**Supplementary Information for:   
Exploring Optimal Reaction Conditions Guided by Graph Neural Networks and Bayesian Optimization**

Youngchun Kwona, b, Dongseon Leeb, Jin Woo Kimb, Youn-Suk Choi\*b, and Sun Kim\*a, c, d, e

*a Department of Computer Science and Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea*

*b Samsung Advanced Institute of Technology, Samsung Electronics Co. Ltd., 130 Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16678, Republic of Korea*

*c Interdisciplinary Program in Bioinformatics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea*

*d Institute of Engineering Research, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Republic of Korea*

*e AIGENDRUG Co., Ltd.*

\* Corresponding author. E-mail: [*ysuk.choi@samsung.com*](mailto:ysuk.choi@samsung.com)*,* [*sunkim.bioinfo@snu.ac.kr*](mailto:sunkim.bioinfo@snu.ac.kr)

# **ABSTRACT**

This supplementary information contains details on: (ⅰ) benchmarked reaction details, including each search condition space for task1 and 2, (ⅱ) additional optimization validation for task1, (ⅲ) optimization details in task2 (ⅳ)distribution of reaction condition data in *Reaxys database* for training graph neural networks, (ⅴ) performance for predictive reaction conditions in graph neural (top-k accuracy), (ⅵ) additional plot for 3a) of task1 (including comparison with baseline and 50 human expertise).

**Contents**

Supplementary Note 1: Benchmarked reaction details including search condition spaceS1

Supplementary Note 2: Additional validation for optimization in task1S2

Supplementary Note 3: Optimization details in task2S3

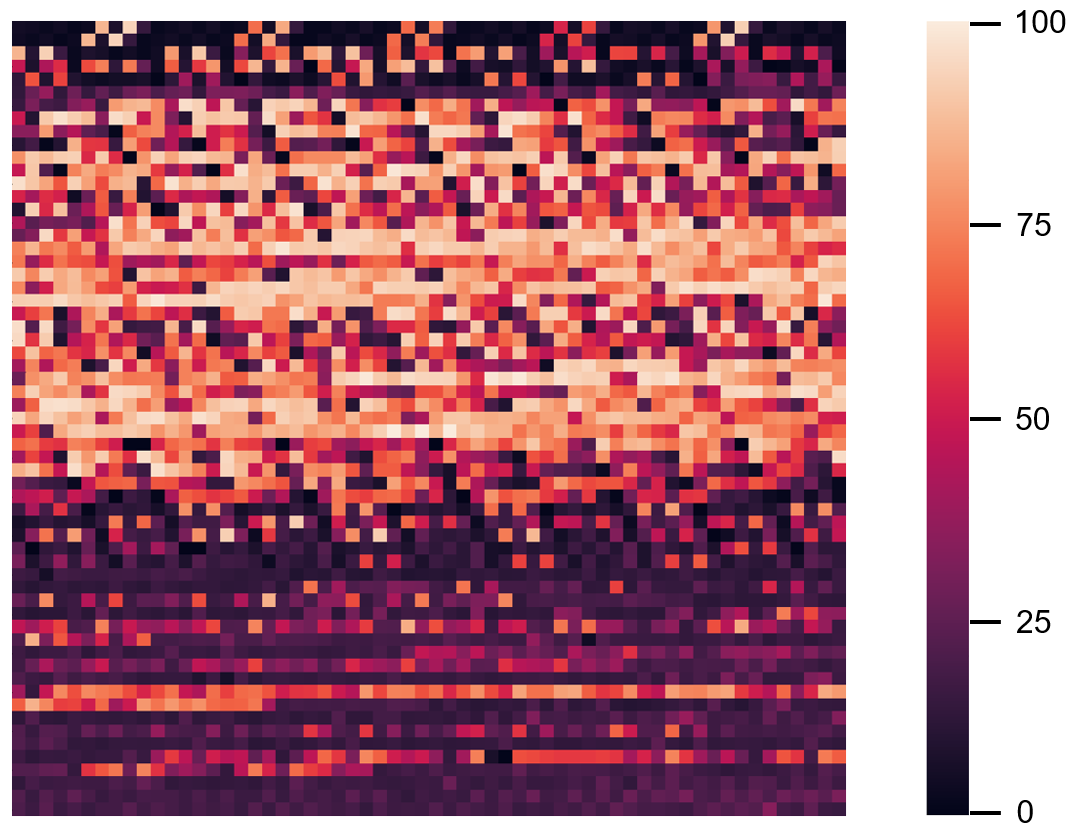
Supplementary Note 4: Distribution for reaction conditions in *Reaxys database*S4

Supplementary Note 5: Top-k accuracy of the MPNNs for predicting conditionsS5

Supplementary Note 6: The HDO’s optimization performance compare to humanS6

# **S. Note 1: Benchmarked reaction details including search condition space**

Reaction condition candidate list for optimal yield are determined in consideration of the subjective judgment and synthesis environment of experts. In particular, the definition of experimental conditions could be a reagent such as a catalyst and a solvent, or an experimental environment such as room temperature and pressure. In this paper, we mainly focuses on reagents. Experiments to find the optimal temperature and equivalent ratio are also included at (3a). In this section, the details of various experimental condition areas are defined.

**Task1 :** In task1, 1a), 2a-e) and 3a), the range set in Shields’ study was benchmarked, and the details are defined as follows. Fully experimental yield results for all combinations of conditions are provided in corresponding <https://github.com/b-shields/edbo>.  
**1a)** Suzuki-Miyaura Reaction  
**2a-e)** Buchwald-Hartwig Reaction

C:\Users\y1000.kwon.CORP\AppData\Local\Microsoft\Windows\INetCache\Content.Word\takeS1.tif

