Supporting information for Democratizing Chemical Machine Learning with Community-Engaged Test Sets

Jason L. Wu,[ab] David M. Friday,[ab] Changhyun Hwang,[bc] Seungjoo Yi,[bd] Tiara C. Torres-Flores,[bc] Martin D. Burke,[abefg] Ying Diao,[abc] Charles M. Schroeder,[abcd] and Nicholas E. Jackson[ab]

a. Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.
b. Molecule Maker Lab, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
c. Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.
d. Department of Materials Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA.
e. Carle R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
f. Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA.
g. Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL, USA.

E-mail: jacksonn@illinois.edu

Contents

Kaggle Playbook

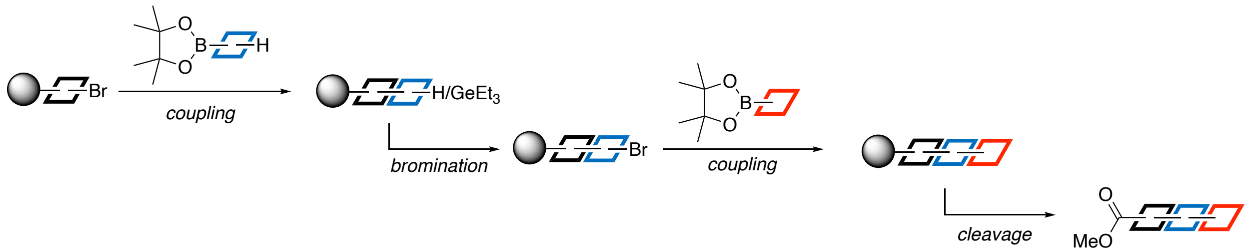
How to use Kaggle: https://www.kaggle.com/docs/competitions-setup

Key points:

- train.csv: features, target value
- test.csv: features, no target value
- solution.csv: target value of test.csv
- Choose an appropriate error metric
    - Mean Absolute Error (MAE): straightforward “average error” in the same units as the target
    - Mean Squared Error (MSE): penalizes large deviations
    - Root Mean Squared Error (RMSE): penalizes large errors but is interpretable in the target’s units
    - Mean Squared Logarithmic Error (MSLE): good for data that spans several orders of magnitude

Synthesis

General Trimer Procedure 1 (GP1)

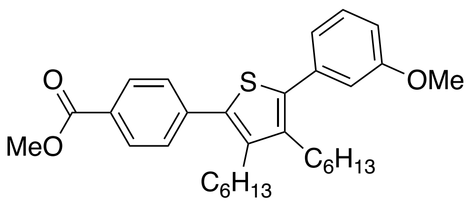


C–C coupling: The appropriate polystyrene-linked aryl bromide (1.0 equiv., 0.15 mmol), TMSOK (1.5 equiv.), and P(t-Bu)$_3$ Pd G3 (10 mol%) were added into a 2.5-mL polypropylene reaction vessel. The aryl pinacol boronic ester (ArBpin, 2.0 equiv.) was added into an Eppendorf tube and dissolved in 1.5 mL of THF. The ArBpin solution was pulled into the reaction vessel, which was capped and shaken on an orbital shaker at 300 rpm for 30 minutes. At the conclusion of the reaction, the resin was washed with DMF, DMF:H_2O 1:1, DMF, MeOH, THF, DCM.

Bromination: N-Bromosuccinimide (NBS, 2.0 equiv.) was added into an Eppendorf tube and dissolved in 1.5 mL of DCM. The NBS solution was pulled into the reaction vessel, which was capped and shaken on an orbital shaker at 300 rpm for 1.5 hours. At the conclusion of the reaction, the resin was washed with DMF, DMF:H_2O 1:1, DMF, MeOH, THF.

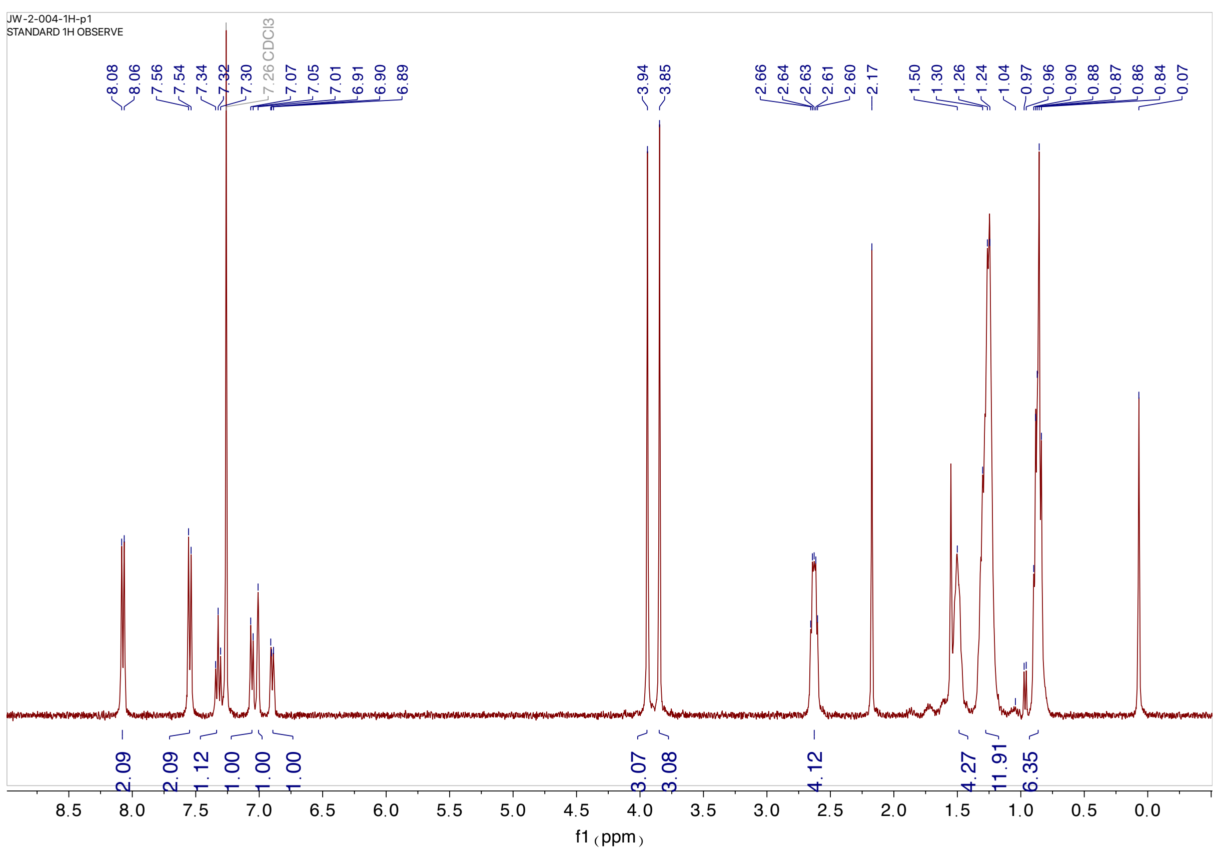
Cleavage: The resin was transferred into a glass solid-phase synthesis vessel. NaOMe (0.1 equiv.) was added into the vessel, and the contents were dissolved in 5 mL Toluene:MeOH 4:1. The vessel was shaken at 300 rpm, 75 °C overnight. At the conclusion of the reaction, the solution was collected, and solvent was removed under vacuum.

