Supporting Information

Experimentally-Based Fe-Catalyzed Ethene Oligomerization Machine Learning Model Provides Highly Accurate Prediction of Propagation/Termination Selectivity

Bo Yang,*[a] Anthony J. Schaefer, Brooke L. Small,[b] Julie A. Leseberg,[b] Steven M. Bischof,[b] Michael S. Webster-Gardiner,*[b] and Daniel H. Ess*[a]

[a]Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah 84602, United

[b]Research and Technology, Chevron Phillips Chemical Company LP, 1862 Kingwood Drive, Kingwood, Texas 77339, United States

AUTHOR INFORMATION

Corresponding Author

*E-mail: b.yang3227@gmail.com
*E-mail: webstm@cpchem.com
*E-mail: dhe@byu.edu

TABLE OF CONTENTS

1. Machine Learning Workflow

The Fe-catalyzed ethylene oligomerization machine learning model was designed and optimized following a workflow (**Figure S1**) that consisted of four main steps:

1) Collection of experimental data.
2) Feature design, generation, and selection.
3) Model training and validation.
4) Evaluation of machine learning algorithms and features.

Steps 2 to 4 are repeated until the accuracy of the machine learning model is maximized with the chosen data set from step 1. Details regarding each step are provided below in subsections.

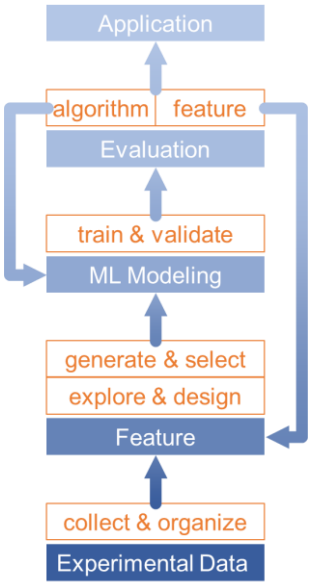

Figure S1. Workflow for experimentally-based Fe-catalyzed ethylene oligomerization machine learning model.

1.1 Experimental data and machine learning data set

All Fe-based ethylene oligomerization catalysts (**Figure S2**) and their corresponding *K* values were collected from litereature as sited in the main text of the manuscript. The selected Fe catalysts mainly consist of four types of ligand motifs, including the pyridine-bisimine, the phenanthroline, the iminopyridine, and the α-diimine ligands.

