Electronic Supplementary Material (ESI) for Reaction Chemistry & Engineering.

Supplementary Information for

Digital Pareto-Front Mapping of Homogeneous Catalytic Reactions

N. Orouji[1], J. A. Bennett[1], S. Sadeghi[1] and M. Abolhasani[1]*

[1]*Department of Chemical and Biomolecular Engineering, North Carolina State University, USA*

Correspondence to: abolhasani@ncsu.edu

S1. Chemicals

All chemicals were used as received unless otherwise specified. 1-Octene, 99+% was purchased from Thermo Scientific Chemicals. Acetylacetonato dicarbonyl rhodium(I) ($Rh(CO)_2(acac)$), 97%, was purchased from Alfa Aesar. Toluene (Anhydrous 99.8%) and tridecane analytical standard were purchased from MilliporeSigma. Carbon monoxide, hydrogen, and nitrogen cylinders were purchased from Airgas at 99.9% purity. The bulky fluorophosphite ligand (L) was provided by Eastman Chemical. Synthesis details and spectra of ligand L are reported in prior work.[1,2]

S2. Experimental Setup

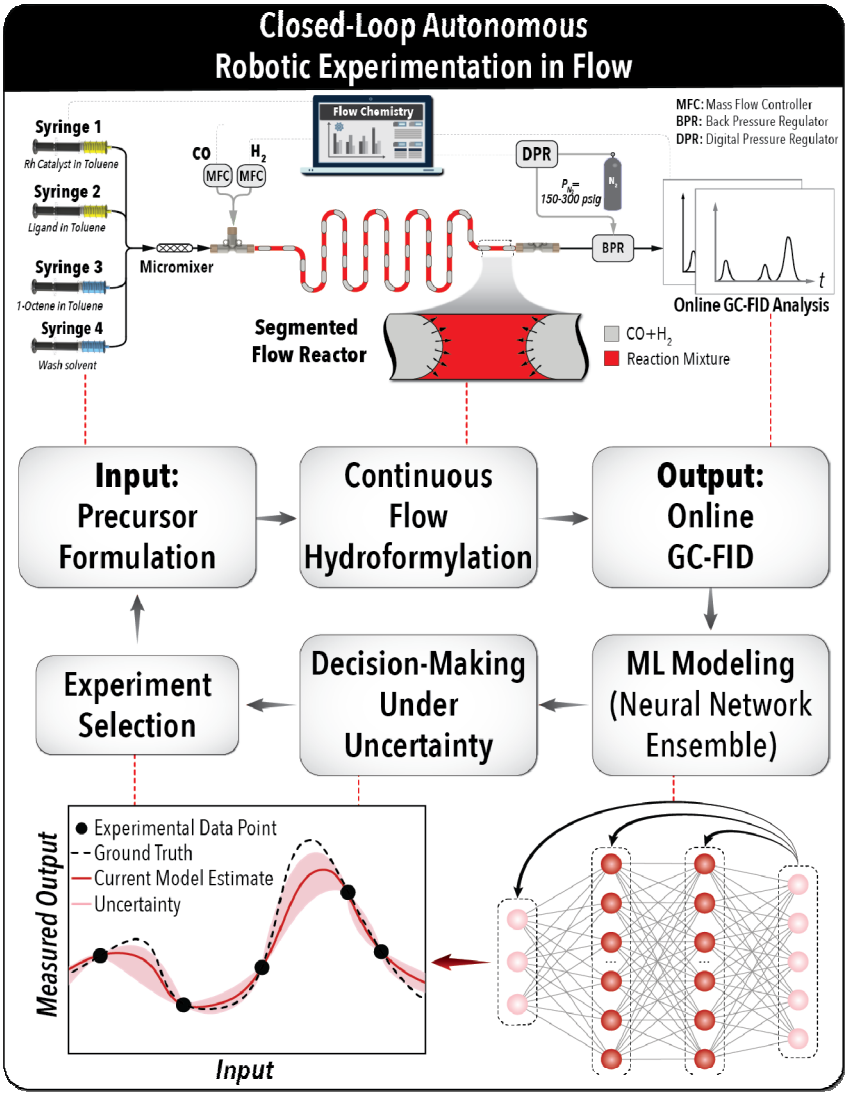


Figure S1. *Schematic illustration of the closed-loop autonomous hydroformylation platform utilized to generate experimental training data for the ML model. Reaction conditions are executed and analyzed automatically, then the experimentally generated data is fed to the model training and new experiment selection algorithms.*