[O1] methyl 4-(3,4-dihexyl-5-(3-methoxyphenyl)thiophen-2-yl)benzoate

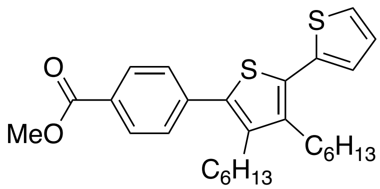


Following GP1 using 2-(3,4-dihexylthiophen-2-yl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 1st coupling and 2-(3-methoxyphenyl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 2nd coupling. The title product was obtained after purification by column chromatography (hexanes/EtOAc 5:1) to give O1 (product was isolated in trace amounts). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.07 (d, *J* = 8.0 Hz, 2H), 7.55 (d, *J* = 8.0 Hz, 2H), 7.32 (t, *J* = 7.9 Hz, 1H), 7.06 (d, *J* = 7.6 Hz, 1H), 7.01 (s, 1H), 6.93 – 6.85 (m, 1H), 3.94 (s, 3H), 3.85 (s, 3H), 2.63 (dt, *J* = 10.8, 5.4 Hz, 4H), 1.50 (s, 4H), 1.27 (t, *J* = 11.5 Hz, 12H), 0.86 (d, *J* = 6.2 Hz, 6H). HRMS (ESI+) calculated for $C_{31}H_{40}O_3S$ $[M]^+$ m/z 492.2698, found 492.2722.

[O1] methyl 4-(3,4-dihexyl-5-(3-methoxyphenyl)thiophen-2-yl)benzoate. [^{1}H 400 MHz; $CDCl_3$]

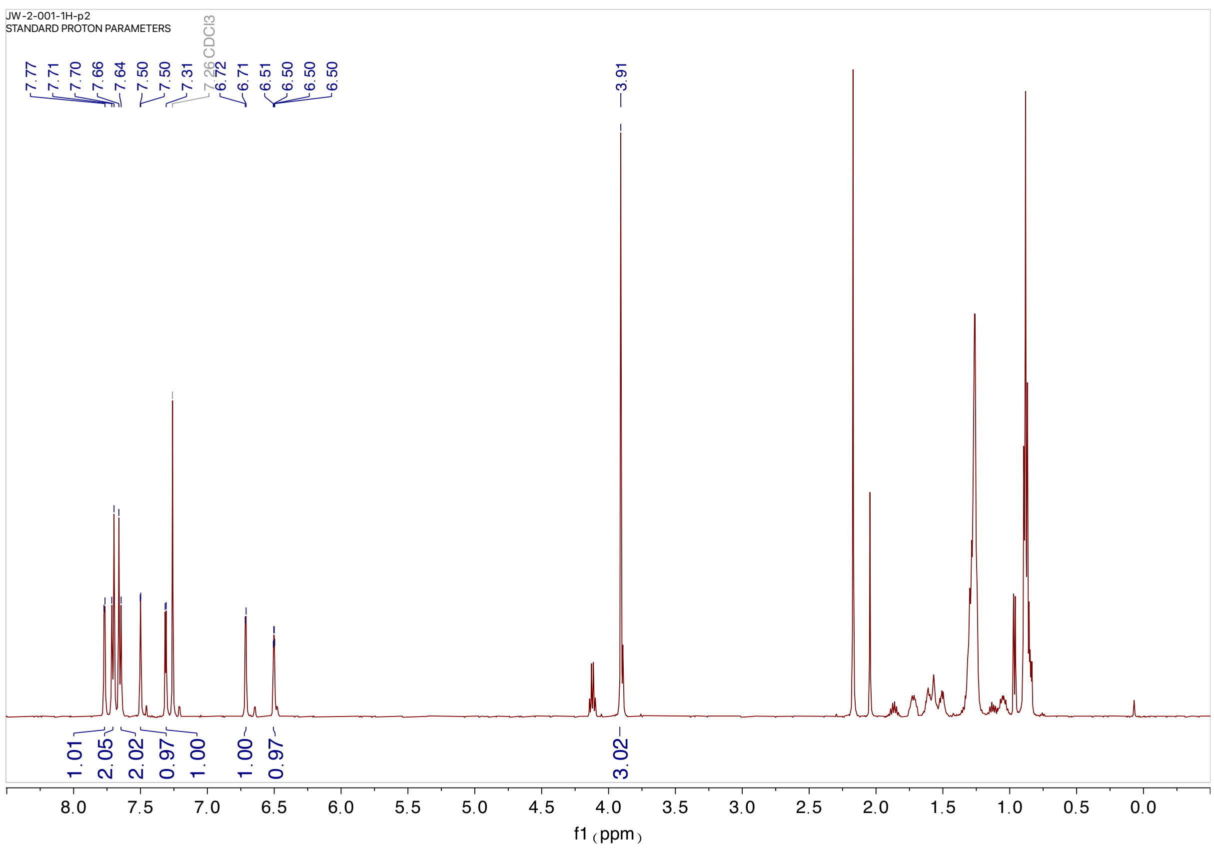


[O2] methyl 4-(3,4-dihexyl-[2,2'-bithiophen]-5-yl)benzoate

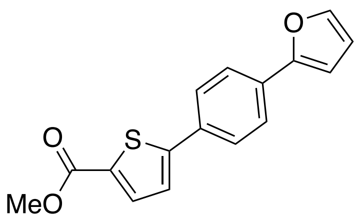


Following GP1 using 2-(3,4-dihexylthiophen-2-yl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 1st coupling and 4,4,5,5-tetramethyl-2-(thiophen-2-yl)-1,3,2-dioxaborolane for the 2nd coupling. The title product was obtained after purification by column chromatography (hexanes/EtOAc 5:1) to give O2 (product was isolated in trace amounts). ^{1}H NMR (400 MHz, $CDCl_3$) δ 8.07 (d, *J* = 7.8 Hz, 2H), 7.53 (d, *J* = 7.7 Hz, 2H), 7.32 (d, *J* = 4.9 Hz, 1H), 7.15 (d, *J* = 2.5 Hz, 1H), 7.07 (t, *J* = 4.4 Hz, 1H), 3.94 (s, 3H), 2.76 – 2.66 (m, 2H), 2.61 (t, *J* = 8.3 Hz, 2H), 1.48 (d, *J* = 7.7 Hz, 2H), 1.41 (s, 2H), 1.35 – 1.11 (m, 12H), 0.87 (dt, *J* = 14.0, 6.6 Hz, 6H). HRMS (ESI+) calculated for $C_{28}H_{36}O_2S_2$ $[M]^+$ m/z 468.2157, found 468.2149.

[O2] methyl 4-(3,4-dihexyl-[2,2'-bithiophen]-5-yl)benzoate. [^{1}H 400 MHz; $CDCl_3$]

