataset. In general, the correlation is more positively linear for the defluorination (here correlation to treatment time and pH), which might be attributed to a higher number of missing values. For degradation, the correlation is mostly negatively linear, especially for the variables such as catalyst amount, initial PFOA concentration and in some cases pH.

Spearman correlation analysis is conducted as well. Compared to Pearson, Spearman reveals fewer negative correlations overall, indicating that the data has significant non-linear relationships not captured by the linear correlation. Furthermore, for all imputation methods (beside of missing value removal) a much stronger negative correlation between degradation and initial PFOA concentration is depicted. This is found for the defluorination response as well in Automatic JMP, mean Python and removal methods. For SWZ, RPCA and smallest square imputations no positive correlations, as in the case with Pearson, can be found and treatment time shows a highly negative correlation with defluorination.

The model screening runs 5 times and model optimization is conducted using 10 loops of hyperparameter optimization (GridSearch). Based on high R^2 and low RMSE values. the best result for predicting the degradation response is achieved with the KNN model (SWZ-imputation) and partly with the Neuronal Boosted model (removal of missing data). The defluorination response is captured well with the Neuronal Boosted model (SWZ, RPCA and Automatic JMP imputations).

Best results for the response degradation are obtained with SWZ imputation and the KNN model (R^2=0.843, RMSE=0.004), while the defluorination response is captured well by the SWZ imputation and Neuronal Boosted model (R^2=0.987, RMSE=0.021), as well as by Automatic JMP imputation with the Neuronal Boosted model (R^2=0.998, RMSE=0.009). For comparison, with a simple linear regression model, an average R^2 of 0.8 for the defluorination response can be reached (RMSE and MAE<1, AIC=-2 to 45).

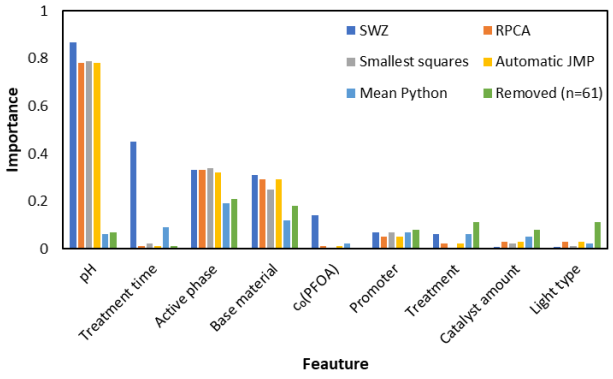

Figure 3. Feature importance with linear regression model for defluorination response with different imputation methods.

The importance of each feature on the defluorination response for the linear regression model is shown in Figure 3, where slight changes can be seen with each imputation method. Main features that influence PFOA defluorination are the pH and the nature of the catalyst (active phase, base material). Linked to the latter is the light type. Interestingly, treatment time plays a minor role in most imputation methods, which highlights the biased data (only small treatment times reported). Beside of the latter (due to bias), the results match other publications on this topic regarding photocatalysis of PFOA [6]. A more theoretical and compound-based approach for the defluorination of Per- and Polyfluoroalkyl Substances (PFAS) has been reported by Raza et al. [7].

FTS for Liquid Fuel Production

A dataset of 96 datapoints is used to model and optimize liquid fuel production using FTS. This dataset includes 26 features consisting of catalyst formulation, characteristics, and operational conditions. As for the targets, CO conversion and C_{5+} selectivity are chosen to represent the liquid fuel production.

To prepare the dataset, after removing the outliers, label encoding is used to change the categorical data to numerical data. Similarly to the previous case study, different methods are used for imputation of the missing data; among them, automated imputation in JMP has shown the best performance and is selected for this dataset. Then data is normalized for the effect of features to be comparable.