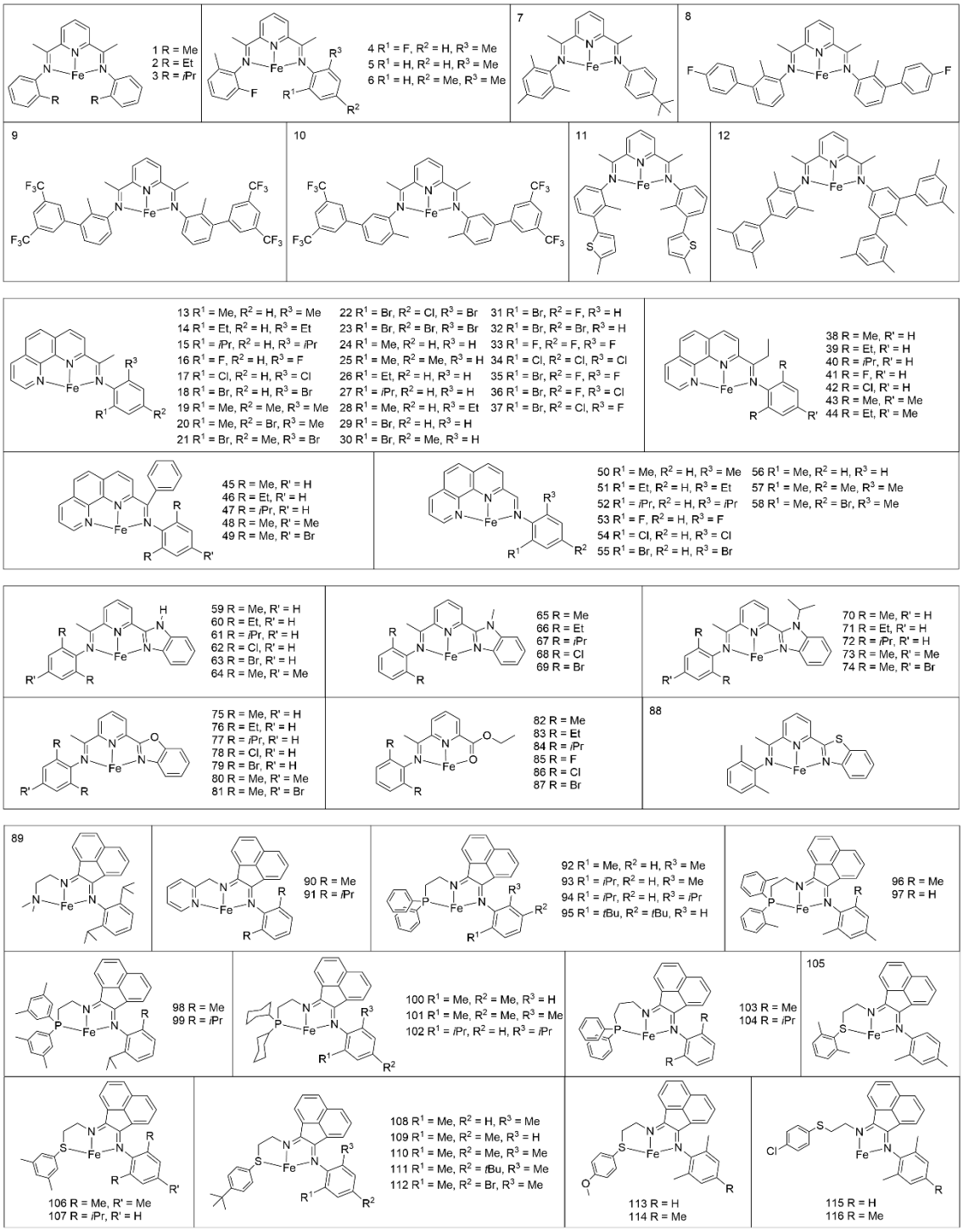

Figure S2. Fe-based ethylene oligomerization catalyst data set. Catalysts are organized based on ligand type, from top to bottom, include the pyridine-bisimine, the phenanthroline, the iminopyridine, and the α-diimine ligands.

The catalyst data was is stored as a Comma Separated Values (CSV) file that was accessed by the Scikit-Learn[1] python library, and consisted of the following entries:

1) 2-dimensional (2D) catalyst structures stored in the Simplified Molecular-Input Line-Entry System (SMILES) format.
2) The experimental $K(C_{12}/C_{10})$ value for each structure; $K(C_{14}/C_{12})$ values were converted to $K(C_{12}/C_{10})$ values when applicable.
3) Reaction conditions reported along with each *K* value for each catalyst. Reaction conditions are used as chemical features in the machine learning model.
4) 2D molecular feature values generated based on 2D structures.

1.2 Features

More than 1500 different 2D molecular features were generated using programs MordRed[2] and RDKit[3] for the catalyst data set. To increase model efficiency, redundant and unrelated features in the machine learning model were removed following four steps:

1) A feature was removed if non-numerical value is generated for any structure within the machine learning data set. This step removed around 400 features.
2) A feature was removed if its corresponding normalized feature importance determined from random forest model was lower than 0.005. Additional 1100 features were removed based on this criterion.
3) A feature was removed if it had high correlation with another feature and if the feature had lower importance. We considered two features highly correlated with each other if the standard correlation coefficient were larger than 0.83. This step removed an additional 6 features.
4) 20 2D features were kept in the model after steps 1 to 3. We then manually selected combinations of features to increase model accuracy based on computed averaged RMSE value. Thirteen additional features were removed, most of which has normalized feature importance lower than 0.015. A correlation heatmap of the final included features is shown in **Figure S3**.