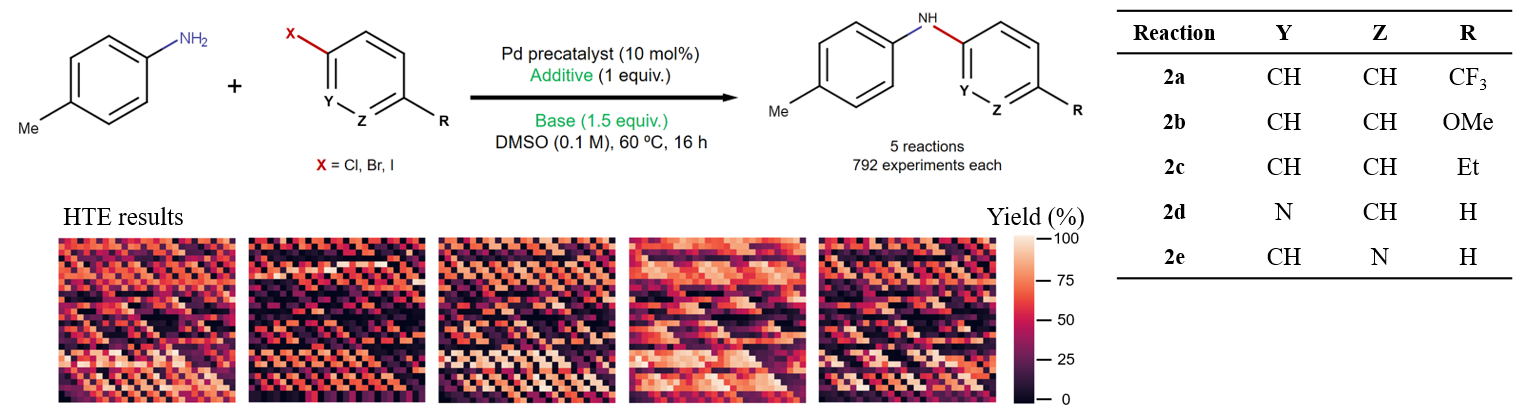
Supplementary Figure 1.

Supplementary Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Count** | **Nucleophiles (reactnant1)** | **Electrophiles (reactnant2)** | **Solvent** | **Base** | **Ligands** |
| **Target** **conditions** | 1 | B(OH)2 | Cl | MeOH/H2O | NaOH | PtBu3 |
| 2 | BPin | Br | THF/H2O | KOH | PPh3 |
| 3 | BF3K | I | MeCN/H2O | NAOH3 | P(o-tol)3 |
| 4 | - | OTf | DMF/H2O | CsF | Butyldi-1-adamantylphosphine |
| 5 | - | - | - | K3PO4 | PCy3 |
| 6 | - | - | - | LiMeO | SPhos |
| 7 | - | - | - | Me3N | XPhos |
| 8 |  |  |  |  | DTBPF |
| 9 |  |  |  |  | DPPF |
| 10 |  |  |  |  | XanPhos |
| 11 |  |  |  |  | Phosphine |

Yield (%)

HTE results



Supplementary Figure 2.

Supplementary Table. 2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Reactant2** | **Ligand** | **Base** | **Additive** |
| **Target** **conditions** | 1 | Cl | XPhos | P2Et | 4-Phenylisoxazole |
| 2 | Br | *t*-BuXPhos | BTMG | 5-phenylisoxazole |
| 3 | I | *t*-BuBrettPhos | MTBD | 3-Phenylisoxazole |
| 4 |  | AdBrettPhos |  | Ethyl 3-Methylisoxazole-5-carboxylate |
| 5 |  |  |  | 3-Methylisoxazole |
| 6 |  |  |  | Ethyl 5-methylisoxazole-3-carboxylate |
| 7 |  |  |  | 5-Phenyl-1,2,4-oxadiazole |
| 8 |  |  |  | 5-Methylisoxazole |
| 9 |  |  |  | Ethyl isoxazole-3-carboxylate |
| 10 |  |  |  | Anthranil |
| 11 |  |  |  | Ethyl 5-Methylisoxazole-4-carboxylate |
| 12 |  |  |  | 3,5-Dimethylisoxazole |
| 13 |  |  |  | Ethyl isoxazole-4-carboxylate |
| 14 |  |  |  | Ethyl isoxazole-5-carboxylate |
| 15 |  |  |  | 1,2-benzisoxazole |
| 16 |  |  |  | 5-(2,6-difluoro-phenyl)-isoxazole |
| 17 |  |  |  | 3-methyl-5-phenylisoxazole |
| 18 |  |  |  | 5-methyl-3-pyrrol-1-yl-1,2-oxazole |
| 19 |  |  |  | 5-furan-2-yl-isoxazole-3-carboxylic acid methyl ester |
| 20 |  |  |  | Methyl 5-phenylisoxazole-3-carboxylate |
| 21 |  |  |  | Ethyl 3-methoxy-1,2-oxazole-5-carboxylate |
| 22 |  |  |  | N,N-dibenzyl-1,2-oxazol-5-amine |

**3a)** Arylation reaction

C:\Users\y1000.kwon.CORP\AppData\Local\Microsoft\Windows\INetCache\Content.Word\take2.tif

Supplementary Figure 3.

Supplementary Table 3.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Count** | **Base** | **Solvent** | **Ligand** | **Concentration(M)** | **Temperature (ºC)** |
| **Target conditions** | 1 | KOAc | BuOAc | PCy3 | 0.057 | 90 |
| 2 | KOPiv | p-Xylene | GloroPhos | 0.100 | 105 |
| 3 | CsOAc | BuCN | CgMe-PPh | 0.153 | 120 |
| 4 | CsOPiv | DMAc | PPh*t*-Bu2 |  |  |
| 5 |  | - | PPhMe2 |  |  |
| 6 |  | - | XPhos |  |  |
| 7 |  | - | BrettPhos |  |  |
| 8 |  |  | *t*-BuPh-CPhos |  |  |
| 9 |  |  | JackiePhos |  |  |
| 10 |  |  | PPh2Me |  |  |
| 11 |  |  | PPh3 |  |  |
| 12 |  |  | P(fur)3 |  |  |

**Task2 :** In task2, a general optimization problem with each Catalyst, Base, Solvent, and Ligand as the main search scope was defined by five organic synthesis experts. In addition, we adapted a wide range of named experimental dataset, from Suzuki-Miyaura reaction which has the most synthetic data based on the Reaxys database, to very small amount of Chan-lam reaction synthetic data. Since there is a limit to conducting an experiment of all combinations of conditions in the task2, it is validated how quickly the yield of the combination of conditions proposed by five experts is achieved with their reference.