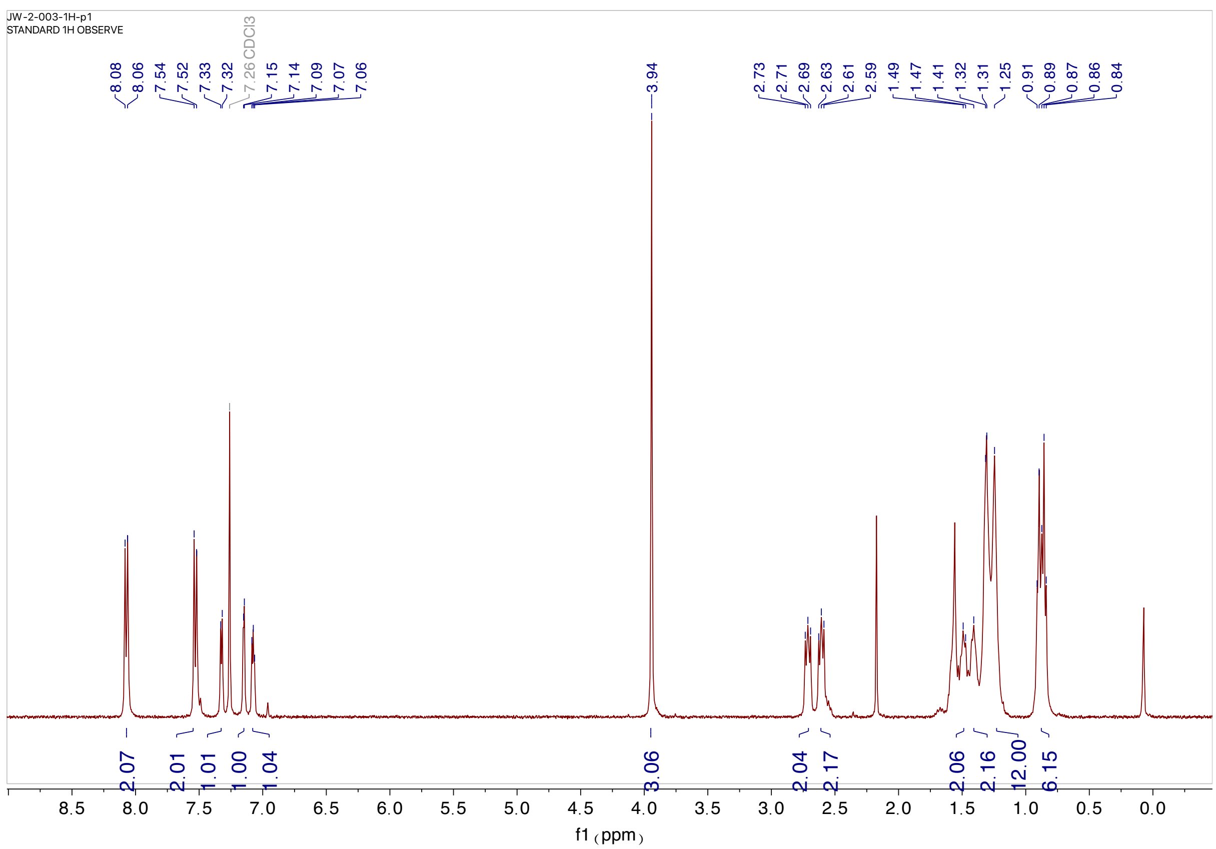


[O3] methyl 5-(4-(furan-2-yl)phenyl)thiophene-2-carboxylate

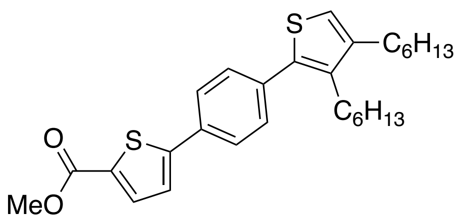


Following GP1 using triethyl(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenyl)germane for the 1st coupling and 2-(furan-2-yl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 2nd coupling. The title product was obtained after purification by column chromatography (hexanes/EtOAc 5:1) to give O3 (product was isolated in trace amounts).^{1}H NMR (500 MHz, $CDCl_3$) δ 7.77 (d, *J* = 3.9 Hz, 1H), 7.71 (d, *J* = 8.5 Hz, 2H), 7.65 (d, *J* = 8.5 Hz, 2H), 7.50 (d, *J* = 1.8 Hz, 1H), 7.31 (d, *J* = 3.9 Hz, 1H), 6.71 (d, *J* = 3.4 Hz, 1H), 6.50 (dd, *J* = 3.4, 1.8 Hz, 1H), 3.91 (s, 3H). HRMS (ESI+) calculated for $C_{16}H_{12}O_3S$ $[M+H]^+$ m/z 285.0507, found 285.0583.

[O3] methyl 5-(4-(furan-2-yl)phenyl)thiophene-2-carboxylate. [^{1}H 500 MHz; $CDCl_3$]

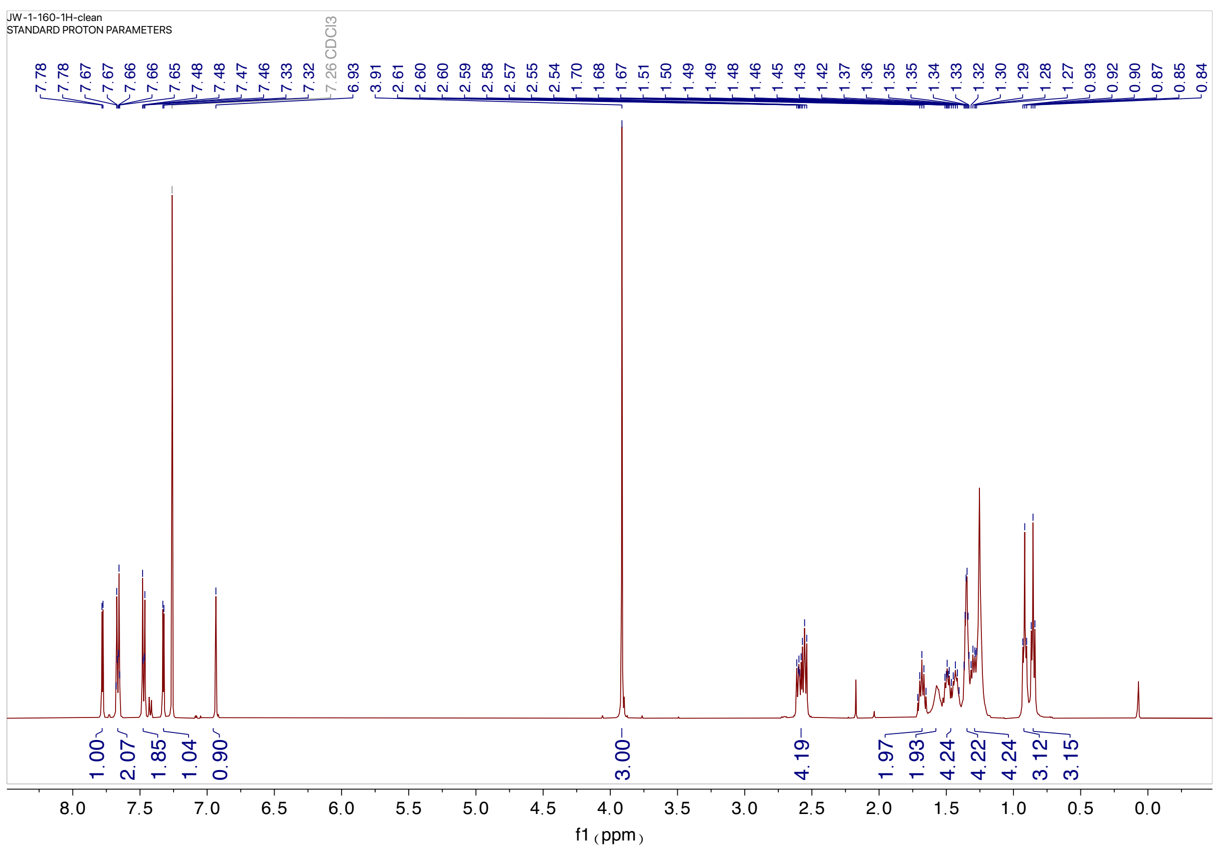


[O4] methyl 5-(4-(3,4-dihexylthiophen-2-yl)phenyl)thiophene-2-carboxylate

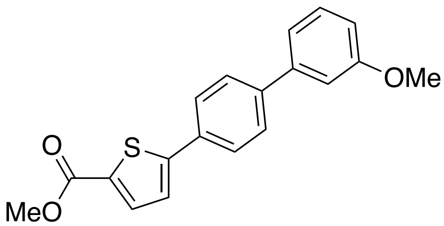


Following GP1 using triethyl(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenyl)germane for the 1st coupling and 2-(3,4-dihexylthiophen-2-yl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 2nd coupling. The title product was obtained after purification by column chromatography (hexanes/EtOAc 5:1) to give O4 (product was isolated in trace amounts). ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.78 (d, *J* = 3.9 Hz, 1H), 7.69 – 7.64 (m, 2H), 7.50 – 7.45 (m, 2H), 7.33 (d, *J* = 3.9 Hz, 1H), 6.93 (s, 1H), 3.91 (s, 3H), 2.64 – 2.53 (m, 4H), 1.68 (p, *J* = 7.5 Hz, 2H), 1.52 – 1.41 (m, 4H), 1.35 (h, *J* = 3.7 Hz, 4H), 1.29 (q, *J* = 6.8 Hz, 4H), 0.95 – 0.90 (m, 3H), 0.87 – 0.83 (m, 3H). HRMS (ESI+) calculated for $C_{28}H_{36}O_2S_2$ $[M]^+$ m/z 468.2157, found 468.2166.

