Supplementary Information for

An Active Representation Learning Method for Reaction Yield Prediction with Small-scale Data

Peng-Xiang Hua[1], Zhen Huang[1], Zhe-Yuan Xu[2], Qiang Zhao[2], Chen-Yang Ye[2], Yi-Feng Wang[2*], Yun-He Xu[2*], Yao Fu[2*], Hu Ding[1*]

[1]School of Computer Science and Technology, University of Science and Technology of China; Hefei, Anhui 230026, China.

[2]Key Laboratory of Precision and Intelligent Chemistry, CAS Key Laboratory of Urban Pollutant Conversion, Anhui Province Key Laboratory of Biomass Clean Energy, Department of Chemistry, University of Science and Technology of China; Hefei, Anhui 230026, China.

*Corresponding authors: yfwangzj@ustc.edu.cn (Y.-F.W.); xyh0709@ustc.edu.cn (Y.-H.X.); fuyao@ustc.edu.cn (Y.F.); huding@ustc.edu.cn (H.D.)

Table of contents:

1. Supplementary Methods

1.1. General information

Unless otherwise noted, all reactions were performed under a nitrogen atmosphere using flame-dried glassware. All new compounds were fully characterized. ^{1}H NMR (400 MHz) spectra were recorded on a Bruker Avance 400 spectrometer and ^{1}H NMR (500 MHz) spectra were recorded on a Bruker Avance 500 spectrometer in $CDCl_3$ [using $CDCl_3$ (for ^{1}H, δ = 7.26, as the internal standard]. ^{13}C NMR (100 MHz) spectra on a Bruker Avance 400 spectrometer and ^{13}C NMR (125 MHz) spectra were recorded on a Bruker Avance 500 spectrometer in $CDCl_3$ [using $CDCl_3$ (for ^{13}C, δ = 77.0), as internal standard]. The following abbreviations were used to explain the multiplicities: s = singlet, d = doublet, t = triplet, q = quartet, dd = doublet of doublet, ddd = doublet of doublet of doublet, dt = doublet of triplet, m = multiplet, s br = single broad. High-resolution mass spectra were obtained with a Water XEVO G2 Q-Tof (Waters Corporation). Commercially available reagents were purchased from Energy Chemical, J & K Scientific, and Adamas-beta.

1.2. Synthesis of starting materials

Compounds **A2-A3** have been reported in our provious work[1], and their spectroscopic data matched those reported in the literature. Tetrachloromethane (**A1**), alkenes **B1-B6**, initiators **D1-D3** and solvents **E1-E4** were commercially available and were used as received. NHC-BH_2CN (**C1**), DMAP-BH_2CN (**C2**), DMAP-BH_3 (**C3**) and NHC−BH_3 (**C4**) were known compounds and prepared according to the literature procedures[1].

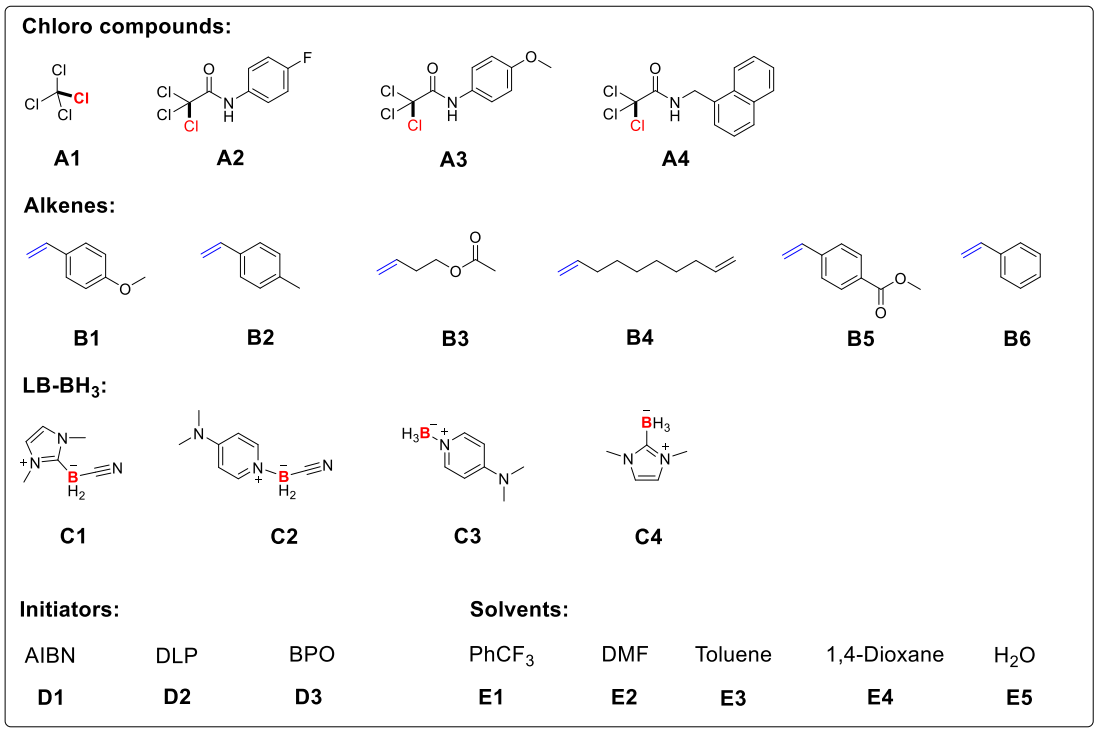
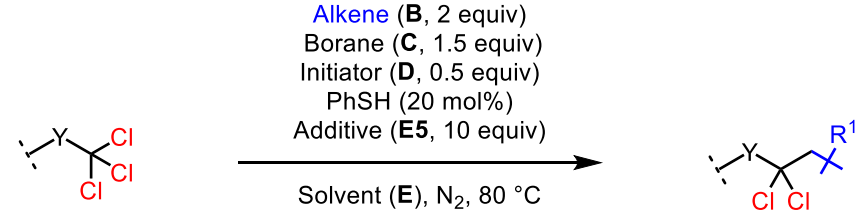


Fig. S1. Components of the reaction space of Lewis base-boryl radicals enabled coupling reaction.

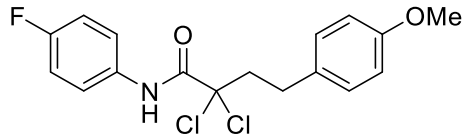
1.3. Dechlorinative coupling reactions of trichloromethyl groups



General Procedure

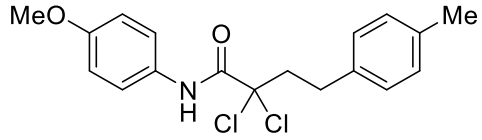
To an N_2 flushed glass tube charged with a magnetic stir bar, substrate (**A**, 0.1 mmol, 1.0 equiv), alkene (**B**, 2.0 equiv), borane (**C**, 1.5 equiv), initiator (**D**, 0.5 equiv), H_2O (10.0 equiv), PhSH (20 mol%), and solvent (**E**, 0.1 M) were added. The reaction mixture was stirred at 80℃ for 2 h. The resulting mixture was cooled down to room temperature then filtered through Na_2SO_4 under reduced pressure. The solvent was evaporated under reduced pressure. The crude mixture was purified by flash column chromatography on silica gel to give pure desired product.

2,2-Dichloro-N-(4-fluorophenyl)-4-(4-methoxyphenyl)butanamide (A)



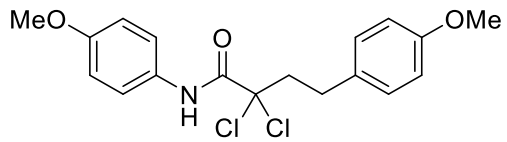
According to the ***general procedure***, the product **A** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). White solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.79-2.83 (2H, m), 2.86-2.91 (2H, m), 3.78 (3H, s), 6.82-6.86 (2H, m), 7.05-7.11 (2H, m), 7.13-7.17 (2H, m), 7.51-7.56 (2H, m), 8.46 (1H, s br); ^{13}C NMR (125 MHz, $CDCl_3$), 30.8, 46.6, 55.3, 86.6, 114.0, 115.9 (d, *J* = 22.8 Hz), 122.1 (d, *J* = 8.1 Hz), 129.5, 131.5, 132.5(d, *J* = 3.1 Hz), 158.2, 160.1 (d, *J* = 243.9 Hz), 163.4; ^{19}F NMR (376 MHz, $CDCl_3$) δ -116.1 (m). ESIHRMS: Found: *m*/*z* 378.0434. Calcd for $C_{17}H_{16}Cl_2FNO_2Na$: $(M+Na)^+$ 378.0434.

2,2-Dichloro-N-(4-methoxyphenyl)-4-(p-tolyl)butanamide (B)



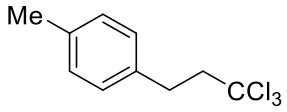
According to the ***general procedure***, the product **B** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). Yellow solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.31 (3H, s), 2.79-2.85 (2H, m), 2.87-2.91 (2H, m), 3.80 (3H, s), 6.88-6.90 (2H, m), 7.08-7.13 (4H, m), 7.44-7.47 (2H, m), 8.44 (1H, s br); ^{13}C NMR (100 MHz, $CDCl_3$) δ 21.0, 31.2, 46.6, 55.5, 86.8, 114.3, 122.1, 128.4, 129.2, 129.5, 135.9, 136.5, 157.3, 163.2. Found: *m*/*z* 374.0685. ESIHRMS: Calcd for $C_{18}H_{19}Cl_2NO_2Na$: $(M+Na)^+$ 374.0685.

2,2-Dichloro-N,4-bis(4-methoxyphenyl)butanamide (C)



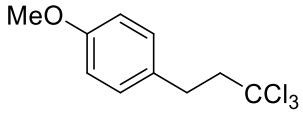
According to the ***general procedure***, the product **C** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). White solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.79-2.83 (2H, m), 2.86-2.90 (2H, m), 3.78 (3H, s), 3.81 (3H, s), 6.82-6.86 (2H, m), 6.89-6.93 (2H, m), 7.14-7.17 (2H, m), 7.46-7.50 (2H, m), 8.41 (1H, s br); ^{13}C NMR (100 MHz, $CDCl_3$) δ 30.8, 46.7, 55.2, 55.5, 86.7, 114.0, 114.3, 122.0, 129.5, 129.5, 131.6, 157.3, 158.1, 163.2. ESIHRMS: Found: *m*/*z* 390.0628. Calcd for $C_{18}H_{19}Cl_2NO_3Na$: $(M+Na)^+$ 390.0634.