**4a-j)** Suzuki-Miyaura Reaction

Supplementary Table 4.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Catalyst** | **Base** | **Solvent** | **Ligand** |
| **Target conditions** | 1 | Pd(PPh3)4 | NaHCO3 | THF/H2O (3:1) | PPh3 |
| 2 | [PdCL(C3H5)]2 | Na2CO3 | Dioxane/H2O (3:1) | *t*-Bu3P |
| 3 | Pd(dppf)Cl2 | K2CO3 | Toluene/Ethanol/H2O (3:1:1) | *t*-Bu3P-HBF4 |
| 4 | Pd(PPh3)2Cl2 | K3PO4 | DMF | PCy3 |
| 5 | Pd(OAc)2 | Cs2CO3 |  | BINAP |
| 6 | Pd(dba)2 | NaOh |  | XantPhos |
| 7 | Pd2(dba)3 | Ba(OH)2 |  | DPEPhos |
| 8 | PdCl2(cod) | KOAc |  | Dppf |
| 9 | PdCl2 |  |  | Dppp |
| 10 |  |  |  | SPhos |
| 11 |  |  |  | XPhos |
| 12 |  |  |  | *t*-BuXPhos |

**5a-h)** Buchwald-Hartwig Reaction

Supplementary Table 5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Catalyst** | **Base** | **Solvent** | **Ligand** |
| **Target conditions** | 1 | Pd(PPh3)4 | NaOtBu | THF | PPh3 |
| 2 | [PdCL(C3H5)]2 | NaOEt | Dioxane | *t*-Bu3P |
| 3 | Pd(dppf)Cl2 | KOtBu | Toluene | *t*-Bu3P-HBF4 |
| 4 | Pd(PPh3)2Cl2 | KOEt | Xylene | PCy3 |
| 5 | Pd(OAc)2 | KOH | DMF | BINAP |
| 6 | Pd(dba)2 | K2CO3 |  | XantPhos |
| 7 | Pd2(dba)3 | K3PO4 |  | DPEPhos |
| 8 | PdCl2(cod) | Cs2CO3 |  | Dppf |
| 9 | PdCl2 | NaOH |  | Dppp |
| 10 |  |  |  | SPhos |
| 11 |  |  |  | XPhos |
| 12 |  |  |  | *t*-BuXPhos |

**6a-b)** Ullmann Reaction

Supplementary Table 6.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Count** | **Catalyst** | **Base** | **Solvent** | **Ligand** |
| **Target conditions** | 1 | Cu(I) | NaOtBu | MeCN | DL-Proline |
| 2 | CuI | NaOEt | Dioxane | Picolinic acid |
| 3 | CuBr | KOtBu | Toluene | DMEDA |
| 4 | CuO | KOEt | DMF | DMCyDA |
| 5 | Cu2O | KOH | DMSO | Phenn |
| 6 | Cu(OAc)2 | K2CO3 |  | diMeO-Phenn |
| 7 | CuSO4 | K3PO4 |  | TetraMe-Phenn |
| 8 |  | Cs2CO3 |  | Dipivloylmethane |
| 9 |  | NaOH |  | Quinolinol |
| 10 |  | NaHCO3 |  |  |
| 11 |  | Na2CO3 |  |  |
| 12 |  | Ba(OH)2 |  |  |
| 13 |  | KOAc |  |  |

**7a, b)** Chan-lam Reaction

Supplementary Table 7.