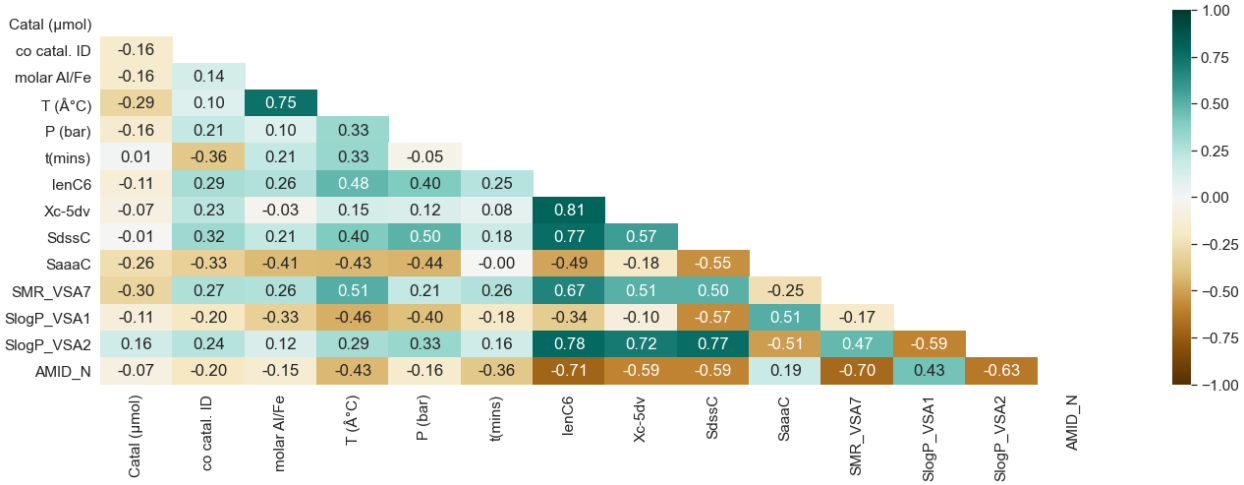

Figure S3. Correlation heatmap of the 14 features used in the machine learning models. The standard correlation coefficient measures the strength and the direction of the linear relationship between two features: a correlation of 0 indicates no linear relationship, a correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation.

1.3 Model training and results evaluation

With the exception of the graph neural network, all machine learning models were trained against experimental *K* values and selected features via the Scikit-Learn[4] Python library. The GNN model was trained using the Spektral Python library. To train a machine learning model, the data set is first randomly divided into training (80% of the data set) and testing (20%) sets. Ten machine learning algorithms (Figure 4 of the main text) were trained using the training set; the accuracy of each algorithm was then evaluated using the testing set. This process was repeated 100 times, each with randomly chosen training and testing sets to avoid overfitting an algorithm to the data set. Overfitted machine learning model, though give high accuracy for the given testing set, often generate poor prediction for new structures outside the

original data set. Feature importance is generated after model training and is used for feature selection as mentioned in the section 1.2 above.

2. Connective Steric Factors (CSF)

The determination of length_C*n*, width_C*n*, depth_C*n* (*n* = 2, 3, 4, 5, or 6) for a given Fe catalyst is described here using several examples.

We first consider complex 49 which has one ligand arm (Figure S4a and Figure S2) which is constituted of one aryl group. Since a methyl (–Me) group occupies the C2 position of the substituted aryl group (Figure S4a), we define features length_C2, width_C2, and depth_C2 to be respectively the length, width, and depth of a methylbenzene molecule (Figure S4b and c). Similarly, length_C4, width_C4, and depth_C4 are defined as respectively the length, width, and depth of a bromobenzene molecule because a bromo group is at the C4 position of phenyl arm. Since the C6 position is also occupied by a methyl group (Figure S4(a)), features length_C6, width_C6, and depth_C6 are the same as length_C2, width_C2, and depth_C2, respectively. For C positions (C3 and C5) where there are no substitution groups, the dimension of a benzene molecule is used.

Note, to determine the bulkiness for the C*n* position (*n* = 2, 3, 4, 5, or 6) of a ligand arm, we use the dimensions of a substituted benzene molecule, instead of the substitute group alone. This is done to ensure the measured length_C*n* always follow the general direction of the phenyl–substitute bond (Figure S4a). To measure the dimensions of substituted benzene molecule, geometry optimizations were performed for the molecule using the PM7[5] semiempirical method as implemented in MOPAC[6] package. Among all the Fe catalysts contained in the data set, there are thirteen different substitution groups that appeared on the aryl group of ligands (Figure S5). Their corresponding dimensions are predetermined and tabulated for the easy calculation of length_C*n*, width_C*n*, depth_C*n* (*n* = 2, 3, 4, 5, 6) for all the Fe catalysts.