1-Methyl-4-(3,3,3-trichloropropyl)benzene (D)



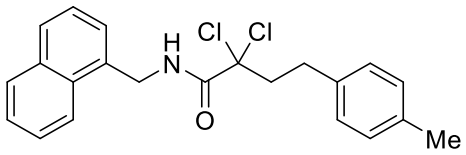
According to the ***general procedure*** with slight modification of the reaction conditions using carbon tetrachloride **A1** (4.0 equiv.), 1-methyl-4-vinylbenzene **B2** (1.0 equiv.), the product **D** was prepared and purified by column chromatography on silica gel (petroleum ether). White solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.34 (3H, s), 2.93-2.97 (2H, m), 2.99-3.10 (2H, m), 7.14 (4H, s); ^{13}C NMR (125 MHz, $CDCl_3$) δ 21.0, 32.3, 56.9, 99.3, 128.4, 129.4, 135.7, 136.2. EIHRMS: Found: *m*/*z* 236.9940. Calcd for $C_{10}H_{12}Cl_3$: $(M+H)^+$ 236.9999.

1-Methoxy-4-(3,3,3-trichloropropyl)benzene (E)



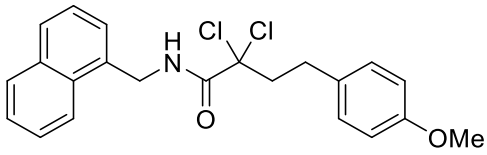
According to the ***general procedure*** with slight modification of the reaction conditions using carbon tetrachloride **A1** (4.0 equiv.), 1-methoxy-4-vinylbenzene **B1** (1.0 equiv.), the product **E** was prepared and purified by column chromatography on silica gel (petroleum ether). White solid. ^{1}H NMR (500 MHz, $CDCl_3$) δ 2.89-2.99 (2H, m), 2.99-3.07 (2H, m), 3.80 (3H, s), 6.82-6.89 (2H, m), 7.12-7.18 (2H, m); ^{13}C NMR (125 MHz, $CDCl_3$) δ 31.9, 55.3, 57.0, 99.3, 114.2, 129.5, 130.9, 158.4. ESIHRMS: Found: m/z 252.9886. Calcd for $C_{10}H_{12}Cl_3O$: $(M+H)^+$ 252.9948.

2,2-Dichloro-N-(naphthalen-2-ylmethyl)-4-(p-tolyl)butanamide (F)



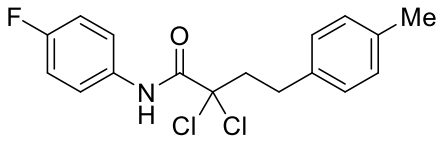
According to the ***general procedure***, the product **F**[1] was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 30 : 1). White solid. ^{1}H NMR (500 MHz, $CDCl_3$) δ 2.33 (3H, s), 2.77-2.81 (2H, m), 2.83-2.86 (2H, m), 4.99 (2H, d, *J* = 5.5 Hz), 7.07 (1H, s br), 7.09-7.13 (4H, m), 7.45-7.50 (2H, m), 7.52-7.58 (2H, m), 7.87 (1H, d, *J* = 8.0 Hz), 7.91 (1H, dd, *J* = 7.5, 1.5 Hz), 7.99 (1H, d, *J* = 8.0 Hz); ^{13}C NMR (125 MHz, $CDCl_3$) δ 21.0, 31.1, 42.9, 46.8, 86.3, 123.2, 125.4, 126.2, 126.8, 126.9, 128.4, 128.9, 129.1, 129.2, 131.3, 132.2, 133.9, 135.9, 136.6, 165.3. ESIHRMS: Found: m/z 386.1077. Calcd for $C_{22}H_{22}Cl_2NO$: $(M+H)^+$ 386.1073.

2,2-Dichloro-4-(4-methoxyphenyl)-N-(naphthalen-2-ylmethyl)butanamide (G)



According to the ***general procedure***, the product **G** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 30 : 1). White solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.75-2.79 (2H, m), 2.80-2.85 (2H, m), 3.80 (3H, s), 4.98 (2H, d, *J* = 5.2 Hz), 6.83-6.86 (2H, m), 7.07 (1H, s br), 7.12-7.15 (2H, m), 7.44-7.59 (2H, m), 7.51-7.60 (2H, m), 7.85-7.87 (1H, m), 7.90-7.92 (1H, m), 7.97-7.99 (1H, m); ^{13}C NMR (125 MHz, $CDCl_3$) δ 30.7, 43.0, 46.9, 55.3, 86.5, 114.0, 123.2, 125.4, 126.2, 126.8, 126.9, 128.9, 129.2, 129.5, 131.3, 131.7, 132.2, 134.0, 158.2, 165.4. ESIHRMS: Found: m/z 424.0840. Calcd for $C_{22}H_{21}Cl_2NO_2Na$: $(M+Na)^+$ 424.0841.

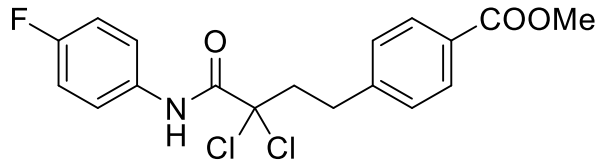
2,2-Dichloro-N-(4-fluorophenyl)-4-(p-tolyl)butanamide (H)



According to the ***general procedure***, the product **H** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). White solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 2.33 (3H, s), 2.81-2.86 (2H, m), 2.89-2.94 (2H, m), 7.05-7.15 (6H, m), 7.51-7.57 (2H, m), 8.47 (1H, s br); ^{13}C NMR (100 MHz, $CDCl_3$) δ 21.0, 31.2,

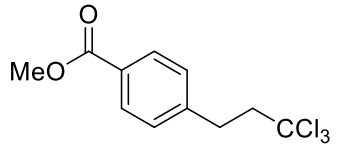
46.5, 86.6, 116.0 (d, J = 22.6 Hz), 122.2 (d, J = 8.0 Hz), 128.4, 129.3, 132.6 (d, J = 2.8 Hz), 136.0, 136.4, δ 161.1 (d, J = 244.0 Hz), 163.5. ^{19}F NMR (376 MHz, $CDCl_3$) δ -116.1 (m). ESIHRMS: Found: m/z 340.0669. Calcd for $C_{17}H_{17}Cl_2FNO$: $(M+H)^+$ 340.0666.

Methyl 4-(3,3-dichloro-4-((4-fluorophenyl)amino)-4-oxobutyl)benzoate (I)



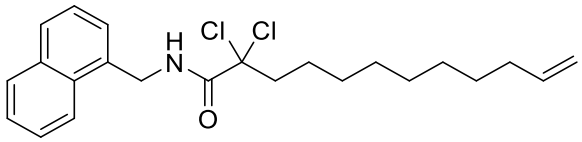
According to the ***general procedure***, the product **I** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). Yellow oil. ^{1}H NMR (500 MHz, $CDCl_3$) δ 2.84-2.86 (2H, m), 2.99-3.02 (2H, m), 3.91 (3H, s), 7.07-7.10 (2H, m), 7.30-7.32 (2H, m), 7.51-7.55 (2H, m), 7.96-7.98 (2H, m), 8.45 (1H, s br); ^{13}C NMR (125 MHz, $CDCl_3$) δ 31.7, 45.7, 52.1, 86.2, δ 116.0 (d, J = 22.8 Hz) , δ 122.1 (d, J = 8.1 Hz), 128.6, 130.0, δ 132.4 (d, J = 3.1 Hz), 144.9, δ 160.1 (d, J = 244.0 Hz), 163.3, 166.9. ^{19}F NMR (376 MHz, $CDCl_3$) δ -116.0 (m). ESIHRMS: Found: m/z 406.0392. Calcd for $C_{18}H_{16}Cl_2FNO_3Na$: $(M+Na)^+$ 406.0383.

Methyl-4-(3,3,3-trichloropropyl)benzoate (J)



According to the ***general procedure*** with slight modification of the reaction conditions using carbon tetrachloride **A1** (4.0 equiv.), methyl 4-vinylbenzoate **B5** (1.0 equiv.), the product **J** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 100 : 1). White solid. ^{1}H NMR (500 MHz, $CDCl_3$) δ 2.97-3.00 (2H, m), 3.14-3.17 (2H, m), 3.92 (3H, s), 7.31-7.32 (2H, m), 7.99-8.01 (2H, m); ^{13}C NMR (125 MHz, $CDCl_3$) δ 32.8, 52.1, 56.1, 99.0, 128.6, 128.7, 130.1, 144.2, 166.9. EIHRMS: Found: m/z 280.9828. Calcd for $C_{11}H_{12}Cl_3O_2$: $(M+H)^+$ 280.9898.

2,2-Dichloro-N-(naphthalen-2-ylmethyl)dodec-11-enamide (K)



According to the ***general procedure*** with slight modification of the reaction conditions using 2,2,2-trichloro-*N*-(naphthalen-2-ylmethyl)acetamide **A4** (1.0 equiv.), deca-1,9-diene **B4** (8.0 equiv.), the product **K** was prepared and purified by column chromatography on silica gel (petroleum ether : ethyl acetate = 50 : 1). Colorless oil.^{1}H NMR (500 MHz, $CDCl_3$) δ 1.25-1.39 (10H, m), 1.49-1.56 (2H, m), 2.01-2.07 (2H, m), 2.47-2.51 (2H, m), 4.92-5.03 (4H, m), 5.82 (1H, ddt, *J* = 16.9, 10.2, 6.7 Hz), 7.06 (1H, t, *J* = 5.1 Hz, br), 7.44-7.49 (2H, m), 7.51-7.58 (2H, m), 7.85-7.87 (1H, m), 7.89-7.91 (1H, m), 7.95-7.97 (1H, m); ^{13}C NMR (125 MHz, $CDCl_3$) δ 25.2, 28.8, 28.9, 29.0, 29.2, 29.3, 33.8, 42.9, 45.0, 87.4, 114.2, 123.2, 125.4, 126.2, 126.7, 126.8, 128.9, 129.1, 131.3, 132.2, 133.9, 139.2, 165.5. ESIHRMS: Found: m/z 428.1516. Calcd for $C_{23}H_{29}Cl_2NONa$: $(M+Na)^+$ 428.1518.