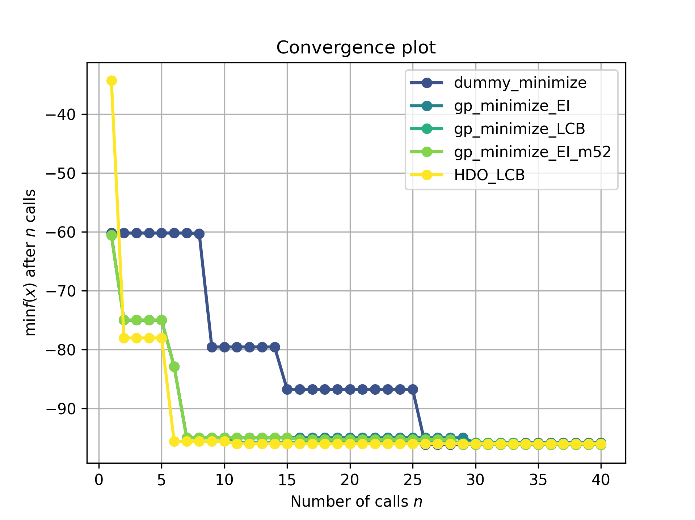
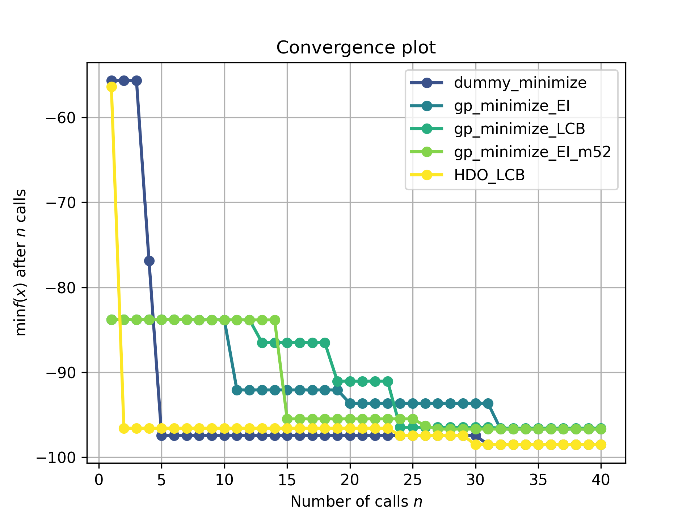
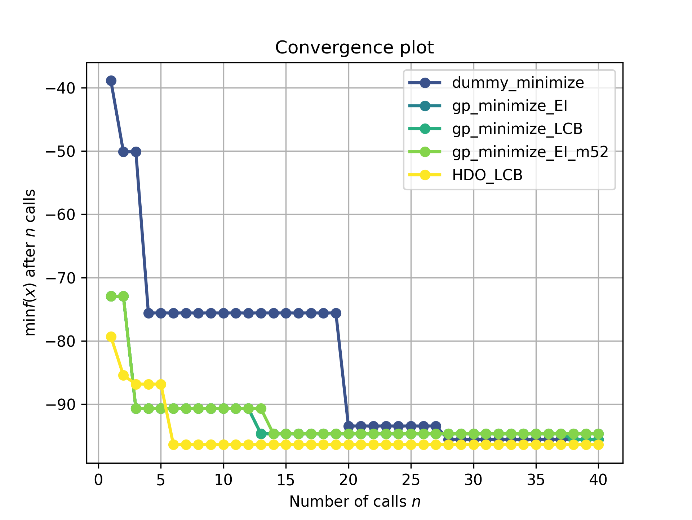
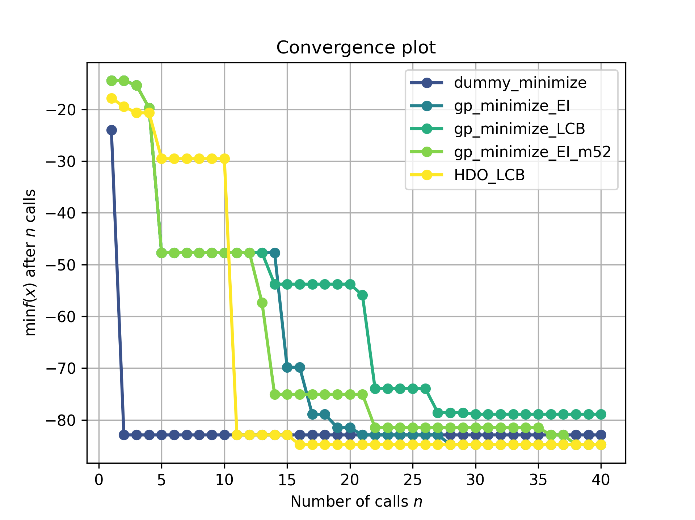
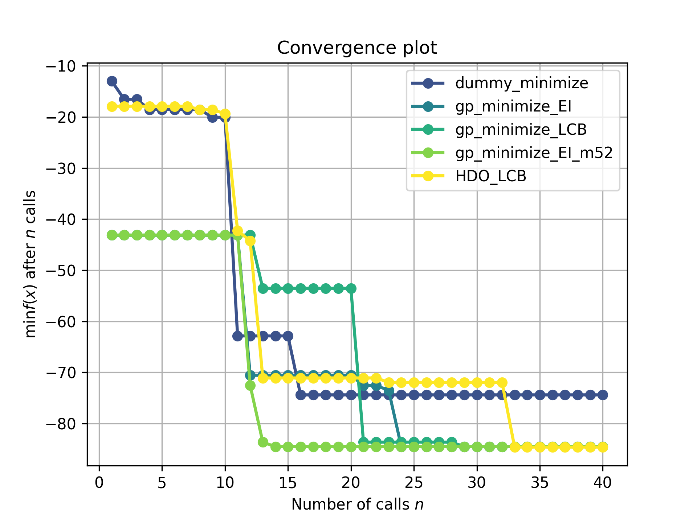
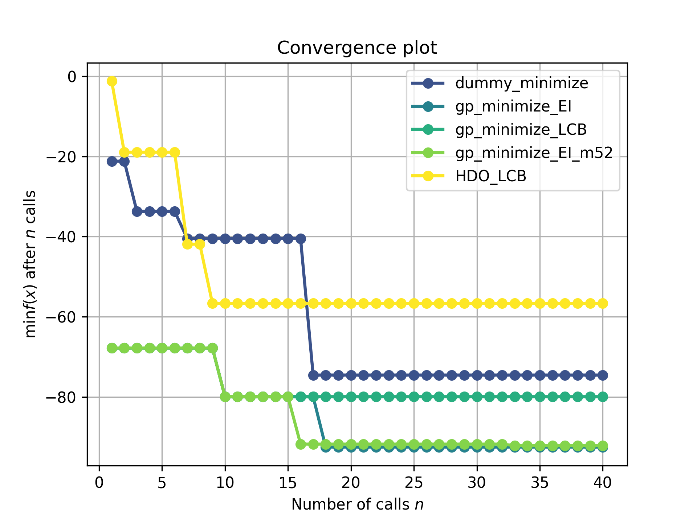
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Catalyst** | **Base** | **Solvent** |
| Target conditions | 1 | Cu(I) | TEA | Methylene chloride |
| 2 | CuI | Pyridine | Methanol |
| 3 | CuBr | DIEA | Ethanol |
| 4 | CuO |  | MeCN |
| 5 | Cu2O |  |  |
| 6 | Cu(OAc)2 |  |  |
| 7 | CuSO4 |  |  |

# **S. Note 2: Additional validation for optimization 1a) and details of 2a-e) reactions**

To validate in the initial selection algorithm of the proposed HDO method, we 1a) reaction experiment as an experiment in which 12 experiments are split into reactants combinations to find the optimal solvent, base and ligand combinations. The redefined areas of the search condition is described in Supplementary Table 8. Through this, the proposed HDO confirmed that the initial experimental condition selection module finds the highest yield on average in nine of the twelve experiments. (Blue area)