[O4] methyl 5-(4-(3,4-dihexylthiophen-2-yl)phenyl)thiophene-2-carboxylate. [^{1}H 500 MHz; $CDCl_3$]

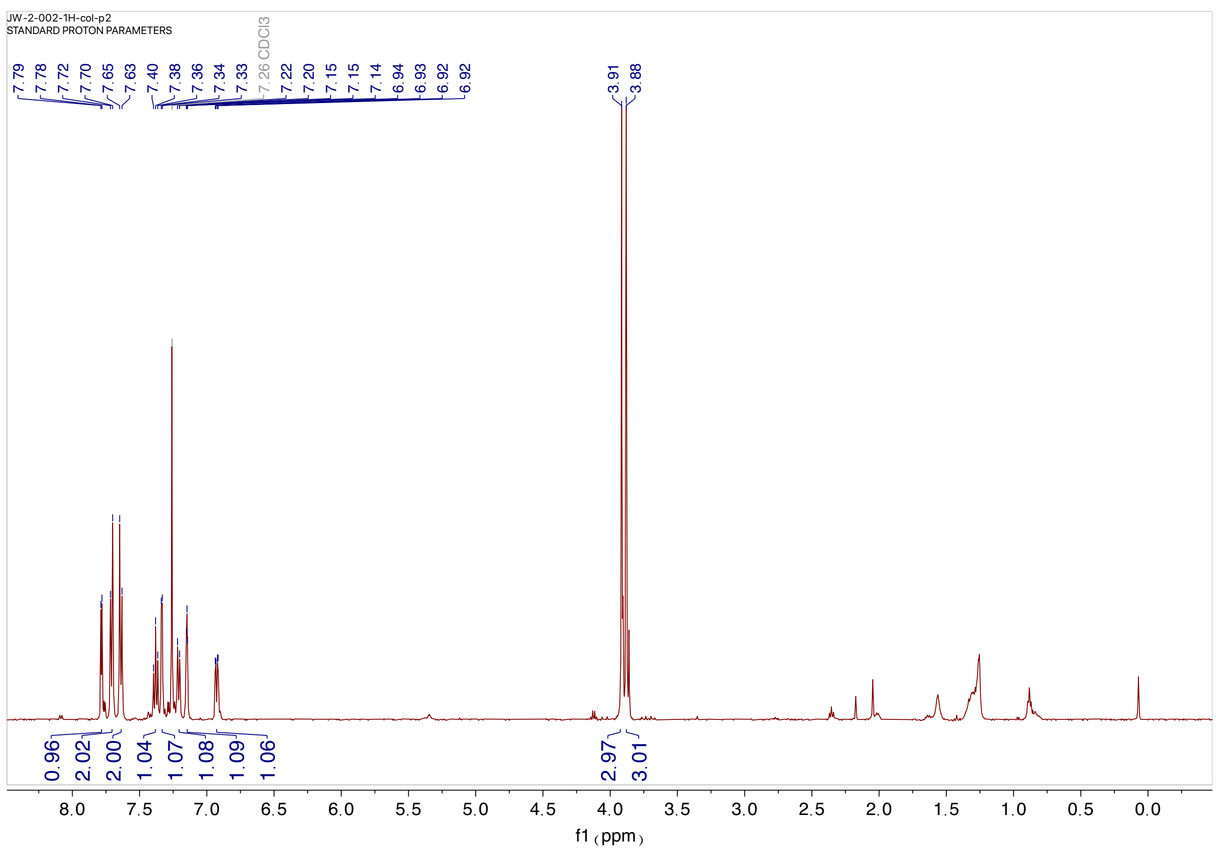


[O5] methyl 5-(3'-methoxy-[1,1'-biphenyl]-4-yl)thiophene-2-carboxylate



Following GP1 using triethyl(4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)phenyl)germane for the 1st coupling and 2-(3-methoxyphenyl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane for the 2nd coupling. The title product was obtained after purification by column chromatography (hexanes/EtOAc 5:1) to give O5 (product was isolated in trace amounts). ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.78 (d, *J* = 3.9 Hz, 1H), 7.71 (d, *J* = 8.3 Hz, 2H), 7.64 (d, *J* = 8.4 Hz, 2H), 7.38 (t, *J* = 7.9 Hz, 1H), 7.33 (d, *J* = 4.0 Hz, 1H), 7.21 (d, *J* = 7.7 Hz, 1H), 7.16 – 7.13 (m, 1H), 6.93 (dd, *J* = 7.8, 2.3 Hz, 1H), 3.91 (s, 3H), 3.88 (s, 3H). HRMS (ESI+) calculated for $C_{19}H_{16}O_3S$ $[M]^+$ m/z 324.0820, found 324.0812.

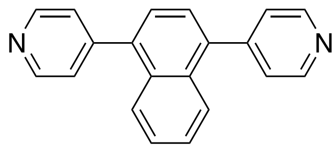
[O5] methyl 5-(3'-methoxy-[1,1'-biphenyl]-4-yl)thiophene-2-carboxylate. [^{1}H 500 MHz; $CDCl_3$]



General Trimer Procedure 2 (GP2)