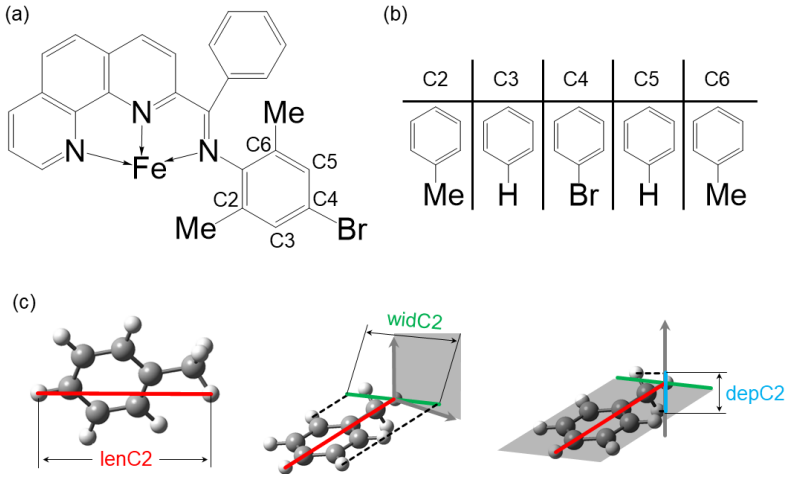

Figure S4. (a) Complex 49; (b) Molecules used for every C position of the phenyl ring to determine the steric bulkiness; (c) Illustrations of the length (length_C2), width (width_C2), and depth (depth_C2) of a methylbenzene molecule. (The length is the maximum distance between any pair of atoms in the molecule; the width is the maximum distance in the plane perpendicular to the length direction between any pair of atoms; the depth is the maximum distance between any two atoms on the line perpendicular to the plane of the length and width dimensions.)

For complexes where there are more than one ligand arm (e.g. complex 1, Figure S2), length_C*n*, width_C*n*, and depth_C*n* (*n* = 2, 3, 4, 5, or 6) were determined via two steps: (1) for each arm *i* (*i* = 1, 2, or 3), $\mathrm{length\_C}_n^i$, $\mathrm{width\_C}_n^i$, and $\mathrm{depth\_C}_n^i$ are first determined; (2) length_C*n* is defined as the sum of $\mathrm{length\_C}_n^i$ (*i* = 1, 2, or 3); width_C*n* and depth_C*n* are defined in the similar manner. For complexes with cyclohexyl arms (complexes 94 to 96, Figure S2), the dimension of a cyclohexane molecule in the chair conformation is used to determine length_C*n*, width_C*n*, and depth_C*n*.

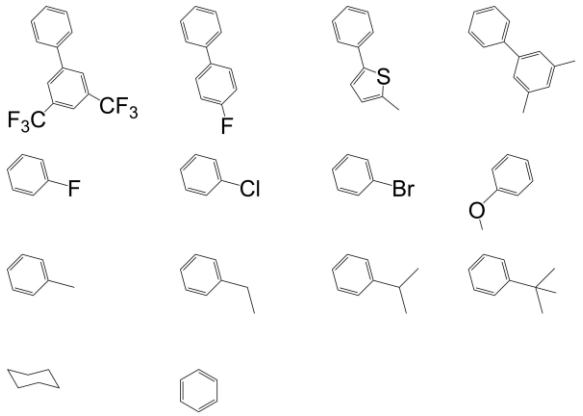

Figure S5. Fourteen molecules used for generating the characteristic size of substitution groups on the polydentate ligand arms.

3. Distribution of Errors for RF Models

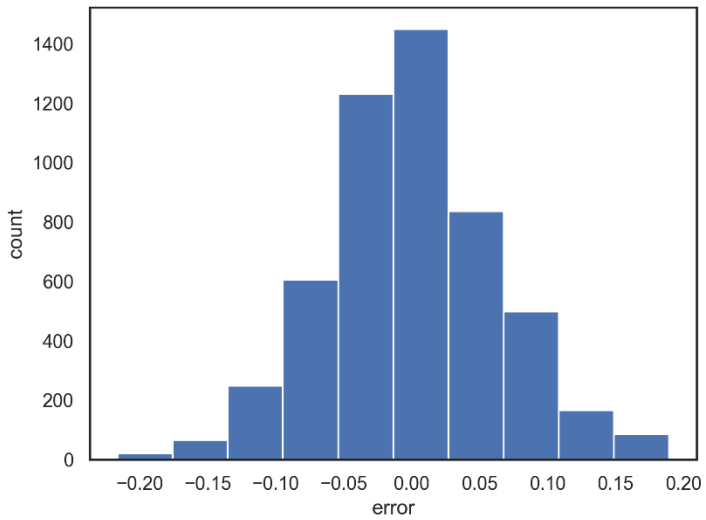

Figure S6. Plot of distribution of errors for the RF model trained without reaction condition features across 100 models.