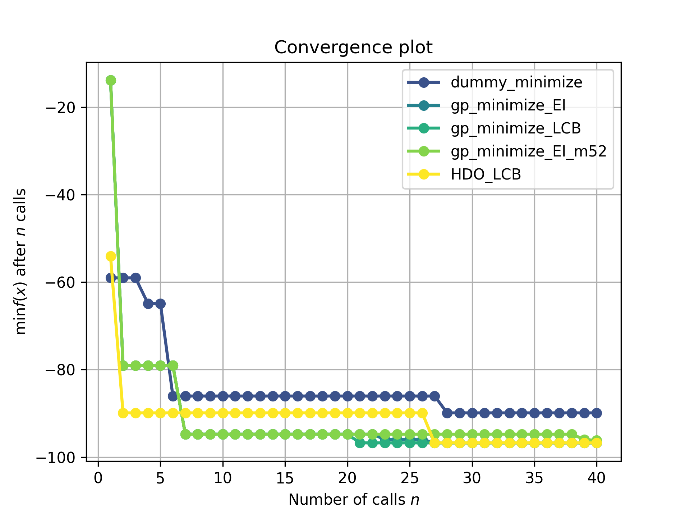
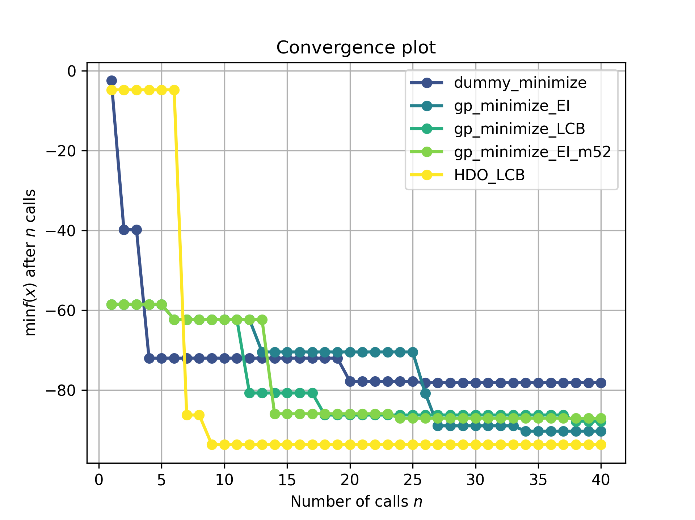
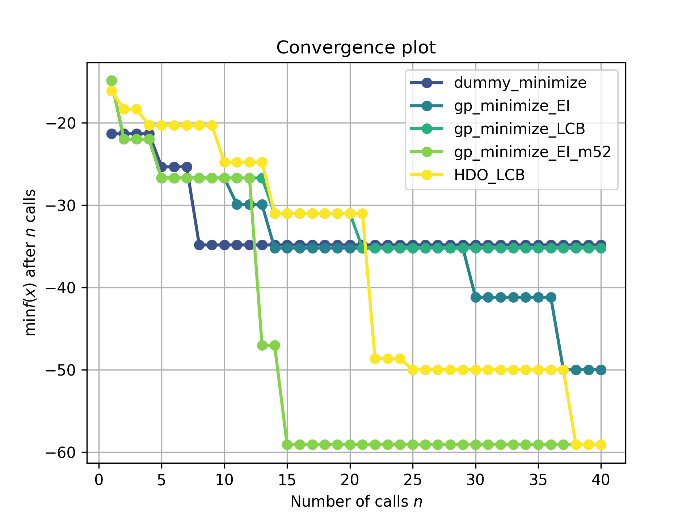
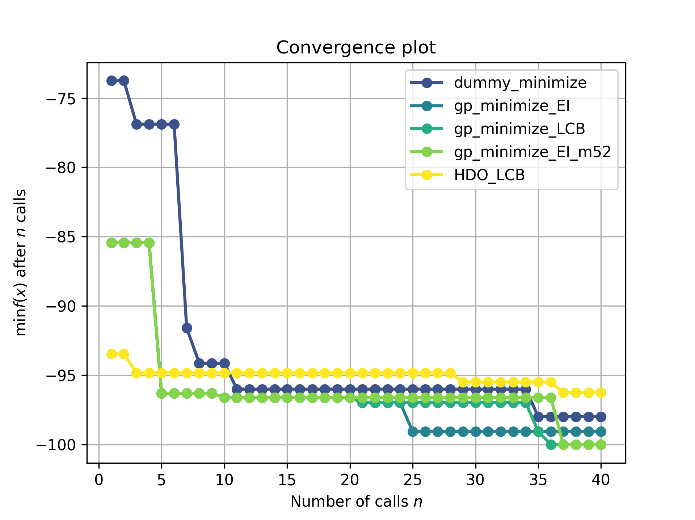
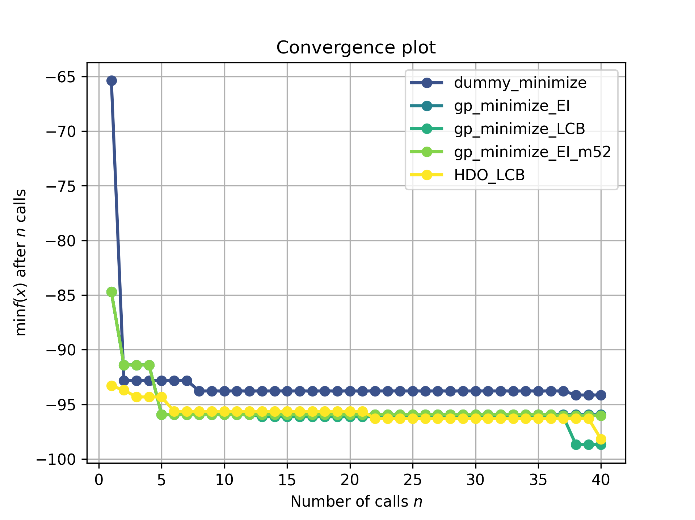
Supplementary Table 8.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Count** | **Solvent** | **Base** | **Ligands** |
| **Target** **conditions** | 1 | MeOH/H2O | NaOH | PtBu3 |
| 2 | THF/H2O | KOH | PPh3 |
| 3 | MeCN/H2O | NAOH3 | P(o-tol)3 |
| 4 | DMF/H2O | CsF | Butyldi-1-adamantylphosphine |
| 5 | - | K3PO4 | PCy3 |
| 6 | - | LiMeO | SPhos |
| 7 | - | Me3N | XPhos |
| 8 |  |  | DTBPF |
| 9 |  |  | DPPF |
| 10 |  |  | XanPhos |
| 11 |  |  | Phosphine |