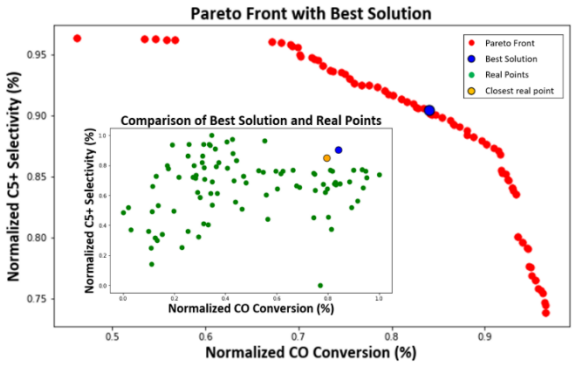

Figure 4. Pareto front with Fuzzy-based solution for normalized CO conversion and C_{5+} selectivity.

To find the best model for describing the behaviour of the dataset, 5 different models are tested and compared, including: XGBoost, LightGBM, Gradient Boosting, CatBoost, and Random Forest. To compare the performance of these models, different criteria, described in the previous sections are used. Among the models, CatBoost had the best performance with R^2, MSE, RMSE, and MAE equal to 0.95, 0.003, 0.057, 0.041, respectively for CO conversion and 0.88, 0.004, 0.063, 0.035 for C_{5+} selectivity.

The CatBoost model is then used for optimization. The aim was to optimize the CO conversion and C_{5+} selectivity. The Pareto front with the optimal Fuzzy solution, as well as the comparison of the best solution and the real data, are depicted in Figure 4.

Since the data was normalized in the preparation step, to recover the real data, de-normalization is carried out. The best Fuzzy solutions for CO conversion and C_{5+} selectivity are 76.24 % and 81.89 %, respectively. The best Fuzzy solution is also compared with the experimental values and the closest point is also detected [8]. The values for these two points are summarised in Table 2.

Table 2. Descriptor values for the best optimum point and closest real point.

Descriptor	Best Solution	Closest Real Point
Support	ZSM-5	ZSM-5
Active metal	Cobalt	Iron
Active metal loading (%)	16.66	18.00
H_2/CO ratio	2	2
Zeolite SiO_2/Al_2O_3 ratio	19.31	280.00
Calcination temperature (°C)	371.55	450.00
Calcination time (h)	3.10	8.00
Pore volume (cm^3/g)	0.58	0.17
Meso-pore volume (cm^3/g)	0.16	0.13
Reduction temperature (°C)	402.66	420.00
Reduction time (h)	1.84	10.00
Reaction temperature (°C)	241.38	280.00
Reaction pressure (bar)	19.91	19.00
Time on stream (h)	67.78	34.46
Loading of catalyst (g)	1.34	0.50
C_{5+} selectivity (%)	81.89	78.20
CO conversion (%)	76.24	72.40

FTS for Jet Fuel Production

To develop a ML method for modelling and optimizing jet fuel production through FTS, a dataset of 41 datapoints from 10 recent research papers was collected. The data was manually extracted and included 21 features related to catalyst formulation and operating conditions. The model targets included C_8-C_{16} selectivity and CO conversion rate.

The data was prepared by removing outliers using box-plot analysis, imputing missing data with methods such as Soft Impute and KNN, and normalization to improve data quality. Cluster analysis and correlation matrix methods were also used to examine the relationship between features.

Several ML algorithms including Random Forest, CatBoost, Boosted Gradient, and ANN were evaluated to predict C_8-C_{16} selectivity and CO conversion. The CatBoost model showed the best performance using metrics such as R^2, RMSE, and MAPE. Important features were

identified using SHAP and PDP techniques.

These analyses showed that the H_2/CO ratio, the reaction temperature, and the catalyst specific surface area had a significant impact on the selectivity and conversion rate. For example, higher reduction temperature and moderate active metal loading led to improved catalyst performance.

The NSGA-II algorithm was used to simultaneously optimize selectivity and conversion. This method helped to identify the optimal points on the Pareto front. Then, the optimal parameters including active metal loading, reduction temperature, reaction temperature, and H_2/CO ratio were determined. The results were evaluated using statistical indices and comparison with experimental data.