2. Supplementary Figures

^{1}H, ^{13}C, and ^{19}F NMR Spectra

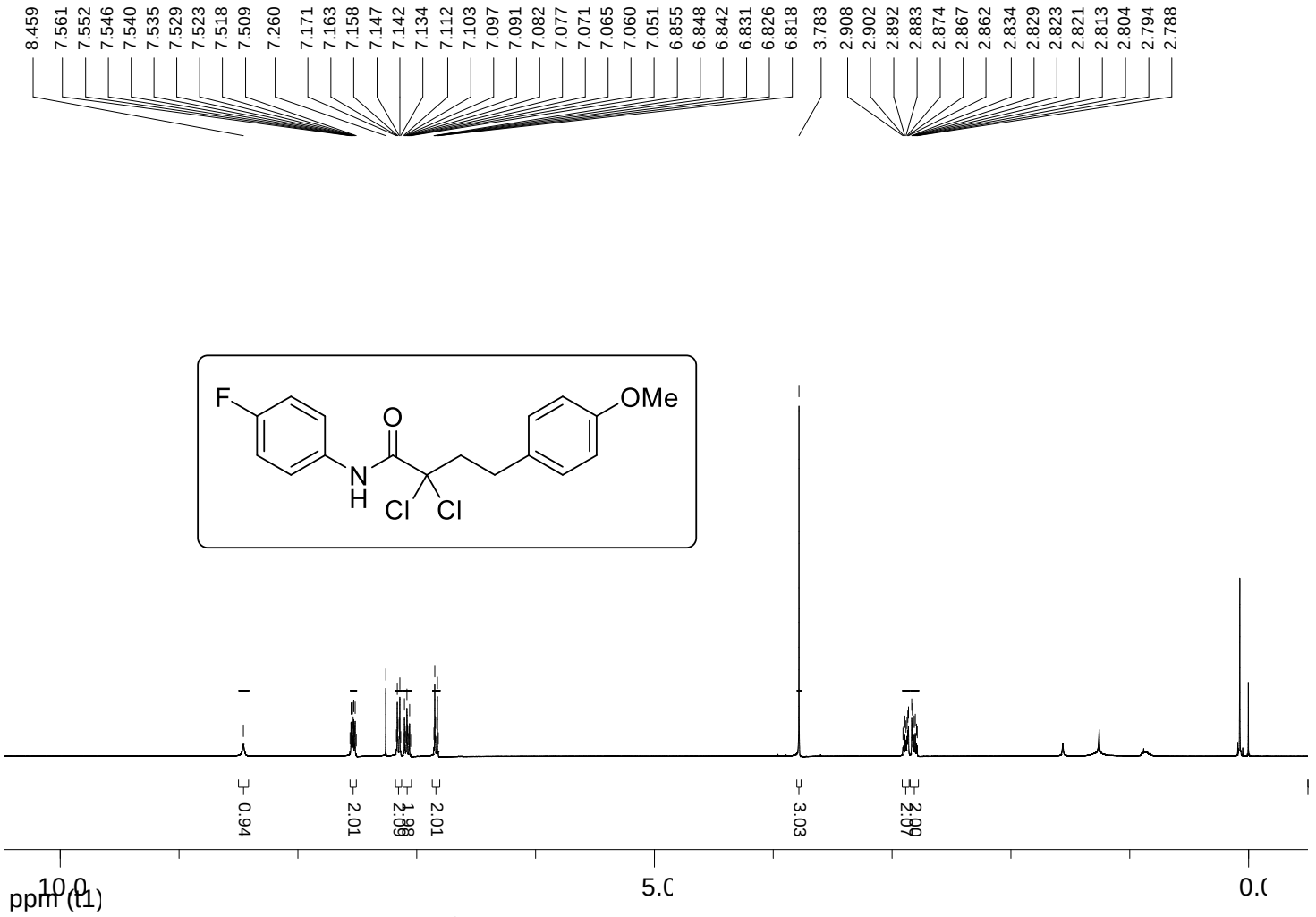


Fig. S1. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of A

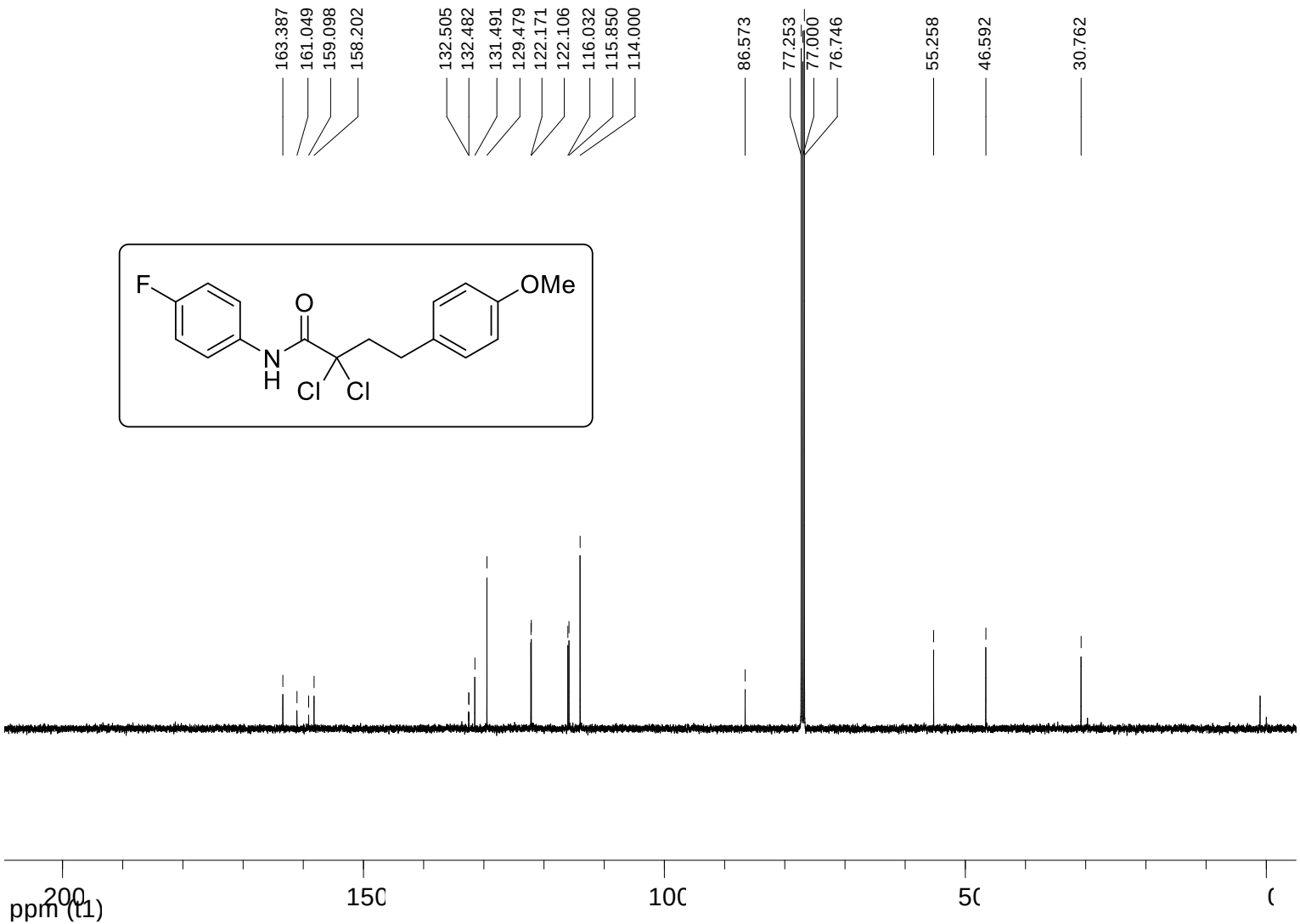


Fig. S2. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of A

5

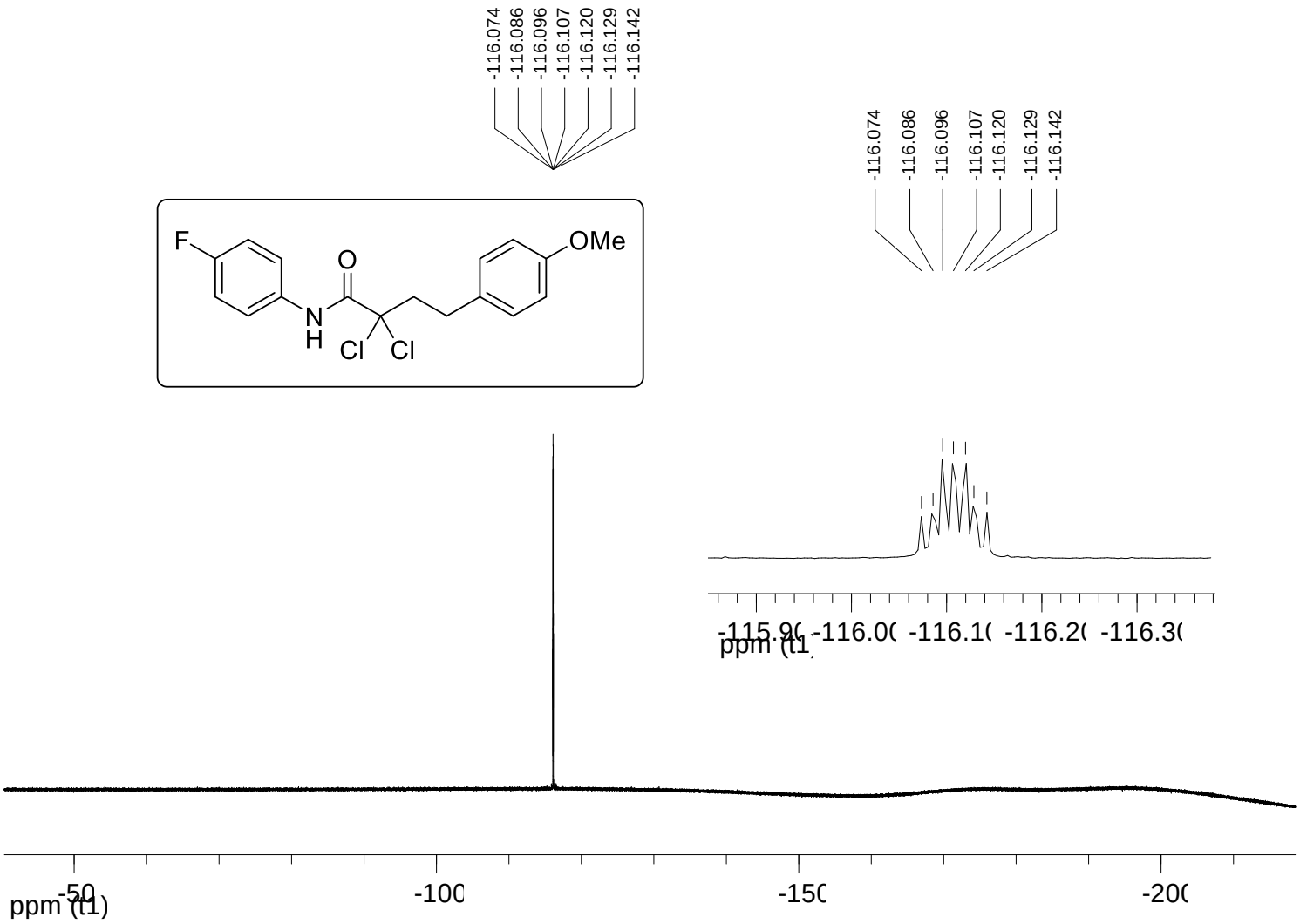