4. Details of Multi-Layer Perceptron Model

A multi-layer perceptron model, a feedforward artificial neural network, was also trained to predict *K* values using the same descriptors as other models. This model used one hidden layer with 100 nodes. Adding more layers and/or nodes did not yield a significant improvement. The logistic activation function significantly outperformed the rectified linear unit (relu) activation function. For all other hyperparameters, we used the default values for scikit-learn's MLPRegressor. The performance of this model averaged over 100 training runs is summarized in the table below.

Table S1. Table of performance metrics for the MLP model

metric	**value**
MAE	0.100 ± 0.002
RMSE	0.124 ± 0.003
R^2	0.11 ± 0.04

5. *K* Values Prediction when a Catalyst Class was Excluded

A RF model was trained using all available data, except the pyridine bisimine catalysts. These were excluded to see how well the model can predict the *K* value of an unseen category of ligands. This model only utilized features based on the structure of the catalyst. The result was significant errors for most pyridine bisimine catalysts. Randomly including three pyridine bisimine catalysts in the training data while excluding the rest noticeably improved the predictions for the pyridine bisimine catalysts that were excluded from the training data.

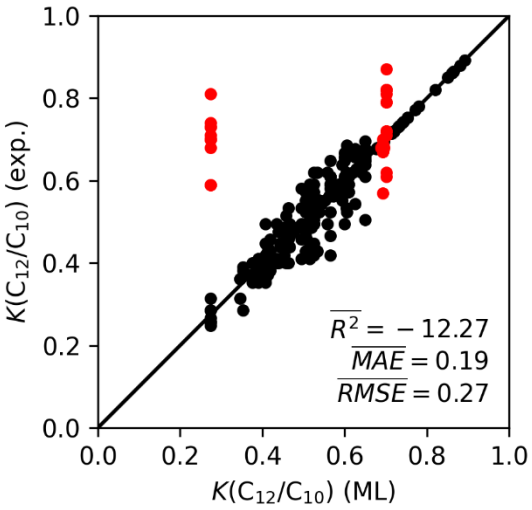

Figure S7. *K* value predictions vs experiment for pyridine bisimine (red), which were excluded from the training data, and all other catalysts (black).

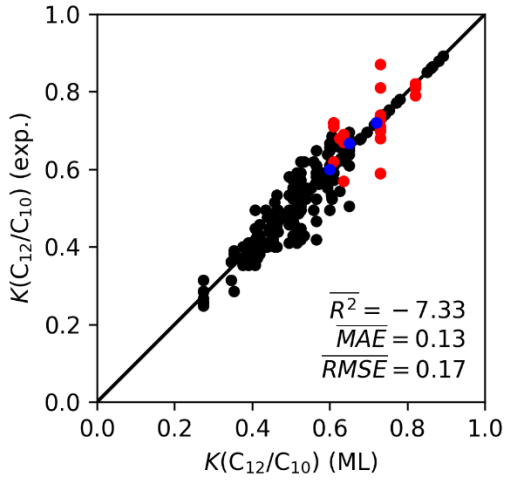

Figure S8. *K* value predictions vs experiment for pyridine bisimine excluded from the training data (red), pyridine bisimine included in the training data (blue), and all other training data (black).