Figure S2. *A picture of the flow chemistry platform utilized to generate the initial experimental training data. The setup includes a liquid delivery module, a reactive gas manifold, a precursor refilling module, two heated coil flow reactors, and an on-line sampling module (GC-FID).*

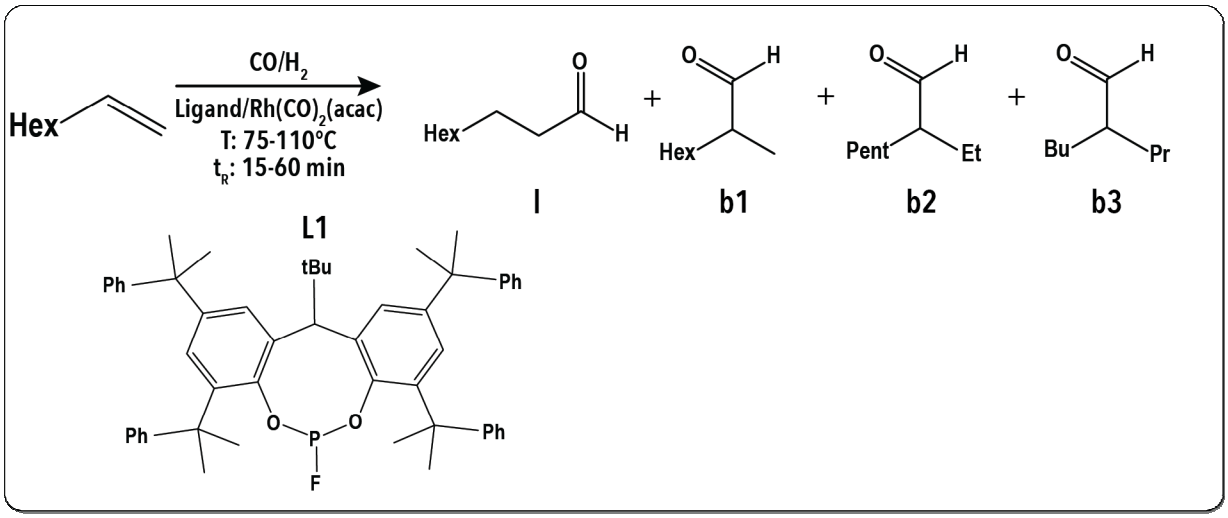
The automated flow chemistry platform consists of six primary modules, including fluid delivery with automated refilling, syngas delivery, heated flow reactor coils, in-line reaction sampling, and in-line characterization using flame ionization gas chromatography (GC-FID). The flow reactor is supplied with four separate stock solutions to independently control concentrations of each reagent species and set the next hydroformylation condition generated by the machine learning (ML) module. The components present in the liquid streams are: solvent stream (solvent + internal standard), olefin stream (olefin + solvent + internal standard), ligand stream (ligand + solvent + internal standard), and metal catalyst stream (rhodium, Rh, salt + solvent + internal standard). The stock solutions are created so an equal volumetric flowrate of all streams results in a standard concentration of 0.5M 1-octene, 0.25 mM $Rh(CO)_2(acac)$, 2.5 mM Ligand, and 20 mM tridecane internal standard in toluene as the solvent (**Table S1**).

Table S1. *Composition of liquid feed solutions.*

Precursor	**Stream Composition**
Solvent	20 mM tridecane in toluene
Olefin	2.0M 1-octene, 20 mM tridecane in toluene
Ligand	10 mM ligand, 20 mM tridecane in toluene
Rhodium	1 mM $Rh(CO)_2(acac)$, 20 mM tridecane in toluene

S3. Reaction Scheme

The hydroformylation reaction occurs with 1-octene in the presence of high pressure carbon monoxide (CO) and hydrogen (H_2, 150-300 psig) with a phosphorous-based ligand and Rh salt. The formyl group can be added to either of the atoms of the double bond depending on the selectivity of the reaction. Competing olefin isomerizations can form internal olefins and certain ligands including L can perform hydroformylation on the internal olefins resulting in further branched products.



Scheme S1. *Hydroformylation of 1-octene to linear (l) and branched (b) aldehyde products with fluorophosphite ligand L.*

S4. Model Inputs and Normalization.