Fig. S3. ^{19}F NMR (376 MHz, $CDCl_3$) spectrum of A

5

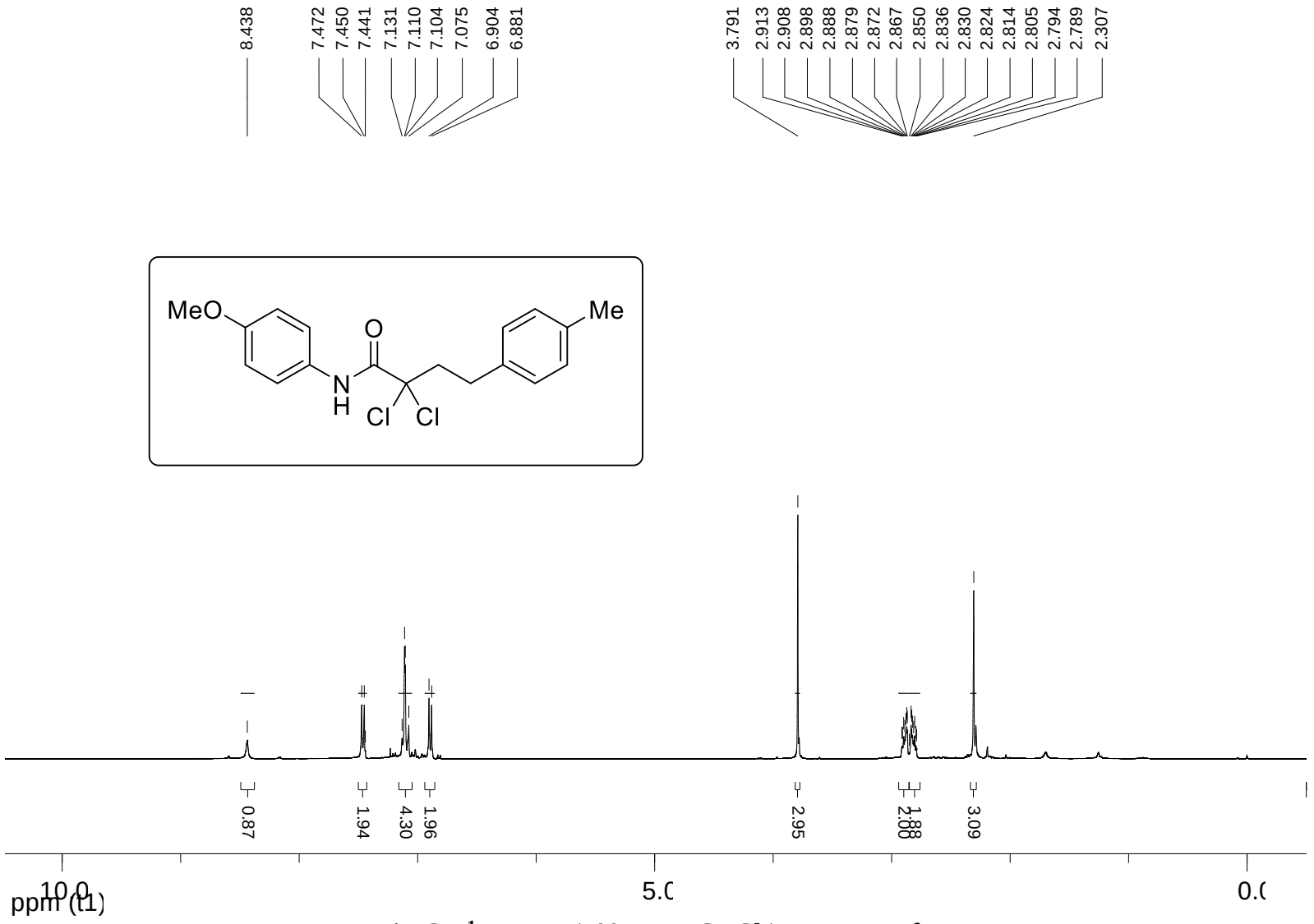


Fig. S4. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of B

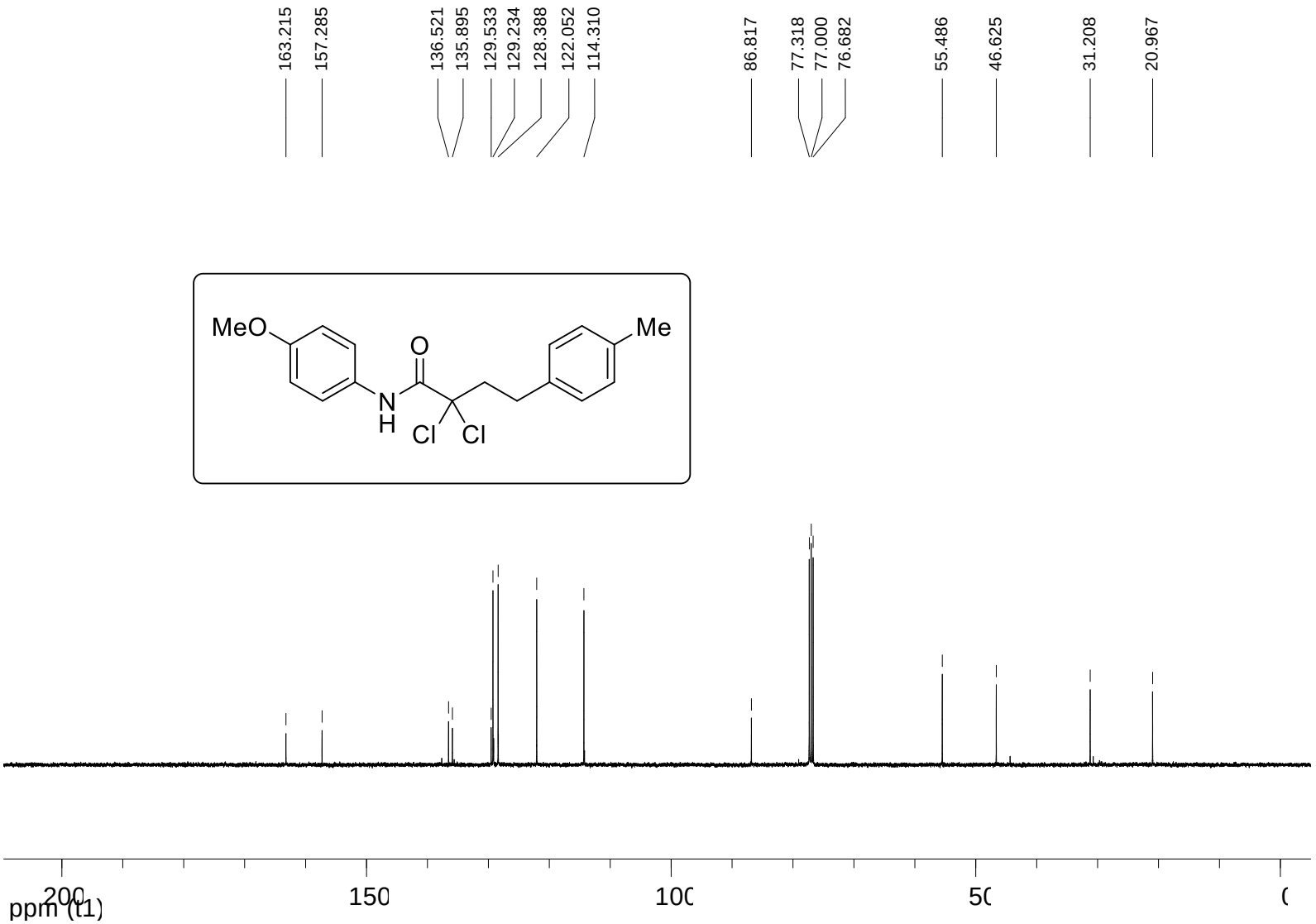


Fig. S5. ^{13}C NMR (100 MHz, $CDCl_3$) spectrum of B

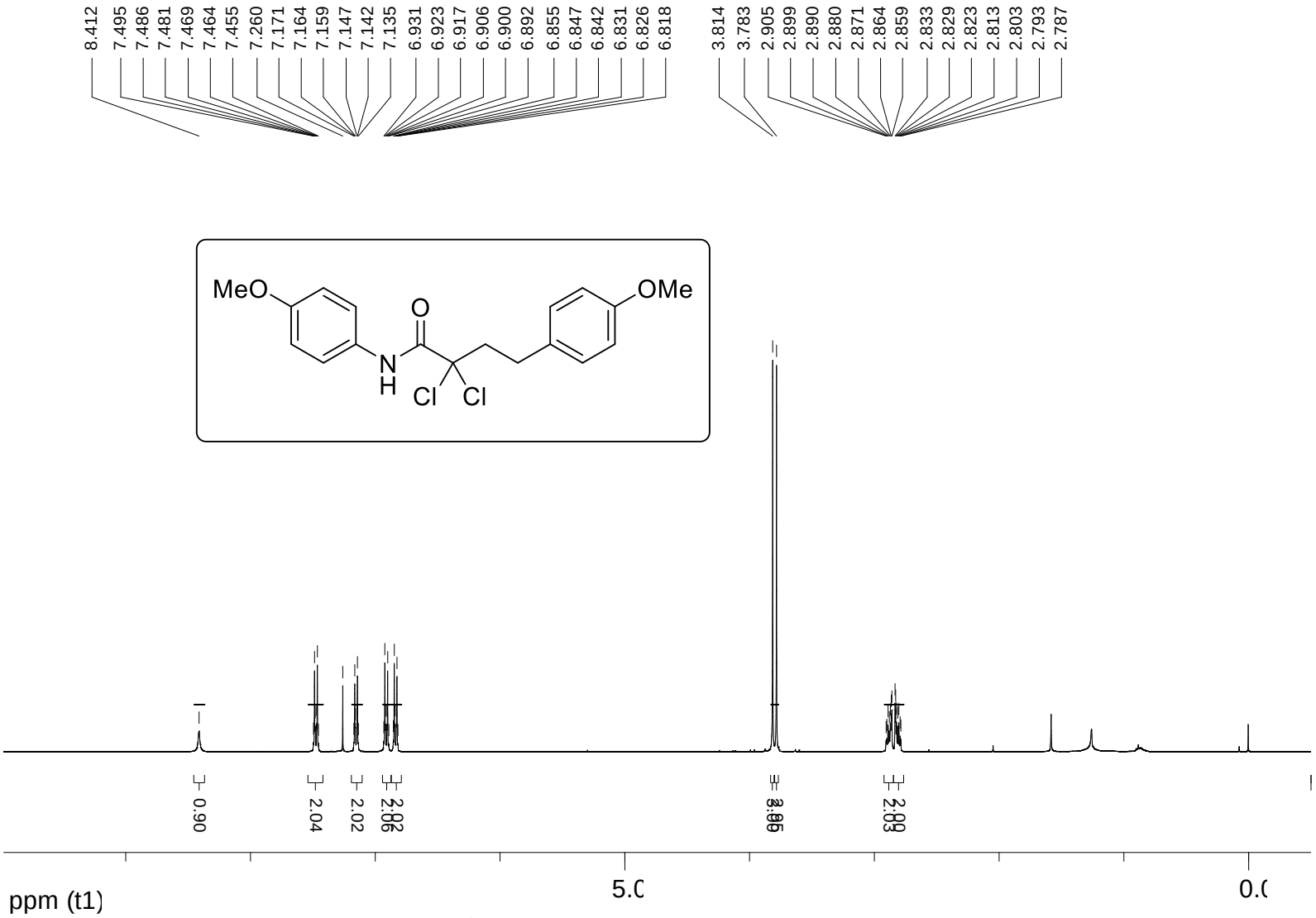