Table 3. Value of the features for best solution point

Parameter	Best solution	Closest Real Point
Active metal	Iron	Iron
Promoter	Yes	Yes
Active metal loading (%)	50.5	100.0
Promoter loading (%)	5	6
Average pore size of catalyst (nm)	4.25	4.24
Specific surface area of catalyst (m^2/g)	283.8	14.6
Specific surface area of support (m^2/g)	504.0	458.4
Acidity of support (μmol/g)	1040	465.8
Active metal particle size (nm)	12.30	11.44
Calcination time (h)	4	4
Calcination temperature (°C)	450	450
Reduction atmosphere	H_2/CO	H_2/CO
Reduction time (h)	6.5	10.0
Reduction temperature (°C)	407.5	350.0
Reaction temperature (°C)	265	250
Reaction pressure (bar)	39.45	20.00
Loading of catalyst (g)	1.6	0.5
H_2/CO ratio	1.25	2.00
Time on stream (h)	99	60
C_8-C_{16} selectivity (%)	36.86	30.00
CO conversion (%)	49.32	50.00

Statistical analyses showed that properties such as active metal loading and specific surface area of the catalyst had a significant effect on catalyst performance. For example, increasing the metal loading by about 15 % increased the selectivity, but loading beyond this value led to a decrease in performance. In addition, properties such as support acidity and surface area played a key role in improving the C_8-C_{16} selectivity.

The CatBoost model showed the best performance with an accuracy of R^2=0.82 for CO conversion and R^2=0.84 for C_8-C_{16} selectivity. This model had higher accuracy than other methods such as Neural Networks and Random Forests and was able to better analyze the complex relationships between properties and responses. Also, the use of model interpretability techniques such as SHAP helped to better understand the effect of the properties.

The H_2/CO ratio had a nonlinear effect on selectivity. Increasing this ratio to 2 improved the C_8-C_{16} selectivity, but higher values produced lighter products. Also, the reaction temperature between 245 °C and 260 °C provided the optimal conditions for jet fuel production.

It was found that the specific surface area of the catalyst and the support acidity played an important role in increasing the selectivity. Catalysts with mesoporous structure and moderate acidity performed best. These features helped to improve the chain reactions and reduce light by-products.

Additionally, the Fuzzy satisfaction levels for solutions obtained using a multi-objective genetic algorithm were evaluated. The optimized parameters achieved a MSE of 0.015 and an R^2 of 0.70, indicating high accuracy. The best compromise solution offered 36.48 % C_8-C_{16} selectivity, 48.95 % CO conversion, and a Fuzzy satisfaction level of 0.99, suggesting a well-balanced trade-off. Further refinement used the minimize function from the SciPy optimization library to match target values of 36.48 % C_8-C_{16} selectivity and 48.95 % CO conversion. The optimized input features, shown in Table 3, closely matched these targets with predicted values of 36.86 % and 49.32 %, respectively. This accuracy validates the robustness of the models and the optimization process. The optimal parameters were identified as a reduction temperature of 407.5 °C, a reaction temperature of 265 °C, a reaction pressure of 39.45 bar, and a H_2/CO ratio of 1.25. These values provided a balance between selectivity and conversion. The results obtained were in good agreement with the experimental data and confirmed the effectiveness of the applied framework.

CONCLUSIONS

In this work, an applied ML framework is used for modeling and optimizing different datasets on photo-catalysis and heterogeneous catalysis.

For the photocatalytic degradation of PFOA the importance of data quality is highlighted, as biased data influenced the efficiency of the modelling procedure. Reliable (in-house generated) experimental data for photocatalysis is recommended for ML processes. Regardless of this, the influence of different imputation methods is presented, which can be considered in future applications. Additionally, the resulting feature importance for defluorination is in accordance with recent literature.