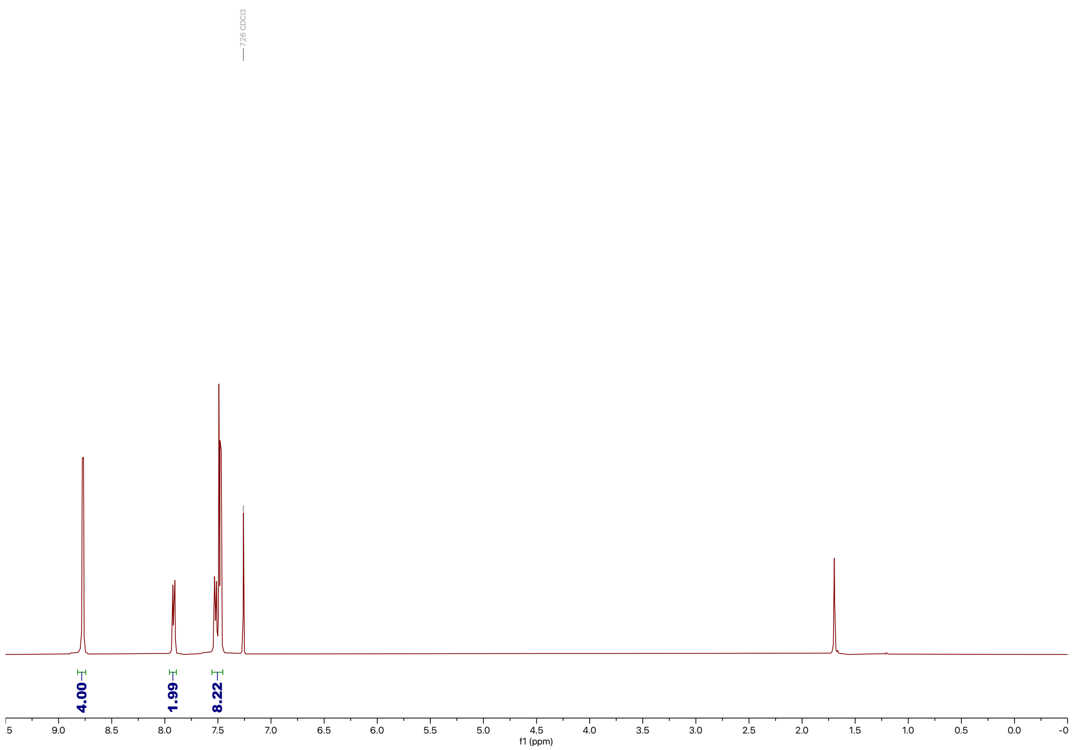
A 40 mL I-Chem vial equipped with a stir bar was charged with aryl dibromide (1 equiv), 4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)pyridine (2.5 equiv), Trimethyl borate (5 equiv), TMSOK (3 equiv), and CataCXium Pd G4 (10 mol%). The vials were loaded onto the heating block inside the hood and the synthesis procedure was executed. A Schlenk-line procedure with 10 cycles, solvent addition (0.2 M, anhydrous toluene), heating (90 °C) and stirring (220 rpm) for the 2 h of reaction time, and then stopping the reaction by cooling the hotplate (20 °C) and ceasing stirring (0 rpm) were performed sequentially. Following reaction completion, the mixture was diluted by the addition of dichloromethane (DCM, 20× volume relative to reaction volume). The resulting solution was subjected to aqueous work-up, involving three successive extractions with water. The organic layer was collected and subsequently acidified by the slow addition of 1 M hydrochloric acid (HCl) until no further precipitation was observed. The resulting precipitate was isolated by vacuum filtration. The solid was then redissolved in water, and the solution was basified by the gradual addition of 1M aqueous sodium hydroxide (NaOH) until precipitation ceased. The final precipitate was collected via filtration and thoroughly rinsed with excess deionized water to remove residual salts and impurities, yielding the purified product as a powder.

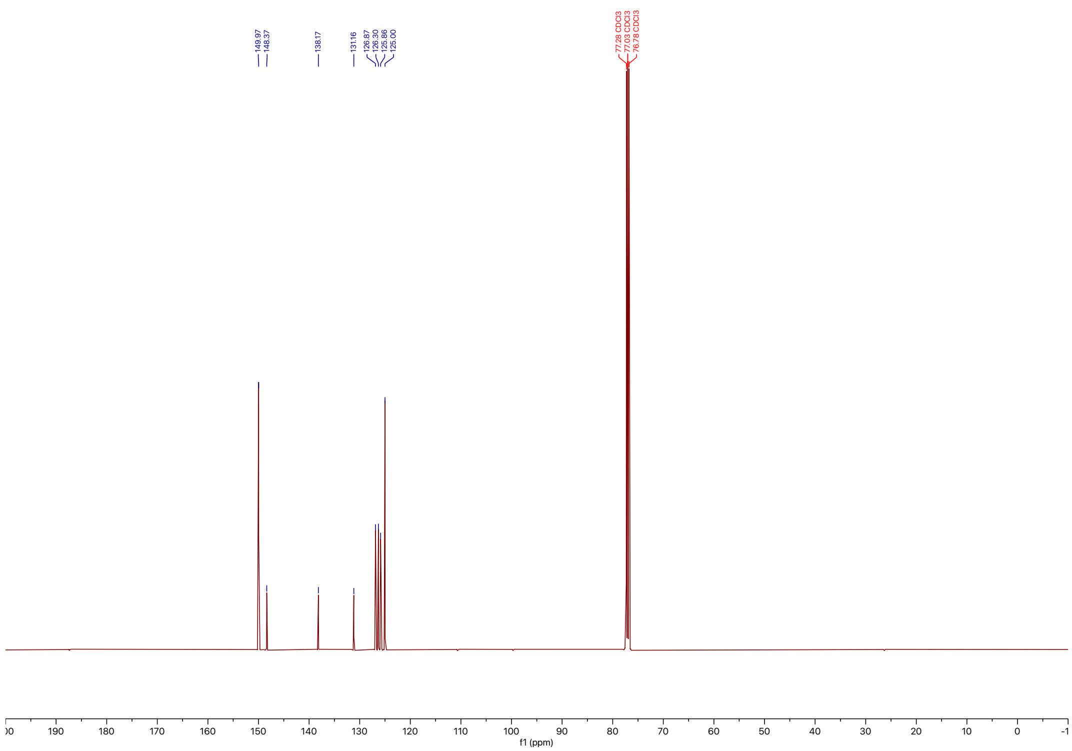
[O6] 1,4-di(pyridin-4-yl)naphthalene



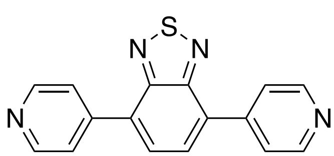
Following slightly modified GP2 using 1,4-dibromonaphthalene and using DME as a solvent gave O6 (222 mg, 39.3% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.80 – 8.75 (m, 4H), 7.96 – 7.88 (m, 2H), 7.53 (dt, *J* = 6.5, 3.1 Hz, 2H), 7.50 – 7.46 (m, 6H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 150.30, 150.10, 149.86, 148.50, 138.30, 131.29, 127.00, 126.43, 125.99, 125.33, 125.13, 124.89. HRMS (ESI+) calculated for C20H15N2 $[M+H]^+$ m/z 283.1235, found 283.1241.

[O6] 1,4-di(pyridin-4-yl)naphthalene [^{1}H 500 MHz; ^{13}C 126 MHz; $CDCl_3$]



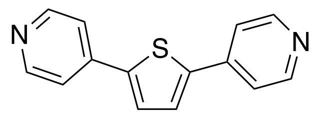


[O7] 2,5-di(pyridin-4-yl)2,1,3-Benzothiadiazole



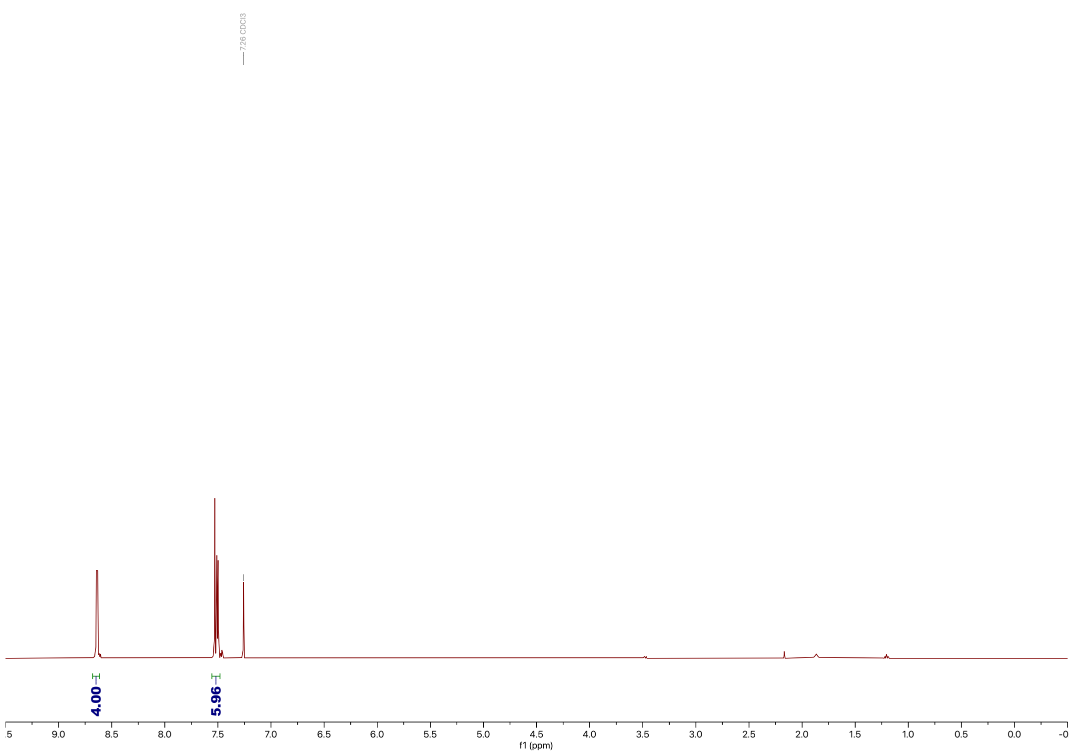
O7 was purchased through Ambeed Cat. No A2022249