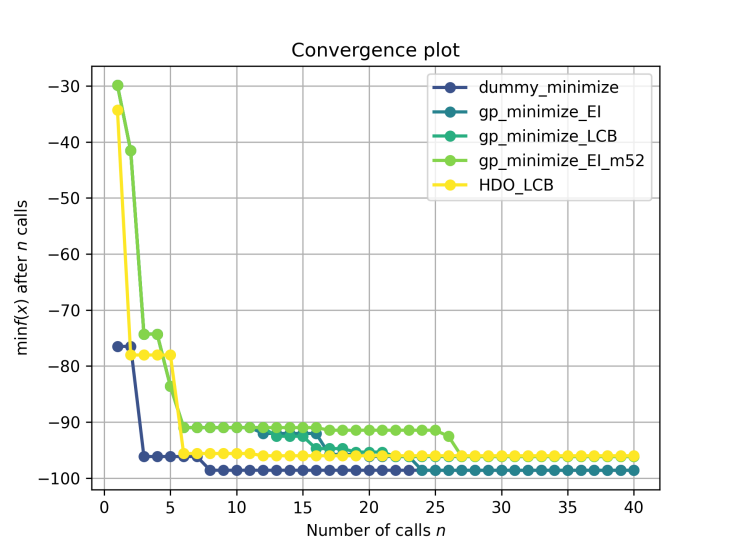
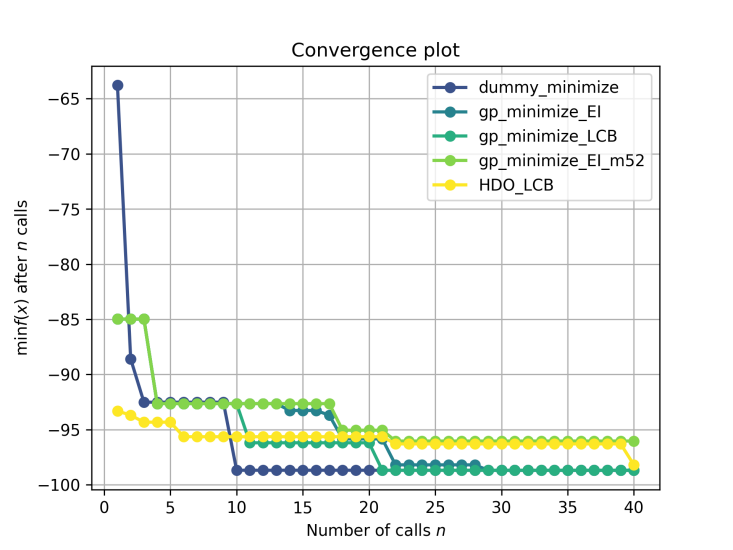
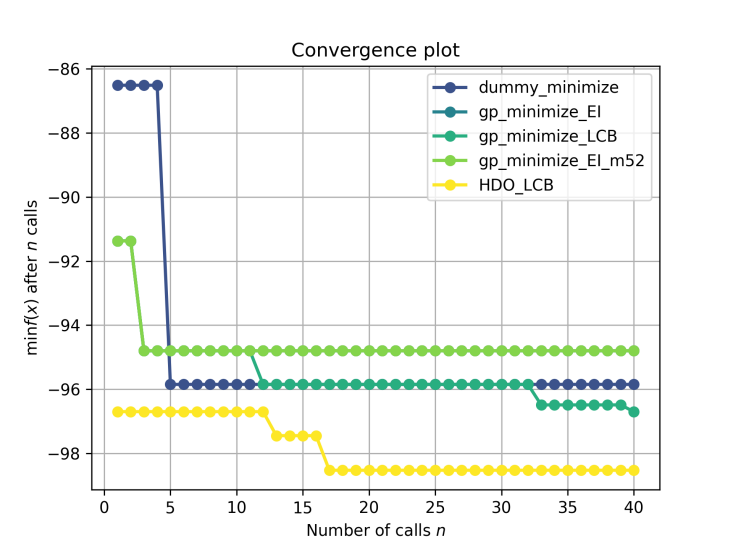
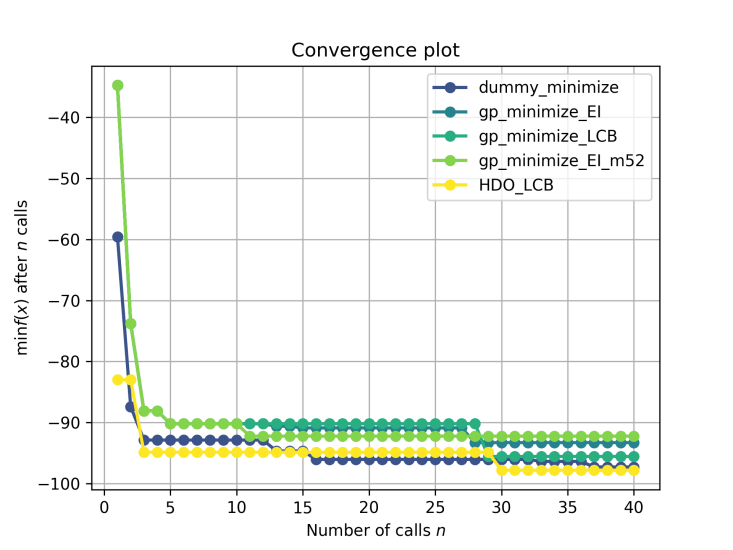
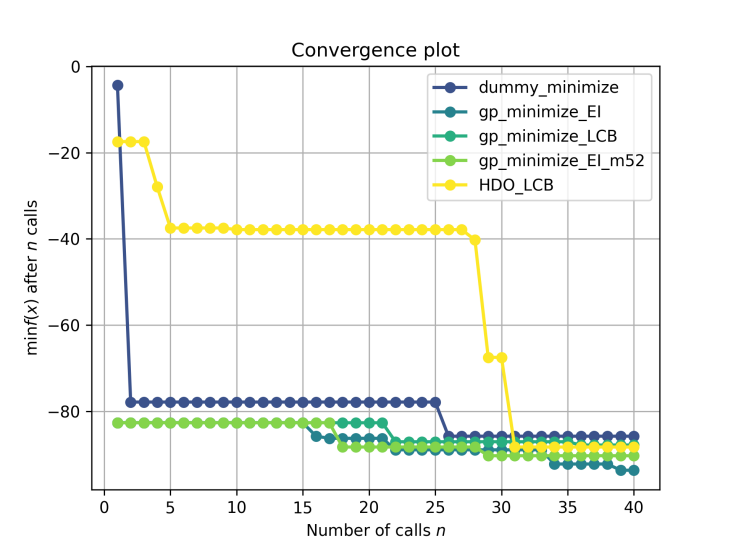
**Fig. 2. MPNN**