6. *K* Value Prediction for Catalysts with Methyl, Ethyl, Isopropyl Groups

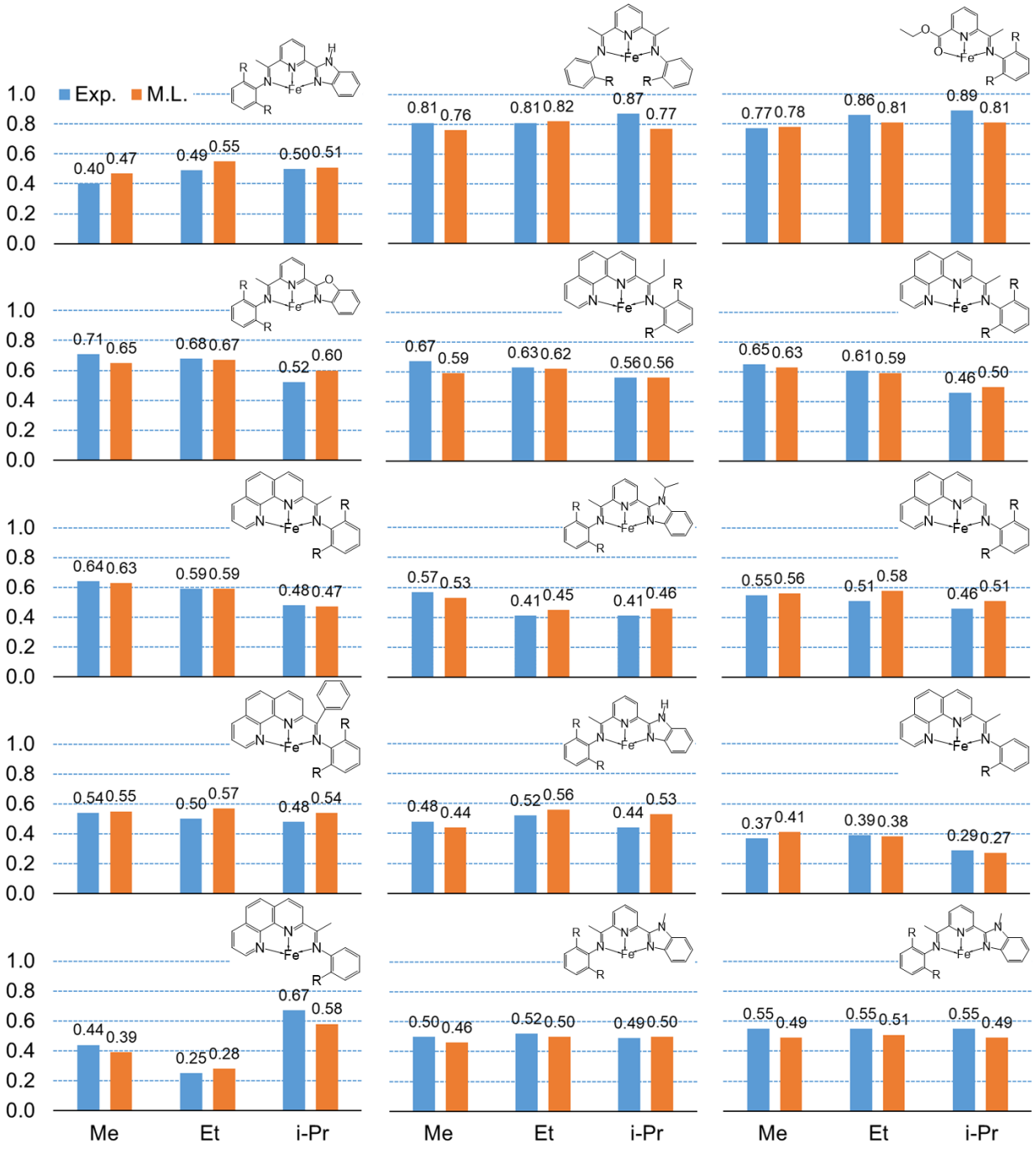

Figure S9. Effect of substitution groups methyl (–Me), ethyl (–Et), and isopropyl (–iPr) on Fe-based ethylene oligomerization catalysts performances. The machine learning predicted $K(C_{12}/C_{10})$ values (orange) are compared with experimental measurements (blue).

7. REFERENCES

1 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 and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

2 H. Moriwaki, Y.-S. Tian, N. Kawashita and T. Takagi, *J. Cheminform.*, 2018, **10**, 4.

3 G. Landrum, RDkit: Open-source cheminformatics, http://www.rdkit.org.

4 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 and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

5 J. J. P. Stewart, *J. Mol. Model*. 2013, **19**, 1−32.

6 J. J. P. Stewart, MOPAC2016 (version 20.284W), Stewart Computational Chemistry, Colorado Springs, CO, USA, 2016. https://OpenMOPAC.net/.