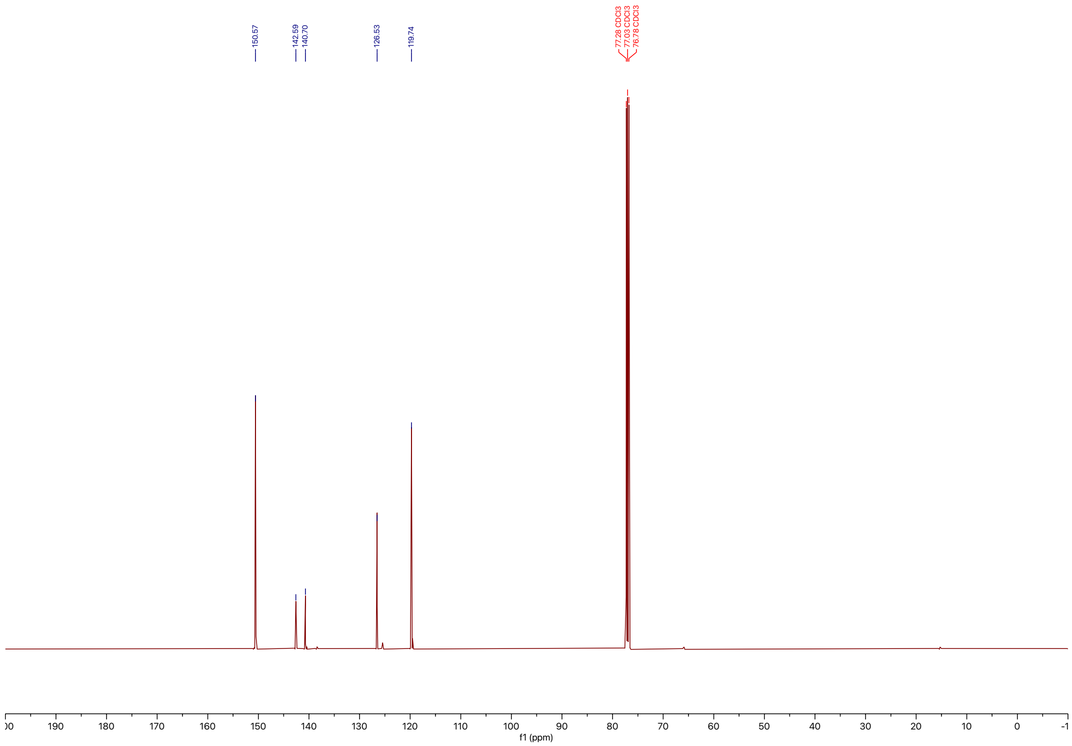
[O8] 2,5-di(pyridin-4-yl)thiophene



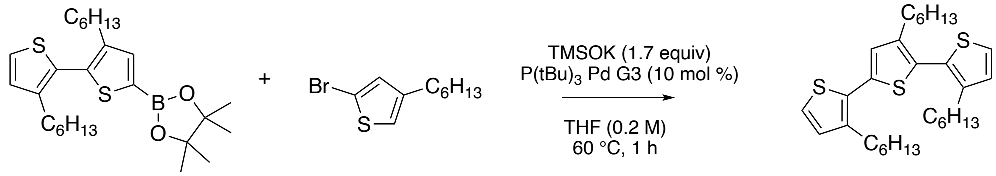
Following GP2 using 2,5-dibromothiophene gave O8 (142 mg, 29.7% yield). ^{1}H NMR (500 MHz, $CDCl_3$) δ 8.69 – 8.64 (m, 4H), 7.56 (d, *J* = 2.6 Hz, 2H), 7.55 – 7.51 (m, 4H). ^{13}C NMR (126 MHz, $CDCl_3$) δ 150.70, 142.72, 140.84, 126.66, 119.87. HRMS (ESI+) calculated for C14H11N2S $[M+H]^+$ m/z 239.0643, found 239.0645.

[O8] 2,5-di(pyridin-4-yl)thiophene [^{1}H 500 MHz; ^{13}C 126 MHz; $CDCl_3$]




150.57
142.59
140.70
126.53
119.74
77.28 CDCl3
77.03 CDCl3
76.78 CDCl3
f1 (ppm)