Fig. S6. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of C

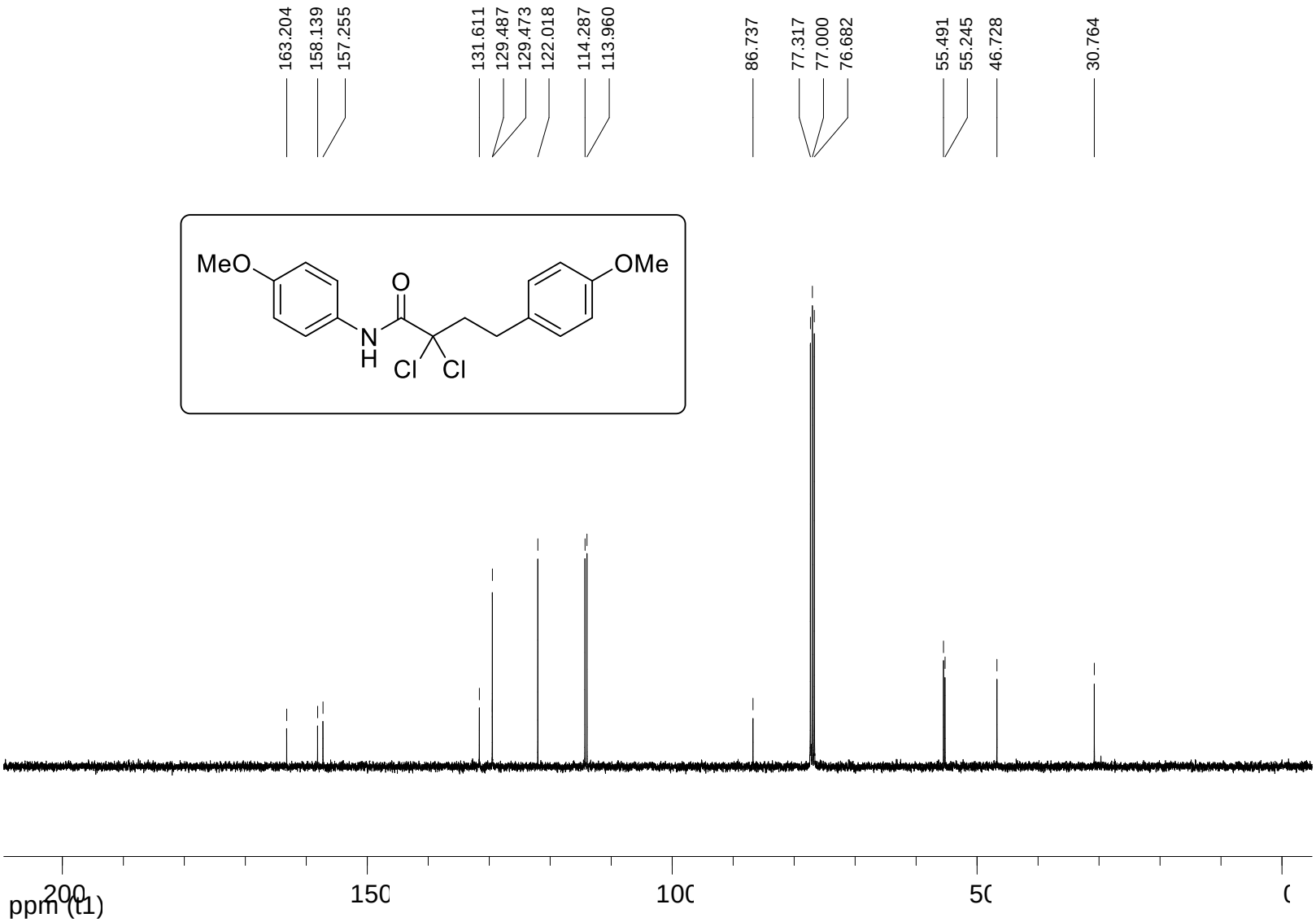


Fig. S7. ^{13}C NMR (100 MHz, $CDCl_3$) spectrum of C

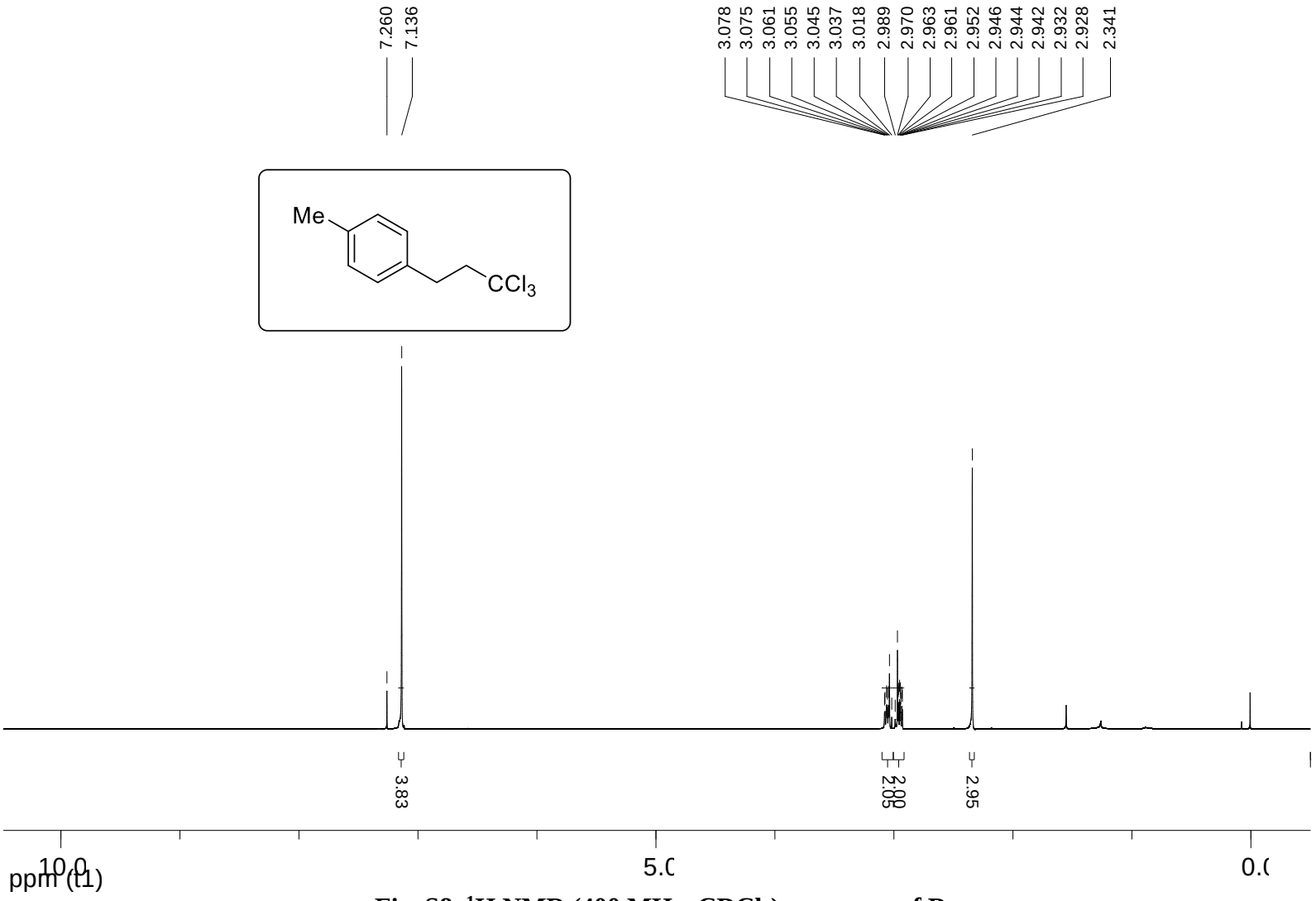


Fig. S8. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of D

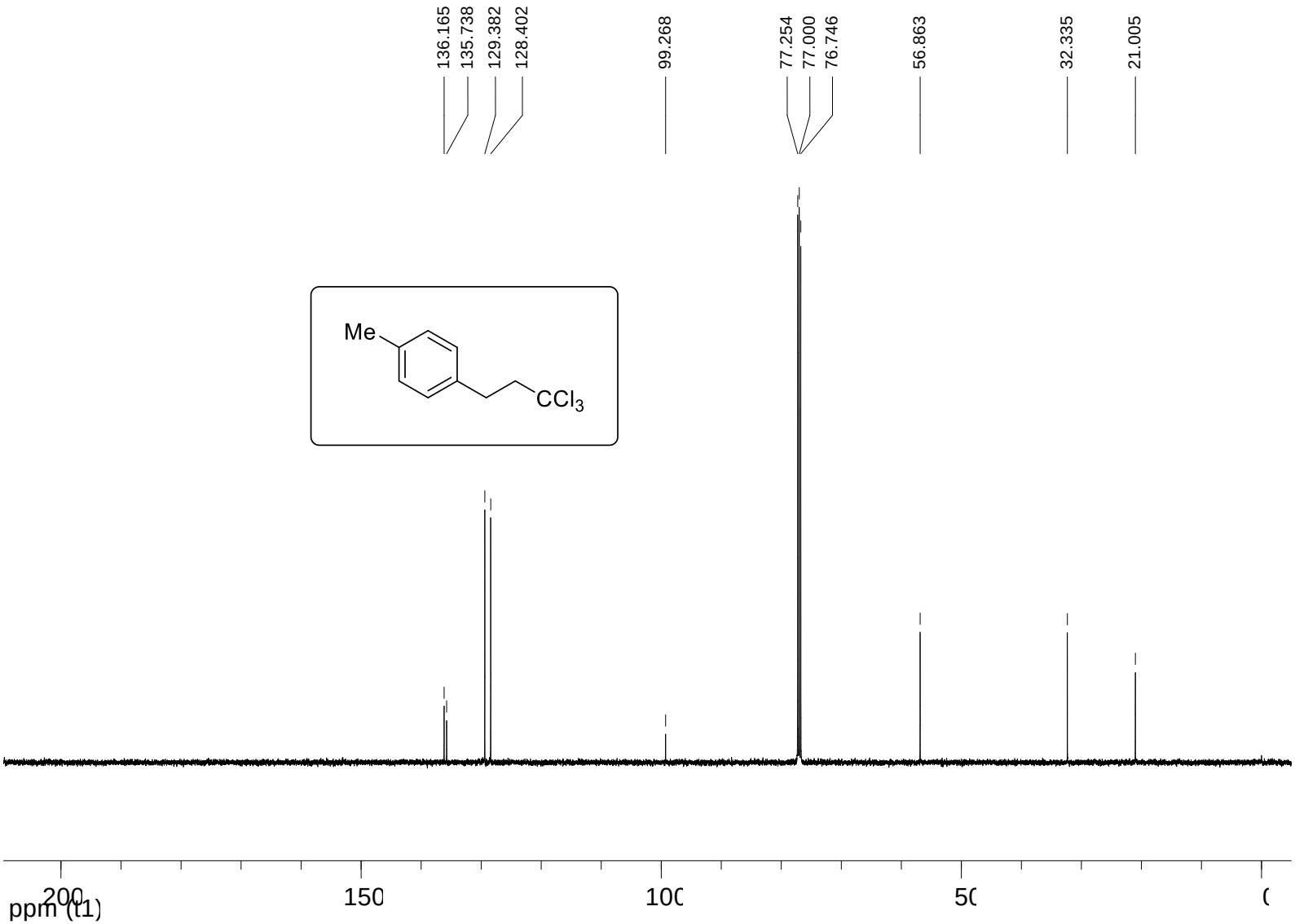