Table S2. *Model input normalization for non-dimensionalized parameters between 0 and 1. Upper and lower bounds are generally hardware (X_1 and X_2: MFC Flow limits) or system limitations ($X_{5\text{-}7}$: stable multi-phase flow).*

	X_1 **CO**	X_2 **H_2**	X_3 **Pressure**	X_4 **Temp.**	X_5 **Dilution**	X_6 **L:Rh**	X_7 **Olefin:Rh**
Min	0.1 mLn/min	0.1 mLn/min	150 psig	75 °C	0.1	0.2	0.2
Max	1.9 mLn/min	1.9 mLn/min	300 psig	110 °C	0.4	0.8	0.8
Formula	$\frac{X-0.1}{1.9-0.1}$	$\frac{X-0.1}{1.9-0.1}$	$\frac{X-150}{300-150}$	$\frac{X-75}{110-75}$	$\frac{X-0.1}{0.4-0.1}$	$\frac{X-0.2}{0.8-0.2}$	$\frac{X-0.2}{0.8-0.2}$
Def.	MFC Flowrate	MFC Flowrate	Total Pressure	Reactor Temperature	Solvent Stream Fraction	Ligand Stream Fraction	Olefin Stream Fraction
					$\frac{\dot{V}_{Solvent}}{\dot{V}_{liquid}}$	$\frac{\dot{V}_{Ligand}}{\dot{V}_{Ligand}+\dot{V}_{Rh}}$	$\frac{\dot{V}_{Olefin}}{\dot{V}_{Olefin}+\dot{V}_{Rh}}$

Model parameters were normalized by hardware or system limitations to limit the potential reaction space to physically accessible new experimental conditions. The other option for defining the gas phase stream composition (total flow rate, pressure, and CO:H_2 ratio) results in physically inaccessible regions of parameter space present in the model input space. For example, at a desired gas flowrate of 3.8 mLn/min, with current hardware, it is only possible to have a 1:1 CO:H_2 ratio as both MFCs would be operating at their maximum value, whereas a total flowrate of 2 mLn/min could allow for ratios of 1:19 to 19:1 (0.1 mLn/min CO and 1.9 mLn/min H_2 and vice versa) and as the total flowrate drops down to 0.2 mLn/min, again, only a 1:1 syngas ratio becomes possible as the MFCs are both operating at their minimum value. In a pressurized reactor with a fixed volume and a gas-to-liquid volumetric ratio of 3:1, the residence time along with the compositions, flowrates, temperature, and pressure fully defines both the gas and liquid streams. A consequence of the selected parameters is that X_1-X_3 contribute to the total residence time in the reactor either as a greater total flow rate or expansion/compression of the gas phase within the reactor.

S5. Experimental Data

Table S3. *Summary of experimental conditions autonomously selected and tested for accelerated Pareto-front mapping of ligand L. Init: initialization run; L: Linear aldehyde campaign; B: Branched aldehyde campaign.*