[O9] 3,3',3''-trihexyl-2,2':5',2''-terthiophene



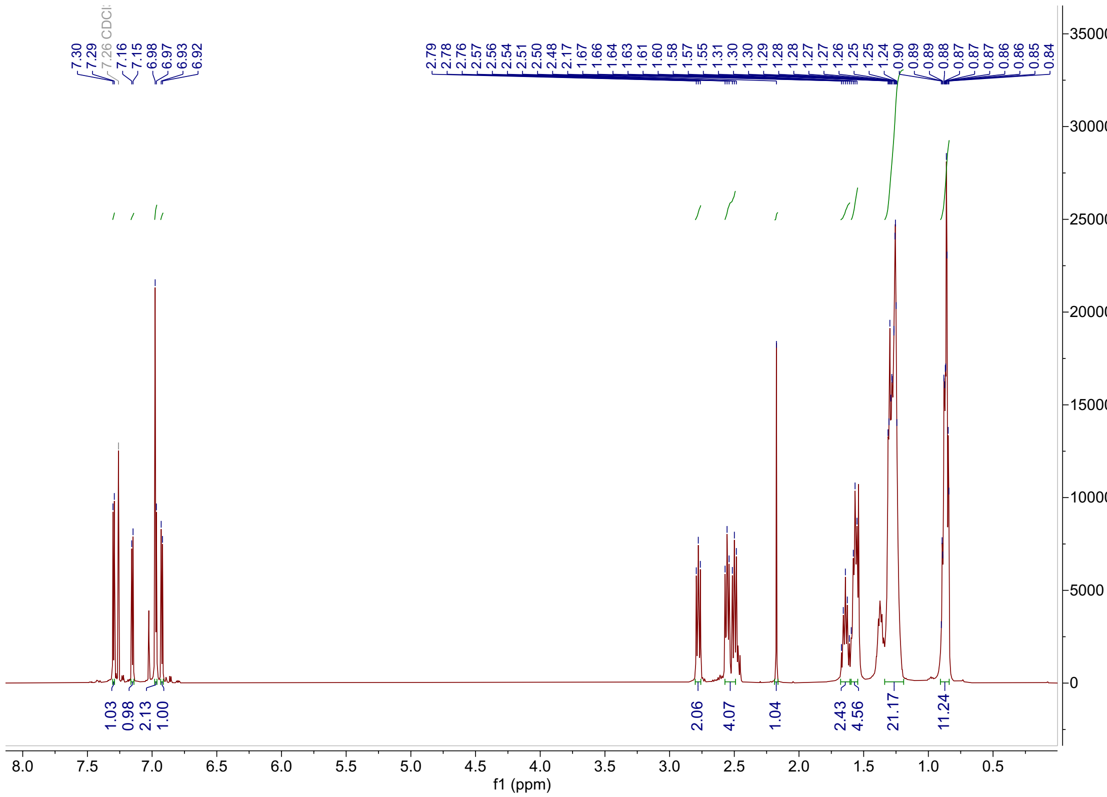
S1 (1.5 equiv) **S2** (1 equiv) **O9**

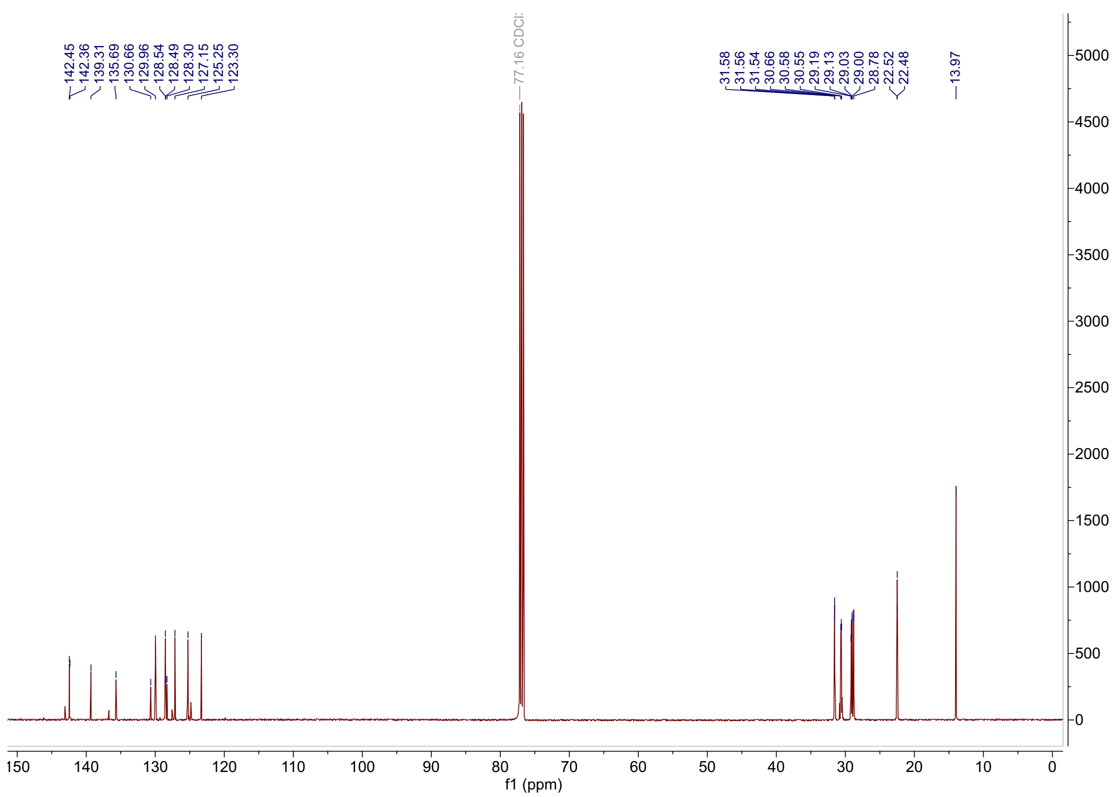
To an oven dried 10 mL round bottom flask (RBF) with a stir bar was added TMSOK (157 mg, 1.22 mmol, 1.7 equiv) and P(tBu)$_3$ Pd G4 (41 mg, 0.072 mmol, 10 mol%). The RBF was sealed with a rubber septum and purged with nitrogen (3x) using a needle attached to a Schlenk line. 2-(3,3'-dihexyl-[2,2'-bithiophen]-5-yl)-4,4,5,5-tetramethyl-1,3,2-dioxaborolane **S1** (500 mg, 1.09 mmol, 1.5 equiv) and 2-bromo-4-hexylthiophene **S2** (179 mg, 0.72 mmol, 1 equiv) were each added to an oven dried 10 mL vial and anhydrous THF (0.2 M) was added by syringe. The solution was transferred into the RBF by syringe. The RBF was heated in an oil bath at 60 °C and stirred for 1 h. The reaction vessel was removed from the hot plate and allowed to cool to room temperature. The resulting mixture was diluted in 10 mL dichloromethane (DCM) and subjected to an aqueous work-up (10 mL, H_2O). The phases were separated, and the organic layer was collected, dried over anhydrous sodium sulfate and concentrated by rotary evaporation at 36 °C. The resulting crude mixture was purified by silica gel column chromatography (100% hexanes), to yield **O9** as a brown oil (76 mg, 0.15 mmol, 21% yield).

^{1}H NMR (500 MHz, $CDCl_3$) δ 7.30 (d, *J* = 5.2 Hz, 1H), 7.15 (d, *J* = 5.2 Hz, 1H), 6.98 (s, 1H), 6.97 (d, *J* = 4.7 Hz, 1H), 6.92 (d, *J* = 5.2 Hz, 1H), 2.78 (t, 2H), 2.53 (dt, *J* = 28.1, 7.7 Hz, 4H), 1.64 (p, J = 7.5 Hz, 2H), 1.57 (p, *J* = 7.1 Hz, 4H), 1.43 – 1.15 (m, 20H), 0.87 (dtd, *J* = 9.8, 6.9, 2.5 Hz, 11H).