Fig. S9. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of D

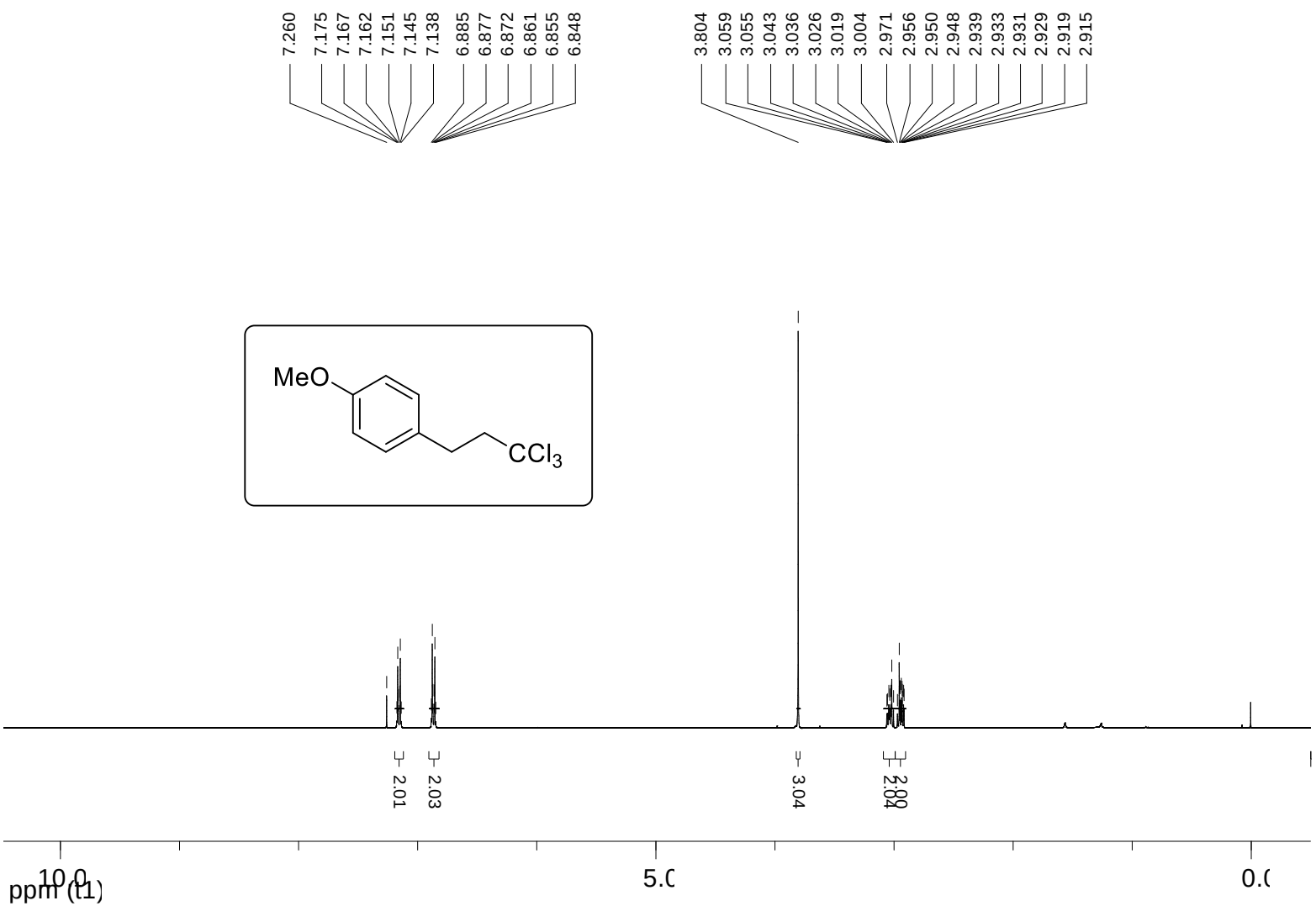


Fig. S10. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of E

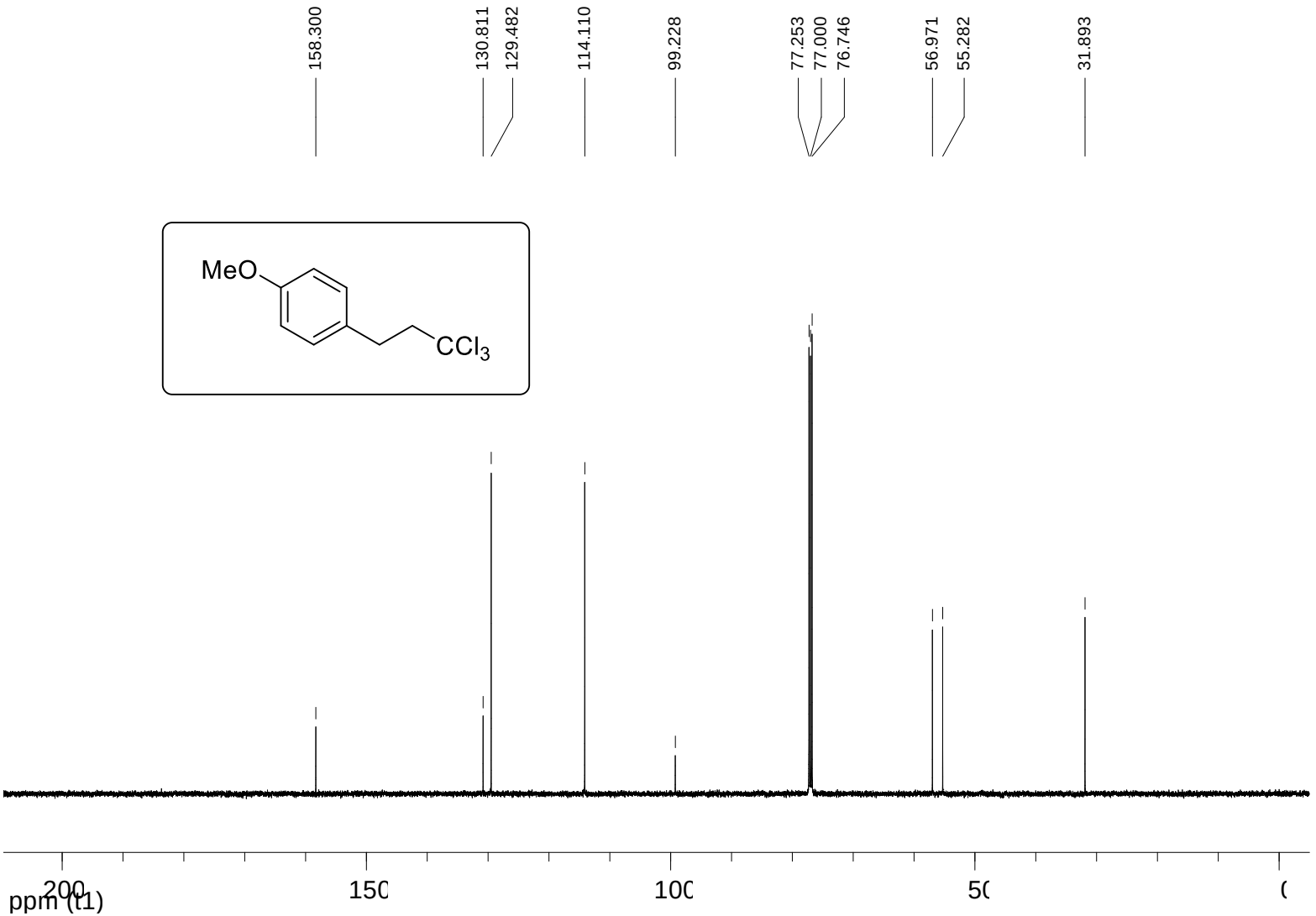


Fig. S11. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of E