X_1 CO	X_2 H_2	X_3 Pressure	X_4 Temp.	X_5 Dilution	X_6 L:Rh	X_7 Olefin:Rh	Y Yield	S_N Selectivity	Opt.
0.309	0.309	0.347	0.000	0.500	0.500	0.500	0.914	0.528	Init.*
0.674	0.674	0.347	1.000	0.500	0.500	0.500	0.684	0.416	Init.*
0.309	0.309	0.347	1.000	0.500	0.500	0.500	0.770	0.376	Init.
0.431	0.917	0.347	0.286	0.500	0.500	0.500	0.687	0.571	Init.*
1.000	0.806	0.966	0.252	1.000	0.096	0.428	0.784	0.626	L*
1.000	0.000	0.359	0.723	0.000	0.343	1.000	0.081	0.709	L
0.793	1.000	0.819	0.000	0.395	0.992	0.980	0.254	0.673	L*
0.000	1.000	1.000	0.000	0.739	0.428	1.000	0.446	0.823	L*
0.504	0.000	0.000	0.119	0.876	1.000	0.869	0.101	0.688	L*
0.999	0.882	0.567	0.481	0.984	1.000	0.320	0.721	0.662	L*
0.000	1.000	0.897	0.990	0.282	1.000	0.591	0.747	0.603	L*
0.263	0.874	0.991	1.000	1.000	0.845	0.597	0.976	0.473	L*
0.000	1.000	0.387	0.000	1.000	1.000	0.176	0.603	0.813	L*
1.000	0.901	1.000	0.000	0.388	1.000	0.000	0.890	0.652	L*
0.748	1.000	1.000	0.825	1.000	0.879	0.000	0.915	0.526	L
0.093	0.539	0.000	0.000	0.861	1.000	1.000	0.541	0.770	L*
0.911	1.000	0.813	0.000	0.944	0.500	0.949	0.218	0.678	L
0.802	0.719	0.000	0.000	1.000	0.595	1.000	0.038	0.701	L*
0.000	1.000	1.000	0.000	0.700	0.838	0.921	0.530	0.838	L
0.090	1.000	0.668	0.000	0.287	1.000	1.000	0.611	0.789	L*
0.000	1.000	0.000	0.458	0.769	0.947	1.000	0.178	0.844	L*
0.000	1.000	1.000	0.000	1.000	1.000	0.900	0.582	0.851	L*
0.936	1.000	1.000	0.773	0.372	1.000	1.000	0.788	0.563	L
1.000	1.000	0.000	0.070	0.000	0.979	0.101	0.340	0.717	L*
0.202	1.000	0.123	1.000	0.918	1.000	0.533	0.432	0.593	L*
1.000	0.707	1.000	0.741	1.000	0.822	0.000	1.000	0.540	L*
0.411	1.000	0.958	0.961	1.000	1.000	0.773	0.851	0.518	L*
0.000	1.000	1.000	0.825	0.000	1.000	1.000	0.561	0.751	L*
1.000	1.000	0.603	0.859	0.542	0.000	0.038	0.776	0.506	L*
0.070	1.000	0.337	0.000	0.392	1.000	1.000	0.525	0.832	L*
0.000	1.000	0.617	0.075	0.113	0.719	0.977	0.362	0.857	L*
0.000	1.000	0.979	0.000	1.000	1.000	0.480	0.682	0.802	L*
1.000	0.000	0.466	0.338	0.922	1.000	1.000	0.105	0.701	L
1.000	0.855	0.000	0.758	1.000	0.827	1.000	0.409	0.647	L

0.439	0.542	1.000	0.638	0.844	0.217	0.000	1.000	0.504	B*
0.533	0.790	1.000	0.528	1.000	0.218	0.000	0.988	0.514	B
0.940	0.629	1.000	0.509	1.000	1.000	0.148	0.834	0.591	B*
0.372	0.586	1.000	0.941	0.392	0.000	0.000	1.000	0.477	B
0.228	1.000	0.712	0.683	0.532	1.000	0.112	0.880	0.578	B*
1.000	0.971	1.000	1.000	0.269	0.847	0.000	0.934	0.506	B
0.439	1.000	0.529	1.000	1.000	1.000	0.337	0.664	0.548	B*
0.550	1.000	0.763	1.000	1.000	0.757	0.221	0.945	0.483	B*
0.360	1.000	1.000	1.000	1.000	0.000	0.132	0.920	0.470	B*
1.000	0.000	1.000	0.356	1.000	0.377	0.000	0.820	0.555	B*
0.815	0.631	1.000	1.000	0.509	0.000	0.000	0.893	0.467	B
0.699	1.000	1.000	0.922	1.000	0.841	0.000	0.962	0.491	B*
0.933	0.932	1.000	0.661	0.041	0.558	0.000	0.994	0.517	B
0.402	0.597	0.977	1.000	0.662	0.282	0.000	0.995	0.463	B*
0.455	0.951	1.000	1.000	0.975	0.384	0.000	1.000	0.446	B*
1.000	1.000	0.000	0.000	0.000	1.000	1.000	0.049	0.712	B*
1.000	0.841	0.066	0.003	0.797	1.000	0.995	0.046	0.702	B*
0.000	1.000	1.000	1.000	0.905	0.365	0.586	0.713	0.551	B
0.922	0.823	0.429	0.852	1.000	1.000	1.000	0.685	0.561	B*
0.074	0.997	1.000	0.000	1.000	0.809	0.152	0.897	0.669	B*
1.000	1.000	1.000	1.000	0.956	0.146	0.074	0.895	0.469	B*
1.000	1.000	1.000	1.000	0.467	0.000	0.000	0.894	0.480	B*
0.303	0.712	1.000	0.000	0.637	0.842	0.742	0.946	0.613	B*
0.379	1.000	1.000	0.147	0.000	1.000	0.000	0.898	0.642	B*
0.634	1.000	0.886	1.000	1.000	1.000	1.000	0.781	0.500	B*
0.497	0.493	0.924	0.135	1.000	0.701	0.922	0.904	0.619	B*
0.635	0.336	0.890	0.172	1.000	1.000	1.000	0.639	0.678	B*
0.178	0.814	1.000	0.777	1.000	1.000	0.232	0.888	0.555	B*
0.000	1.000	1.000	0.052	1.000	0.631	0.789	0.650	0.808	B
0.685	1.000	0.875	0.000	1.000	0.234	0.497	0.889	0.635	B
1.000	0.822	1.000	0.000	1.000	1.000	1.000	0.090	0.663	B