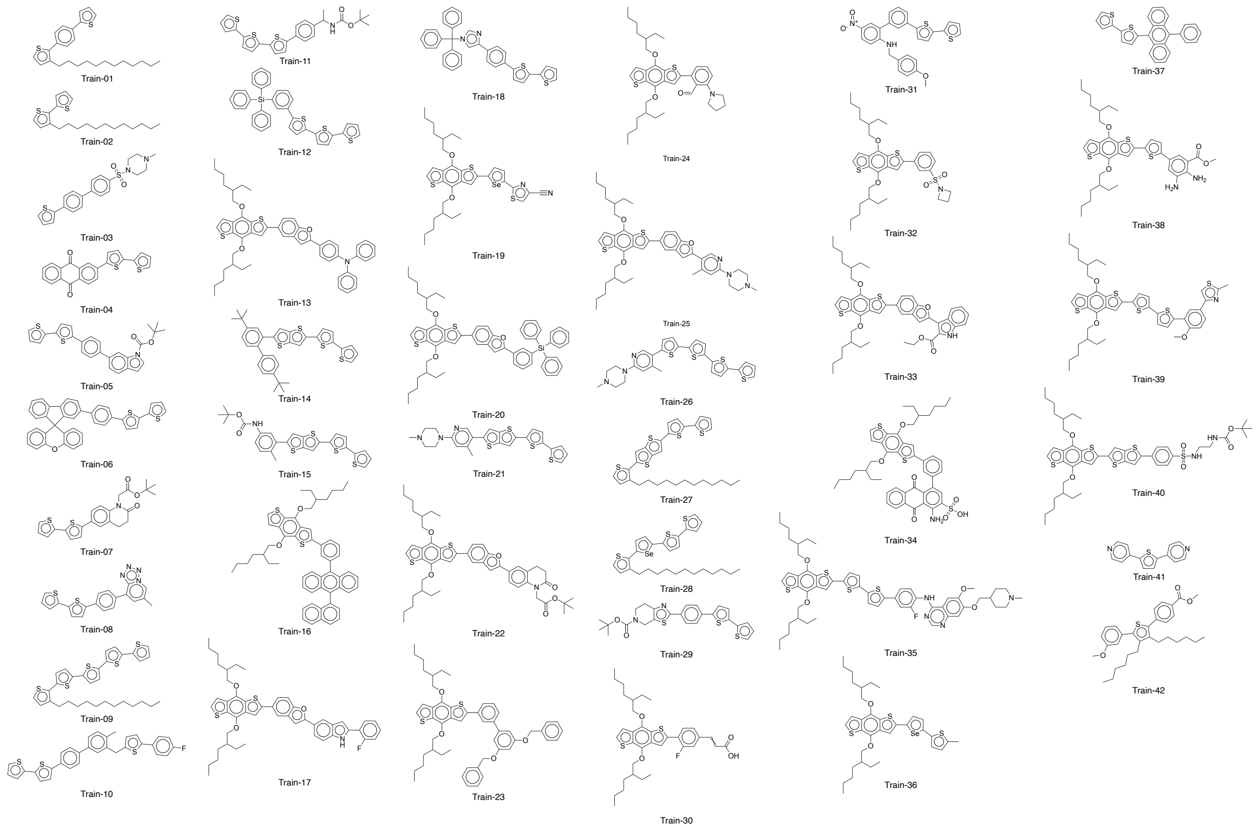
^{13}C NMR (126 MHz, $CDCl_3$) δ 142.40 (d, *J* = 11.9 Hz), 139.31, 135.69, 130.66, 129.96, 128.54, 128.49, 128.30, 127.15, 125.25, 123.30, 31.59 – 31.52 (m), 30.66, 30.57 (d, *J* = 4.3 Hz), 29.19, 29.13, 29.01 (d, *J* = 3.6 Hz), 28.78, 22.50 (d, *J* = 4.8 Hz), 13.97.

HRMS (ESI+) Calculated for $C_{30}H_{45}S_3$ $[M]^+$ m/z 501.2684, found: 501.2683.

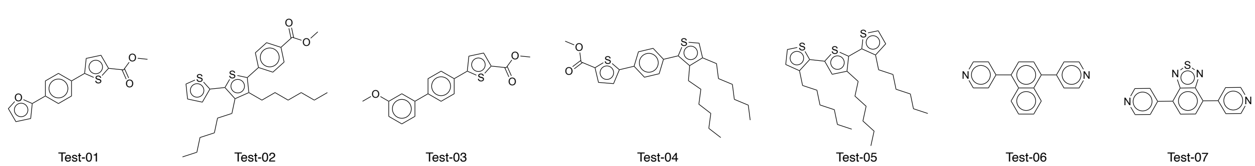

7.30
7.29
7.26 CDCl3
7.16
7.15
6.98
6.97
6.93
6.92
2.79
2.78
2.76
2.57
2.56
2.54
2.51
2.50
2.48
2.17
1.67
1.66
1.64
1.63
1.61
1.60
1.58
1.57
1.55
1.31
1.30
1.30
1.29
1.28
1.28
1.27
1.27
1.26
1.25
1.25
1.24
0.90
0.89
0.89
0.88
0.87
0.87
0.87
0.86
0.86
0.85
0.84
1.03
0.98
2.13
1.00
2.06
4.07
1.04
2.43
4.56
21.17
11.24
f1 (ppm)


142.45
142.36
139.31
135.69
130.66
129.96
128.54
128.49
128.30
127.15
125.25
123.30
77.16 CDCl3
31.58
31.56
31.54
30.66
30.58
30.55
29.19
29.13
29.03
29.00
28.78
22.52
22.48
13.97
f1 (ppm)

Training set molecules



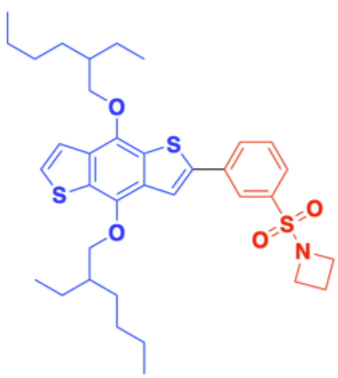
Test set molecules



Solution Testing

General solution testing procedures.

All the synthesized molecules were transferred into the N_2-filled glove box and measured under nitrogen atmosphere unless otherwise indicated. Synthesized molecules were dissolved into anhydrous chlorobenzene (99.8%) purchased from Sigma-Aldrich, stirred overnight, and diluted into 0.01mg mL^{-1} before measurement. Solutions were transferred into a quartz cuvette purchased from FireflySci, Inc, and degraded under 1 Sun illumination. Before starting a new round of molecules, DB_15_A_053 was measured each time as an internal reference to make sure minor fluctuations of environmental conditions (oxygen, moisture content, and temperature in the glovebox) have minimal impact on photodegradation rate.



[DB_15_A_053] 1-((3-(4,8-bis((2-ethylhexyl)oxy)benzo[1,2-b:4,5-b']dithiophen-2-yl)phenyl)sulfonyl)azetidine. See https://doi.org/10.1038/s41586-024-07892-1 for synthesis of **DB_15_A_053**

Solution degradation

A ScienceTech SciSun-300 solar simulator (Class AAA) with AM 1.5G illumination (100 mW cm-2) was used to degrade the synthesized molecules dissolved in chlorobenzene. We observed that the degradation rate depends on solvent choices. Therefore, the same solvent is used for testing all molecules. A ScienceTech SOL-METER-D reference detector with power meter was used as a reference cell to calibrate the illumination intensity. Entire area of solutions in the cuvette were illuminated under the solar simulator and tracked by periodically measuring the UV-Vis absorbance using an Agilent Cary 60 UV-Vis spectrometer. The temperature of the samples was monitored and kept at below 40 ℃. Error bars represent standard deviation of experiments performed in triplicate.

SO calculation

Upon measuring the UV-Vis absorbance of freshly diluted samples, their Spectral Overlap (SO) with ASTM G-173-03 reference spectra provided by NREL from 285nm to 800nm was calculated to quantify the capacity of molecules to absorb actual solar irradiation using the following formula:

$$SO = \int_{285}^{800} \left(1 - 10^{-A(\lambda)}\right) * I(\lambda) d\lambda$$

A is the absorbance of molecules, I is normalized solar irradiance in the UV-Vis region ($285 - 800\, nm, \int_{285}^{800} I(\lambda) d\lambda = 1$), and λ is the wavelength. To make the database consistent and interpretable for Gryffin, we defined SO of uncharacterizable molecules as 0.1% when the solution state absorbance spectra of molecules in chlorobenzene solution was detected at or below the UV cutoff of chlorobenzene (287nm).

T80 calculation

By periodically tracing the UV-Vis absorbance spectra during degradation, the overall weighted change in UV-Vis absorbance spectra over time was defined as Spectral Decay (SD) and calculated using the following formula:

$$SD = \int_{285}^{800} \frac{|A_I(\lambda) - A_{t_1}(\lambda)|}{\int_{285}^{800} A_I(\lambda) d\lambda}$$

$A_I(\lambda)$is the initial absorbance, and $A_{t_1}(\lambda)$is the degraded absorbance of the sample after time t_1. T_{80} of the samples were determined as the time when SD reaches 20 %, as a metric to quantify the lifetime of the molecules in the solution-state. To make the database consistent and interpretable for BO, we defined T_{80} less than 30 minutes as 0.1 hours.