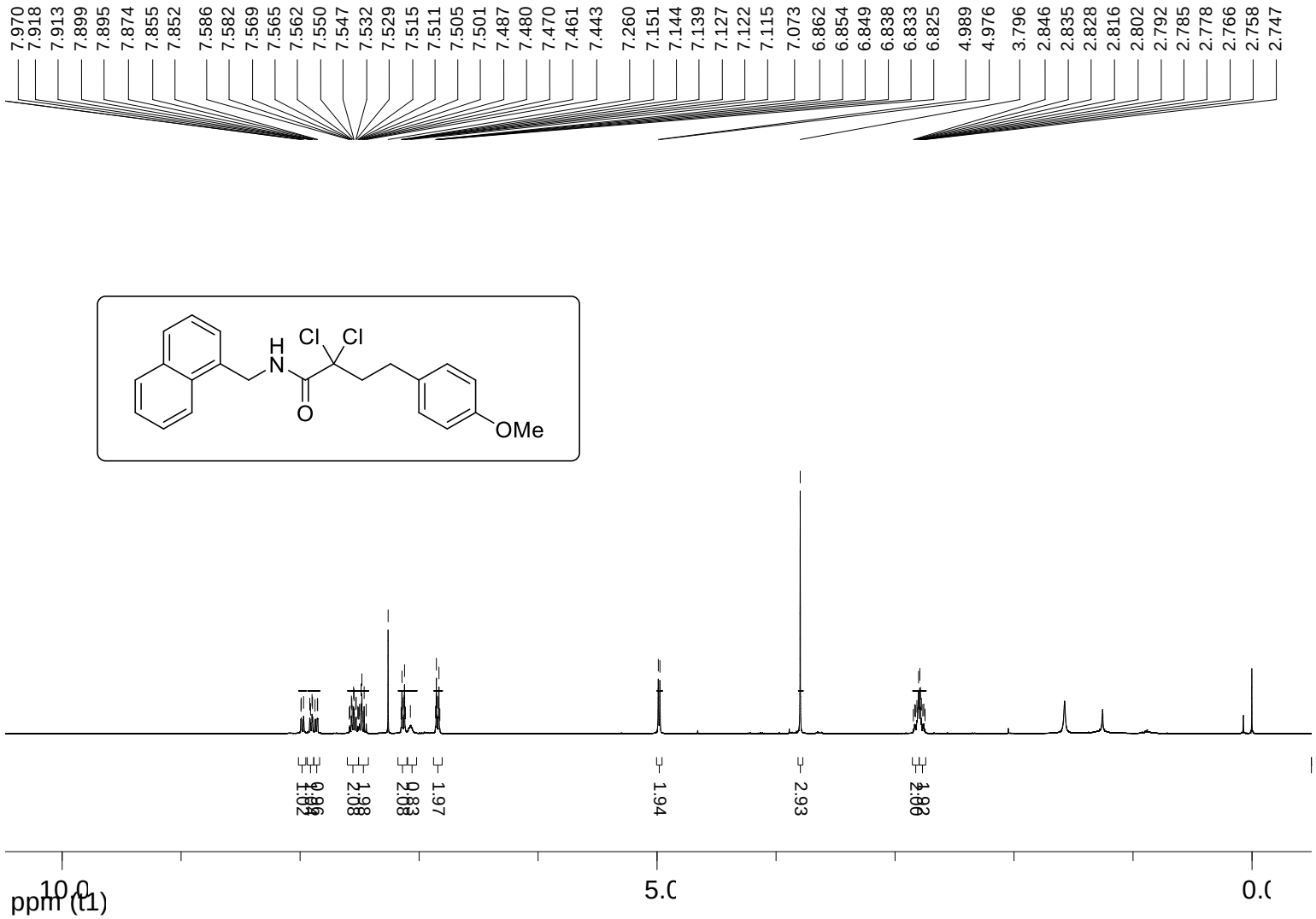


Fig. S12. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of G

5

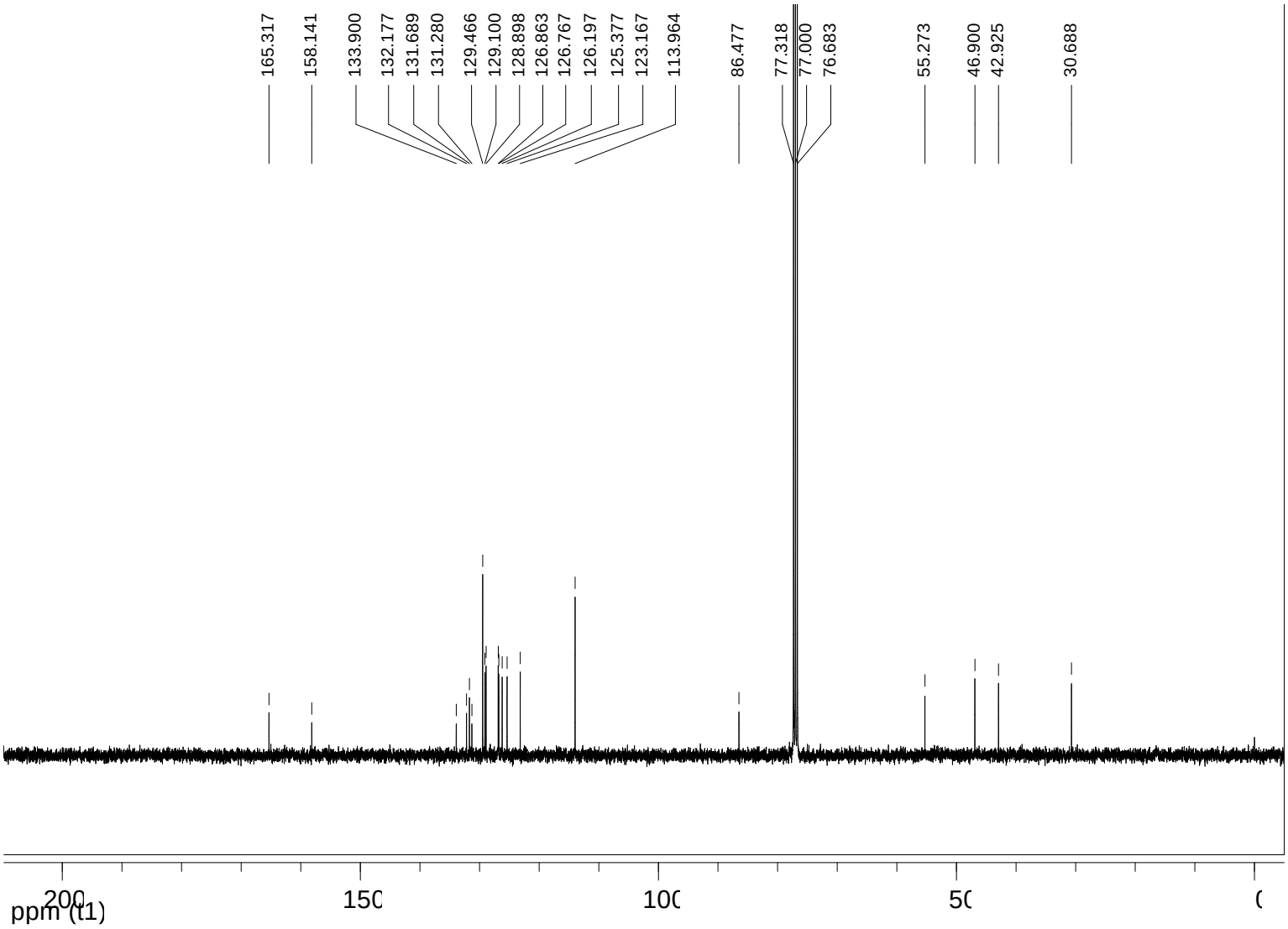


Fig. S13. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of G

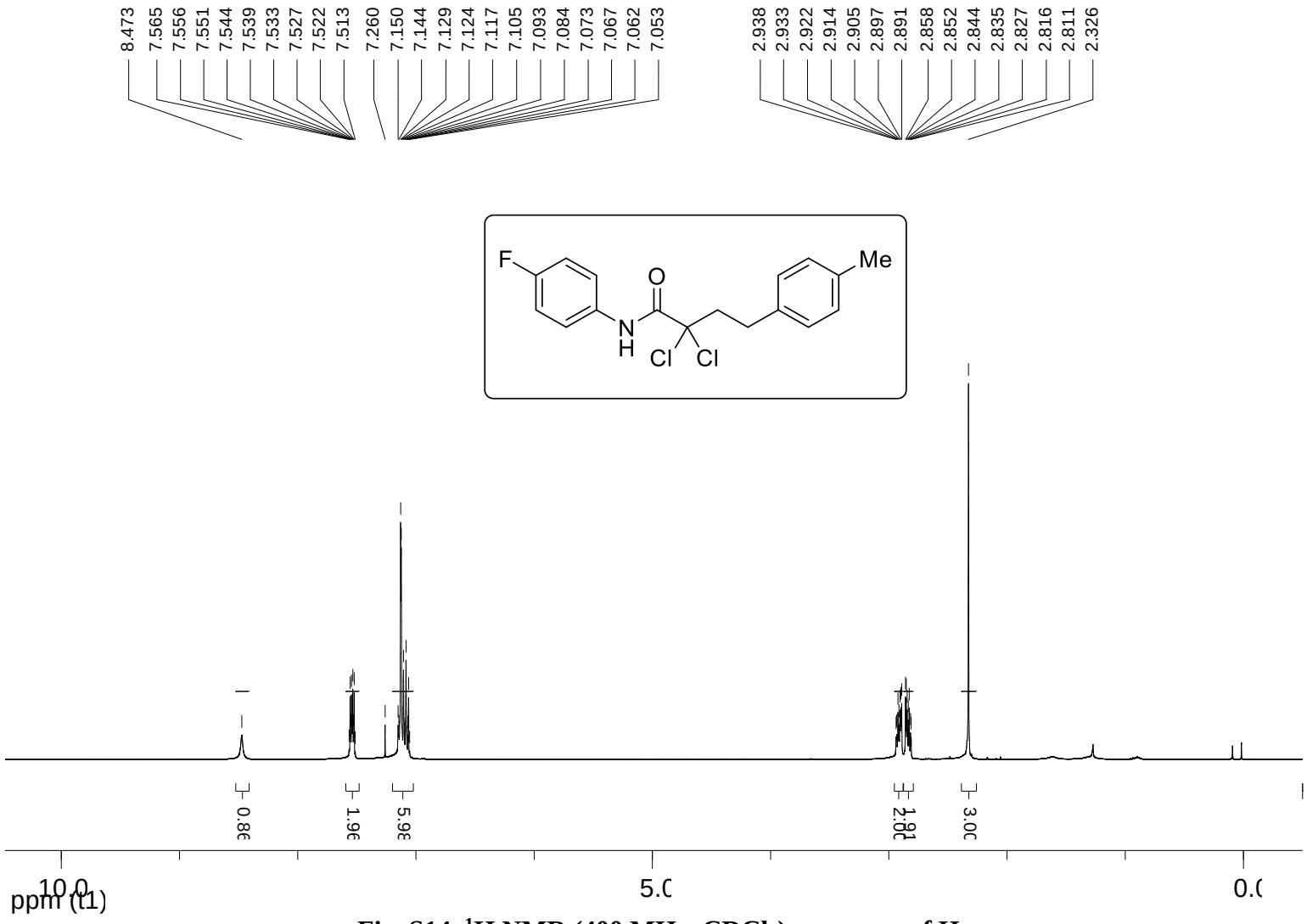


Fig. S14. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of H