Supplementary Figure 4-1. (1a)



Supplementary Figure 4-2. (1a)

Supplementary Figure 4-2. (1a)

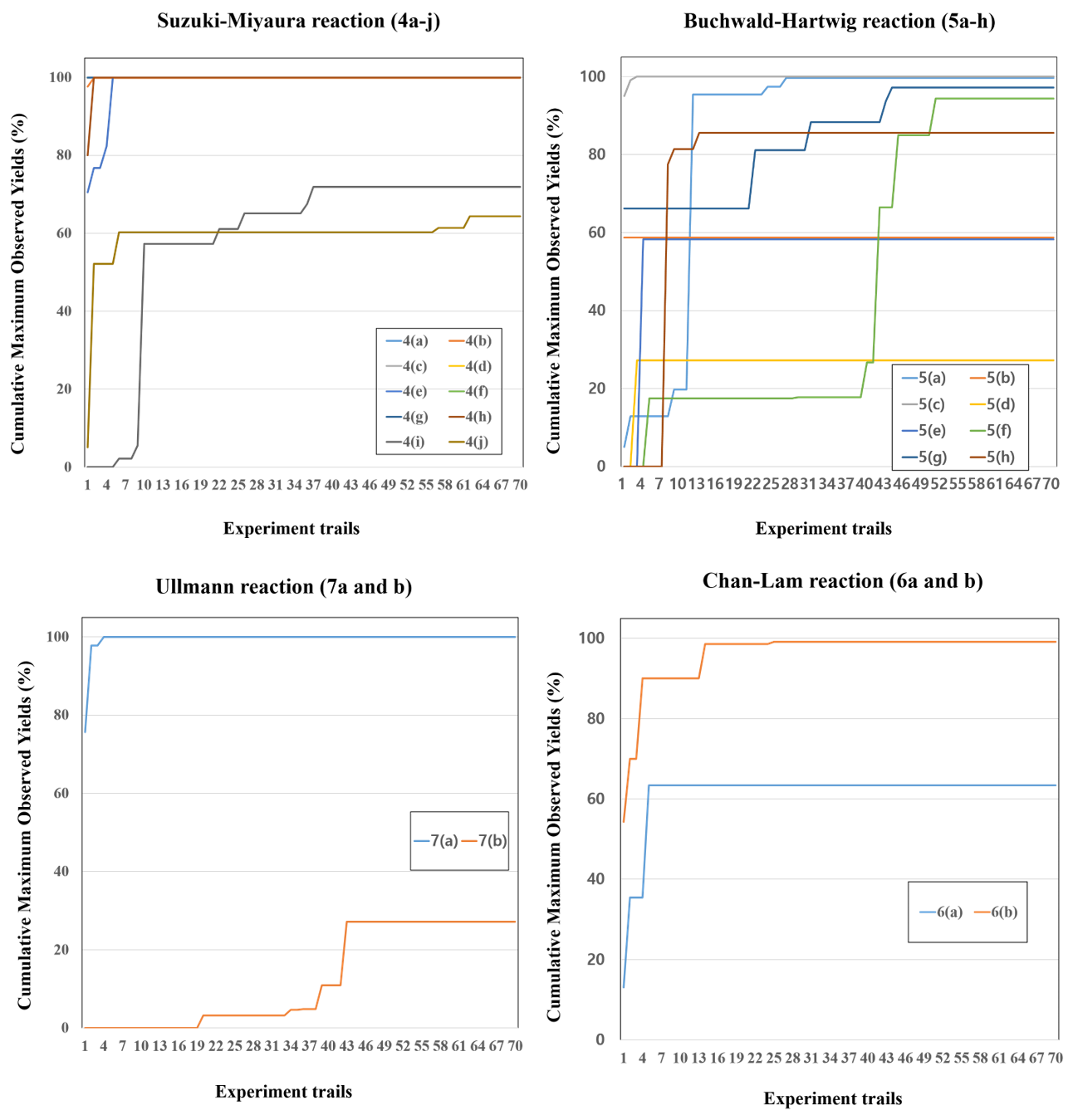


Supplementary Figure 5. (2a-e) : Detailed results for Buchwalt-Hartwig experiments (2a-e)

# **S. Note 3: Optimization results in task 2**

All reaction results for an individual experiment corresponding to the number of trials in Task2.

In case of Suzuki-Miyaura reaction (4a-j), based on vast experimental data for Graph Neural Networks, the HDO found an optimal combination of conditions that conversion yield close to 100% in most 6 trails (initial experiments) in 8 of 10 experiments. In Ullmann and Chan-lam reaction, which has relatively small experimental data, it could be seen that the optimal conditions are not initially found, but the optimal yield is found while repeating. On average, HDO finds conditions within the 4.7 trials that satisfy the yield of the combination of conditions proposed by five organic synthesis experts.



Supplementary Figure 6. (4a-j, 5a-h, 6a-b and 7a-b)

# **S. Note 4: Distribution of reaction conditions in *Reaxys database***

The *Reaxys database* has unbalanced data because it was extracted the experimental information described in the synthetic research papers. Supplementary Figure 7 below shows the number of biased data. Since the classification of reactions is ambiguous or overlapping, and the missing information is included, all reaction condition data are retrieved for Graph Neural Networks training for prediction suitable conditions given reactants with product structures. Plus, a newly redefined a dataset based on notation and CAS-number is used as training.