Moreover, the applied ML framework provides a

powerful tool for the rational design of FTS catalysts and operating conditions to maximize liquid and jet fuel production, respectively. For liquid fuels, zeolite SiO_2/Al_2O_3 ratio, pore volume, and reaction temperature and pressure are identified as key parameters. In the case of jet fuel production, H_2/CO ratio, temperature, and surface area are key to jet fuel selectivity. Multi-objective optimization was used to identify optimal catalysts and FTS operating conditions.

In the case of liquid fuel production, the CatBoost model is the best model and can predict the dataset satisfactorily. After using this model for optimization, the best solution for maximizing the CO conversion and C_{5+} selectivity are also obtained.

The ML framework not only predicts outcomes but also guides experimental efforts by identifying key parameters for catalyst optimization. It reconciles conflicting trends in literature and contributes to the FTS mechanism debate. Moreover, for experimentalists, it offers a practical tool for understanding performance limits, focusing on achievable improvements, and predicting optimal parameters without extensive trial-and-error.

For the next step, the identified optimal catalysts with the best optimized solutions will be tested experimentally to check their performance. This can be used both for model verification/validation and for actual catalysis optimization. For further work, the suggested framework will also be used for datasets of other catalytic reactions, expanding its applicability and ensuring robustness in diverse reaction conditions.

ACKNOWLEDGEMENTS

This project has been funded as part of the PyroBioFuel project within the Long-Term Joint European Union – African Union Research and Innovation Partnership on Renewable Energy (LEAP-RE), co-financed by the Horizon 2020 programme of the European Union and the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung – BMBF), under grant agreement number 03SF0678 and MyReacAvFu project, financed by the German Federal Ministry of Economic Affairs and Climate Action (Bundesministerium für Wirtschaft und Klimaschutz – BMWK), under grant agreement number 03EI5447A.

REFERENCES

1. Choung S., Park W., Moon J., Han JW. Rise of Machine Learning Potentials in Heterogeneous Catalysis: Developments, Applications, and Prospects. *Chem. Eng. J* 2:152757 (2024) https://doi.org/10.1016/j.cej.2024.152757
2. Benavides-Hernández J., Dumeignil F. From Characterization to Discovery: Artificial Intelligence, Machine Learning and High-Throughput Experiments for Heterogeneous Catalyst Design. *ACS Catal.* 14:11749-79 (2024) https://doi.org/10.1021/acscatal.3c06293
3. Shafiee P., Dorneanu B., Arellano-Garcia H. Improving Catalysts and Operating Conditions Using Machine Learning in Fischer-Tropsch Synthesis of Jet Fuels (C_8-C_{16}). *Chem Eng J Adv* 9:100702 (2025) https://doi.org/10.1016/j.ceja.2024.100702
4. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput 6(2), 182–197 (2002) https://doi.org/10.1109/4235.996017
5. Sánchez-Orgaz, S., Pedemonte, M., Ezzatti, P., Curto-Risso, P. L., Medina, A., & Hernández, A. C. Multi-objective optimization of a multi-step solar-driven Brayton plant. Energy Convers. Manage., 99, 346-358(2015) https://doi.org/10.1016/j.enconman.2015.04.077
6. Navidpour AH., Hosseinzadeha A., Huang Z., Li D., Zhou JL. Application of Machine Learning Algorithms in Predicting the Photocatalytic Degradation of Perfluorooctanoic Acid. *Catal Rev* 66:687-712 (2022) https://doi.org/10.1080/01614940.2022.2082650
7. Raza A., Bardhan S., Xu L., Yamijala SS., Lian C., Kwon H., Wong BM. A Machine Learning Approach for Predicting Defluorination of Per-and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environ Sci Technol Lett* 6:624-9 (2019) https://doi.org/10.1021/acs.estlett.9b00476
8. Baranak M., Gürünlü B., Sarıoğlan A., Ataç Ö., Atakül H., Low Acidity ZSM-5 Supported Iron Catalysts for Fischer–Tropsch Synthesis. *Catal today* 207:57-64 (2013) https://doi.org/10.1016/j.cattod.2012.04.013