*The data set used for training the model. The remaining data was allocated for testing and validation purposes.

S6. Machine Learning Model Hyperparameter Tuning

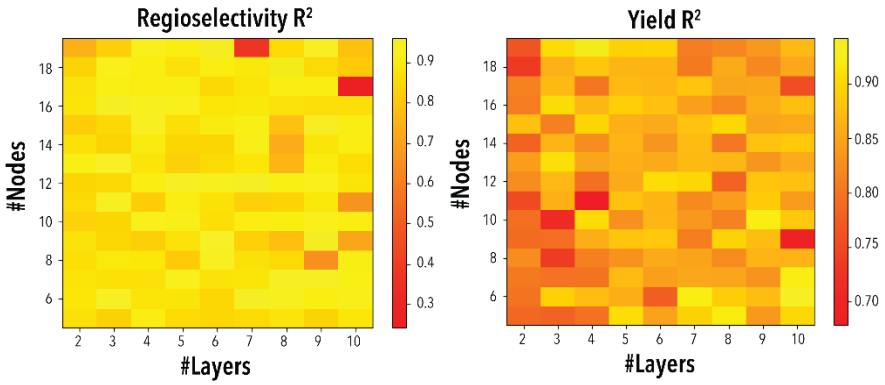


Figure S3. *The ML model architecture study on the cascade forward NN.*

This section discusses the construction of a ML model of the hydroformylation reaction. To select the most suitable model for this problem, we conducted a comprehensive benchmarking analysis, comparing the performance of linear regression, K-nearest neighbors (KNN), and Random Forest models against an ensemble of neural networks (ENNs). In **Table S4**, we present the R^2 values, summarizing the performance of each model on both the training and evaluation datasets.

To optimize the Random Forest model, we performed a grid search to identify the best architecture. It is worth noting that ENNs outperformed both linear regression and KNN models. While the performance of the optimized Random Forest model and ENNs is comparable, we opted for ENNs due to their versatility in handling various types of inputs. ENNs can not only handle tabular data but can also accommodate graph representations of molecular structures, thus allowing facile incorporation of additional features with the presented model in this work in the future. This ML model, which is an ensemble of deep neural networks (ENNs), is built using the experimental data (provided in S5) generated by the experimental setup shown in **Fig. S2**. The ML model is designed to mimic the intricate dynamics of the reaction space, thereby providing a robust tool for analysis and prediction. It leverages the power of ENNs to learn from the experimental data and generate insights about the reaction space.

To assess the performance of the ML model, we utilized the coefficient of determination (R^2). This metric is derived from the parity plot of the ML model validation. The R^2 value quantifies how well the regression model fits the data, serving as an indicator of the model's explanatory power.

The coding and training of the ML model were executed in Python 3.9.13, utilizing libraries such as numpy, pandas, and tensorflow. In the initial step, we conducted a comparative analysis of the performance between a feedforward and a cascade forward NN. The cascade structure has previously been demonstrated to exhibit superior performance compared to the feedforward structure for modelling in sparse data environments. [2] The superior performance of the cascade NN architecture can be attributed to its multi-stage feature extraction, error propagation, ability to manage high-dimensional data, non-linear mapping capabilities, and resilience to data noise. The objective of the NN architecture search was to rapidly identify the most suitable number of layers and nodes, by maximizing the coefficient of determination. **Fig. S3** illustrates the R^2 values of aldehyde yield and regioselectivity (in the presence of ligand L), derived from the complete dataset for an ensemble of 5 cascade feed forward NN models. Each model comprises n layers (represented on the x-axis) and each layer consists of m nodes (represented on the y-axis).