Supplementary Figure 7

# **S. Note 5: Top-k accuracy of the MPNNs for predicting conditions**

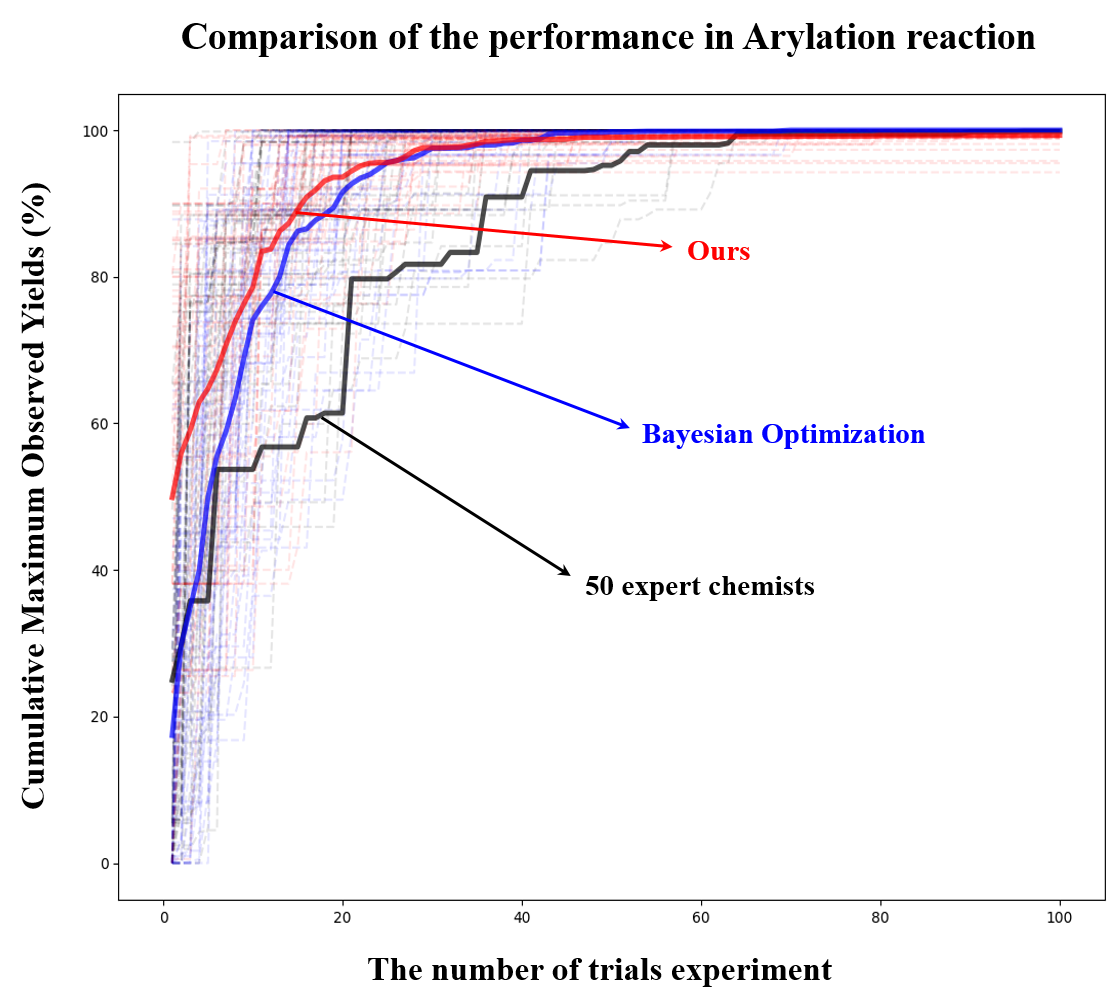
This supplementary for a given search space, we train the MPNN (Message Passing Neural Networks) based on graph-type representation models to predicting the best conditions. The supplementary table 9 below shows the prediction top-k accuracy of the MPNN model according to the experimental type and conditions with DNN (Deep Neural Networks) based ECFP (Extended Connectivity FingerPrint) representation. The table represents the individual prediction performance of the condition prediction model corresponding to task2. Trained on ~ 10 milion examples from Reaxys, the MPNN model is able to propose conditions where a close match to recorded catalyst, solvent, and reagent is found within the top-3 predictions 99.5%.

Supplementary Table 9

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Top-K Accuracy (%)** | | | | | | | | | | | | | | | | | | | | | | | |
|  | **Deep Neural Networks (ECFP)** | | | | | | | | | | | | **Message Passing Neural Networks (Graph)** | | | | | | | | | | | |
|  | **Suzuki-Miyaura** | | | **Buchwald-Hartwig** | | | **Chan-lam** | | | **Ullmann** | | | **Suzuki-Miyaura** | | | **Buchwald-Hartwig** | | | **Chan-lam** | | | **Ullmann** | | |
| **Top-K** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** | **1** | **3** | **5** |
| **Catalyst** | 75.15 | 87.58 | 89.23 | 77.34 | 84.35 | 92.78 | 85.42 | 92.33 | 95.45 | 88.22 | 92.33 | 96.26 | 83.22 | 94.11 | 99.50 | 83.22 | 94.11 | 99.50 | 90.22 | 95.71 | 98.18 | 90.22 | 95.71 | 98.88 |
| **Base** | 77.15 | 88.24 | 88.47 | 78.32 | 84.49 | 89.55 | 92.32 | - | - | 65.44 | 74.25 | 84.35 | 84.71 | 91.54 | 98.47 | 81.12 | 92.38 | 97.87 | 97.72 | - | - | 70.4 | 81.65 | 87.15 |
| **Solvent** | 78.35 | 94.11 | - | 77.34 | 88.11 | 100.0 | 78.44 | 89.88 | - | 85.12 | 92.32 | 100.0 | 88.30 | 98.91 | 100.0 | 87.34 | 97.11 | 100.0 | 81.31 | 98.19 | - | 88.85 | 97.11 | - |
| **Ligand** | 68.87 | 79.22 | 82.34 | 67.12 | 89.88 | 88.84 | - | - | - | 62.32 | 74.26 | 89.32 | 77.1 | 87.32 | 92.84 | 77.1 | 87.32 | 92.84 | - | - | - | 70.76 | 79.54 | 95.15 |
| **Top-3’s Average** | **87.28** | | | **86.70** | | | **94.07** | | | **83.29** | | | **92.97** | | | **92.73** | | | **97.96** | | | **88.50** | | |

# **S. Note 6: The HDO’s optimization performance compare to human**

Finally, the proposed HDO was designed to solve the cold-start problem of Bayesian optimization (BO) in application for finding optimal conditions of synthetic reactions. Supplementary Figure 8 below shows that it trains vast amounts of reaction data compared to BO and Humans, so it is possible to find a combination of high-yield conditions early on. Repeat the experiment, the HDO modify the object function for direction dynamically based on the results of on-going experiments. This is caused to efficiently exploring for optimal conditions.



Supplementary Figure 8.