ENN structure was specifically selected to incorporate an element of uncertainty into the ML model representation of the hydroformylation reaction studied in this work. The development of the ML model entailed the random selection of the number of layers and nodes from a predefined range established in the preceding step (see **Fig. S3**). To optimize the performance of the ML model, an outer NN with two layers was introduced to aggregate the outputs of each individual model instead of linearly averaging in to improve model prediction.[3] The influence of the ensemble size on the ML model's performance was investigated, utilizing the average R^2 value derived from 50 ensemble NNs of equivalent size as the benchmark metric. As depicted in **Fig. S4**, an ensemble size of 10 cascade NNs, chosen randomly, exhibited satisfactory performance (R^2>0.7) when applied to experimental. A further increase in ensemble size was observed to lead to prolonged training times without yielding additional benefits. In order to further enhance the performance of the ML model, various hyperparameters such as the learning rate and optimizer were fine-tuned, with the aim of achieving a smoothly descending learning curve (see **Fig. S5**). Given the regression nature of the task, the loss function was set as the root-mean-squared error, which the ENN model aimed to minimize during the training process. The rectified linear unit (ReLU) activation function was applied to all hidden layers, while the output layers used a linear activation function. The *RMSprop* optimizer was used for optimization. A tolerance level was set on the output layer that rounded results higher than one to one ensuring valid predictions, as the fractional aldehyde yield and regioselectivity values should not exceed one. This step was taken to prevent the model from making predictions outside the feasible experimental range and to comply with the physical constraints of the hardware.

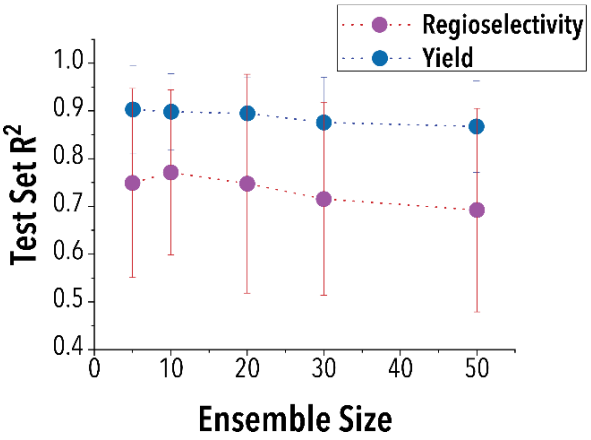


Figure S4. *Average performance of an ensemble of 50 randomly selected cascade NNs of a size 5- 50.*

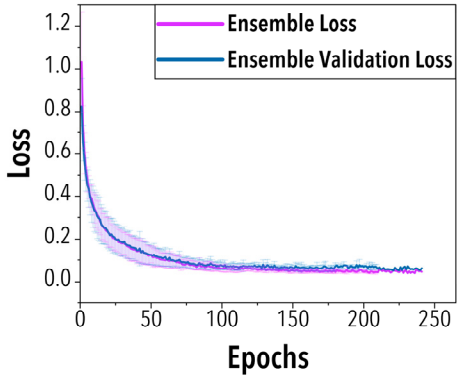


Figure S5. *The Average learning curve (i.e., the improvement in the ML model's performance) of the ensemble model over time. To address the variability introduced by shuffling inputs, the simulation was executed five times. The graph displays the mean outcomes and their associated uncertainty. Early stopping will monitor changes in the validation set error to avoid overfitting. and after passing 20 epochs without significant improvement, it will terminate the training process.*

The ML model was then built using 65 in-house generated experimental data, with 75% data allocated for training, 15% for validation, and the remaining 10% for testing. The test dataset, was not seen by the ML model during training. An early-stopping function was used to monitor the loss function of the validation set to avoid overfitting. Overall, test and validation set together was used to evaluate the performance of the ML model in the reaction space. **Fig. S6** provides a visual representation of the ML model's performance, showing the average of the model-predicted values on the training and evaluation dataset compared to the actual values obtained from the experiments. **Table S4** summarizes R^2 values for training and evaluation dataset. Finally, we have provided a simple representation of our ML model in **Fig. S7**. This illustrative representation serves as a visual aid to enhance the understanding of our model's architecture and its components.

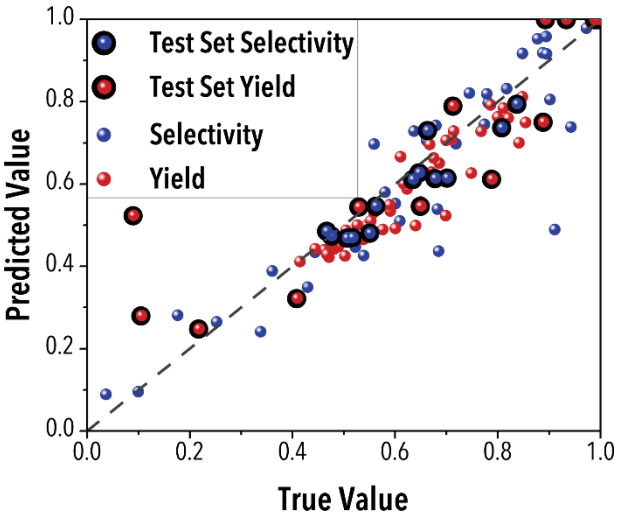


Figure S6. *Predicted values by the ML Model vs. the actual values from the experiment with ligand L for training and test set.*

Table S4. *R^2 Values for training and evaluation dataset*

R^2				
Data / Model	**Yield**		**Regioselectivity**	
	Training	Evaluation	Training	Evaluation
Linear Regression	0.71	0.64	0.75	0.61
KNN	0.73	0.70	0.81	0.83
Random forest	0.98	0.55	0.99	0.81
ENN	0.89	0.80	0.93	0.91

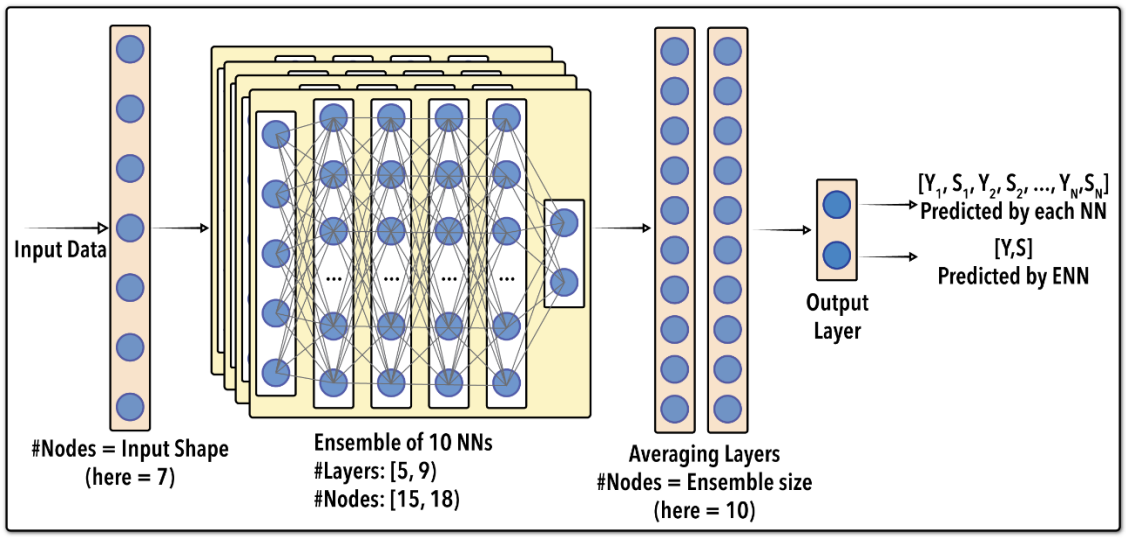


Figure S7. *Illustration of our machine learning model's architecture, offering a visual insight into its components and interactions for enhanced comprehension.*

S7. Supplementary References

1. Ibrahim, M. Y., Bennett, J. A., Mason, D., Rodgers, J. & Abolhasani, M. Flexible homogeneous hydroformylation: on-demand tuning of aldehyde branching with a cyclic fluorophosphite ligand. *Journal of Catalysis* **409**, 105-117 (2022). https://doi.org:https://doi.org/10.1016/j.jcat.2022.03.030

2. Epps, R. W., Volk, A. A., Reyes, K. G. & Abolhasani, M. Accelerated AI development for autonomous materials synthesis in flow. *Chemical science* **12**, 6025-6036 (2021). https://doi.org/10.1039/D0SC06463G

3. Shahhosseini, M., Hu, G. & Pham, H. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications* **7**, 100251 (2022). https://doi.org/10.1016/j.mlwa.2022.100251