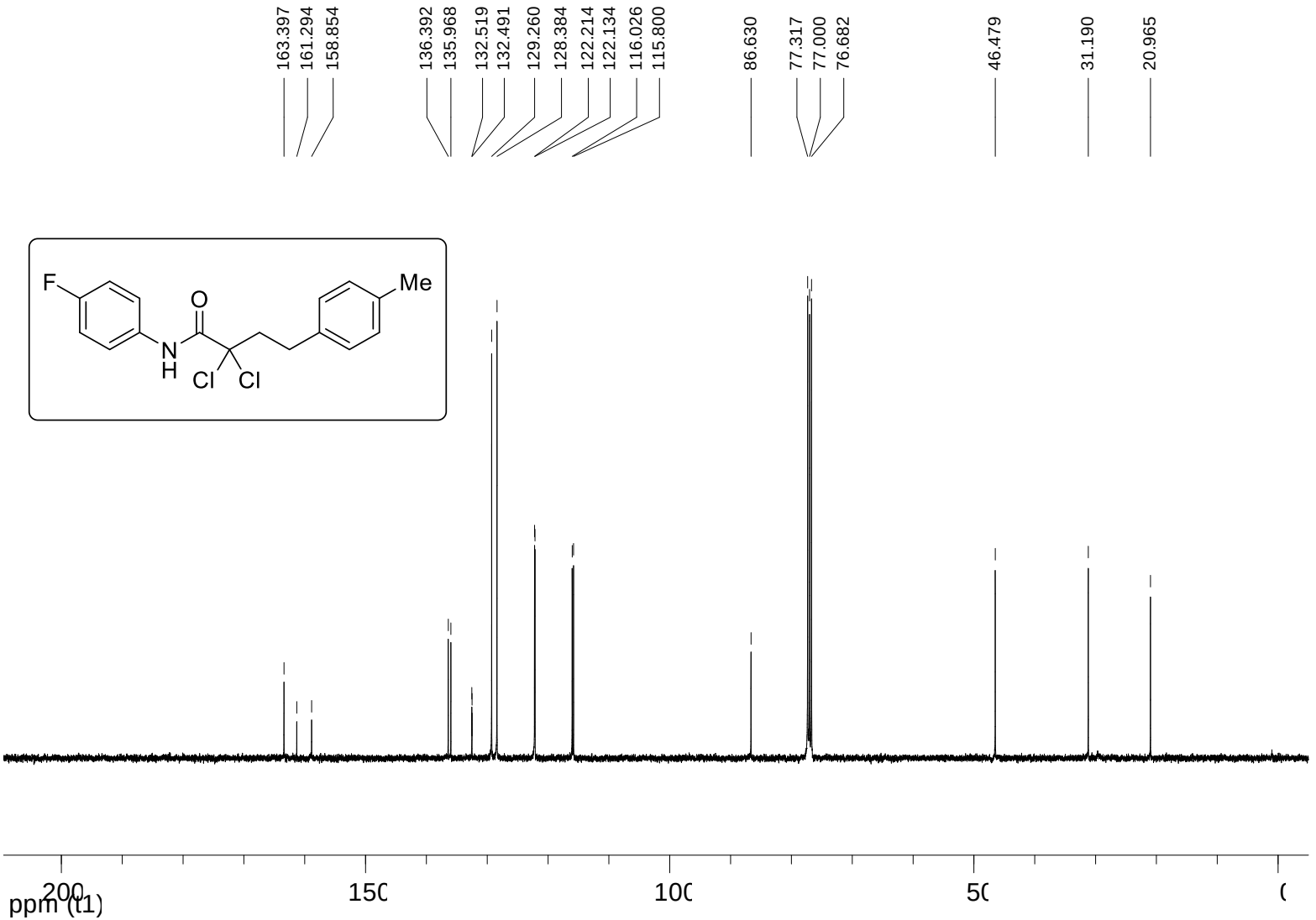


Fig. S15. ^{13}C NMR (100 MHz, $CDCl_3$) spectrum of H

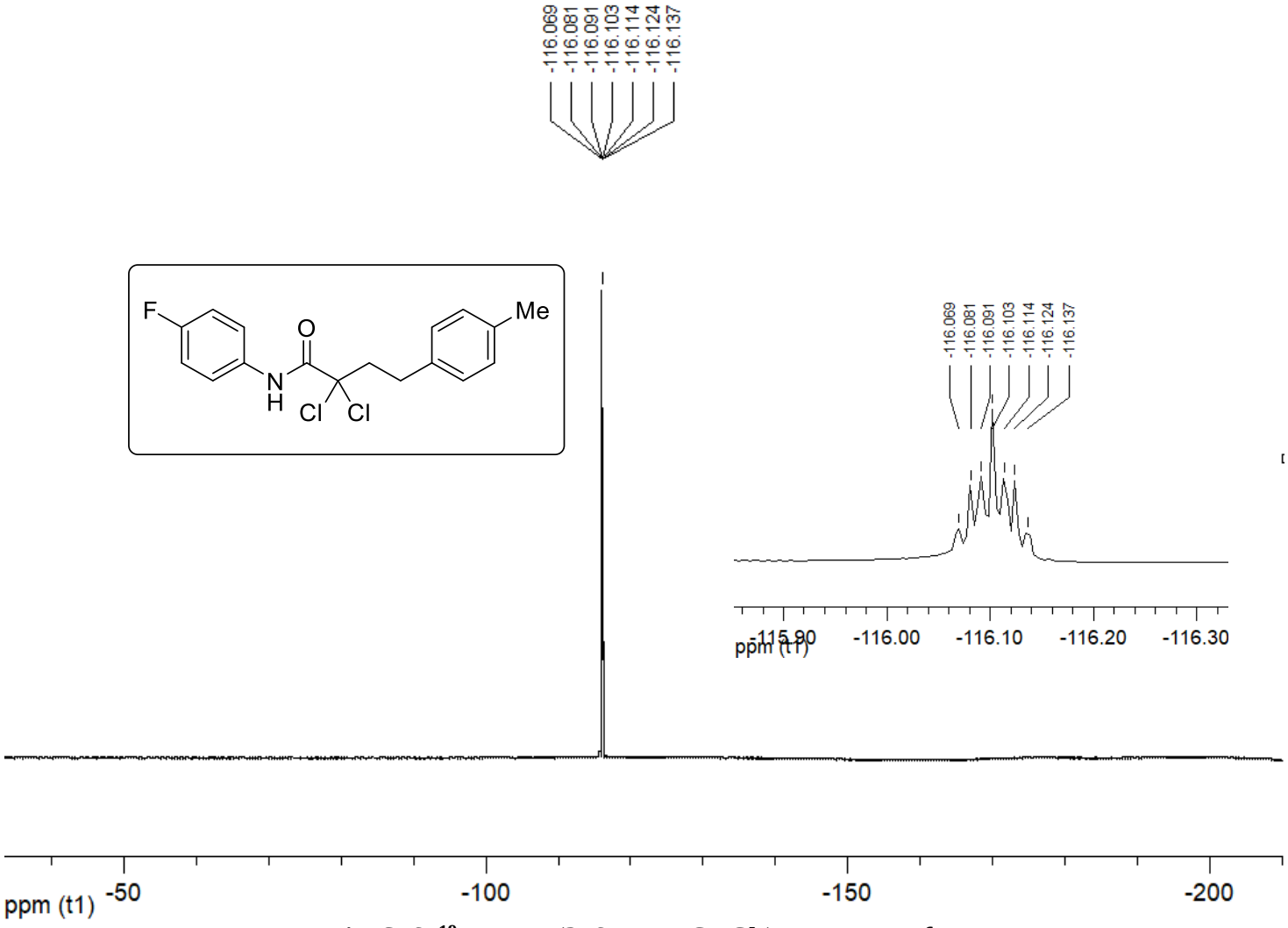


Fig. S16. ^{19}F NMR (376 MHz, $CDCl_3$) spectrum of H

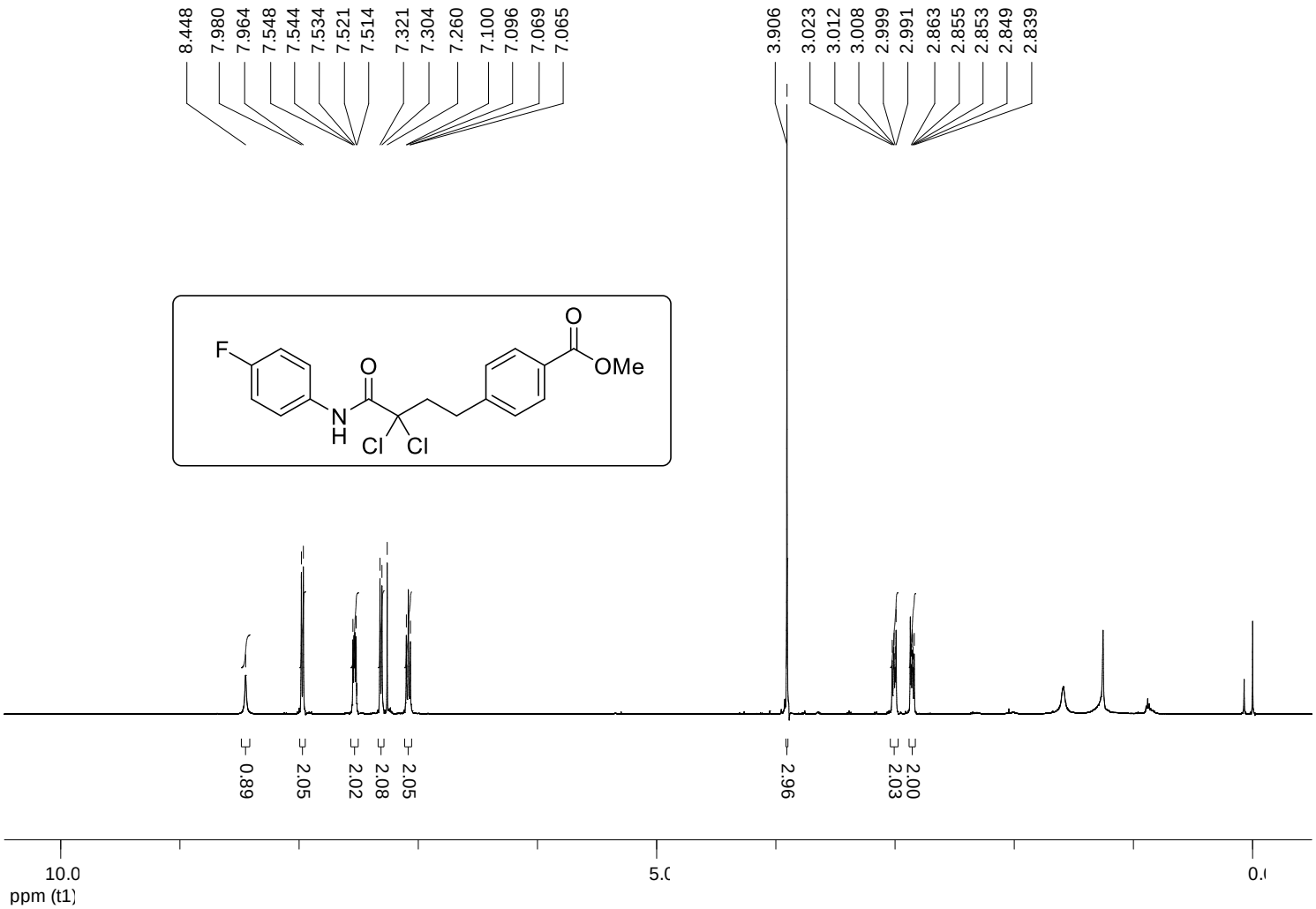


Fig. S17. ^{1}H NMR (500 MHz, $CDCl_3$) spectrum of I

5

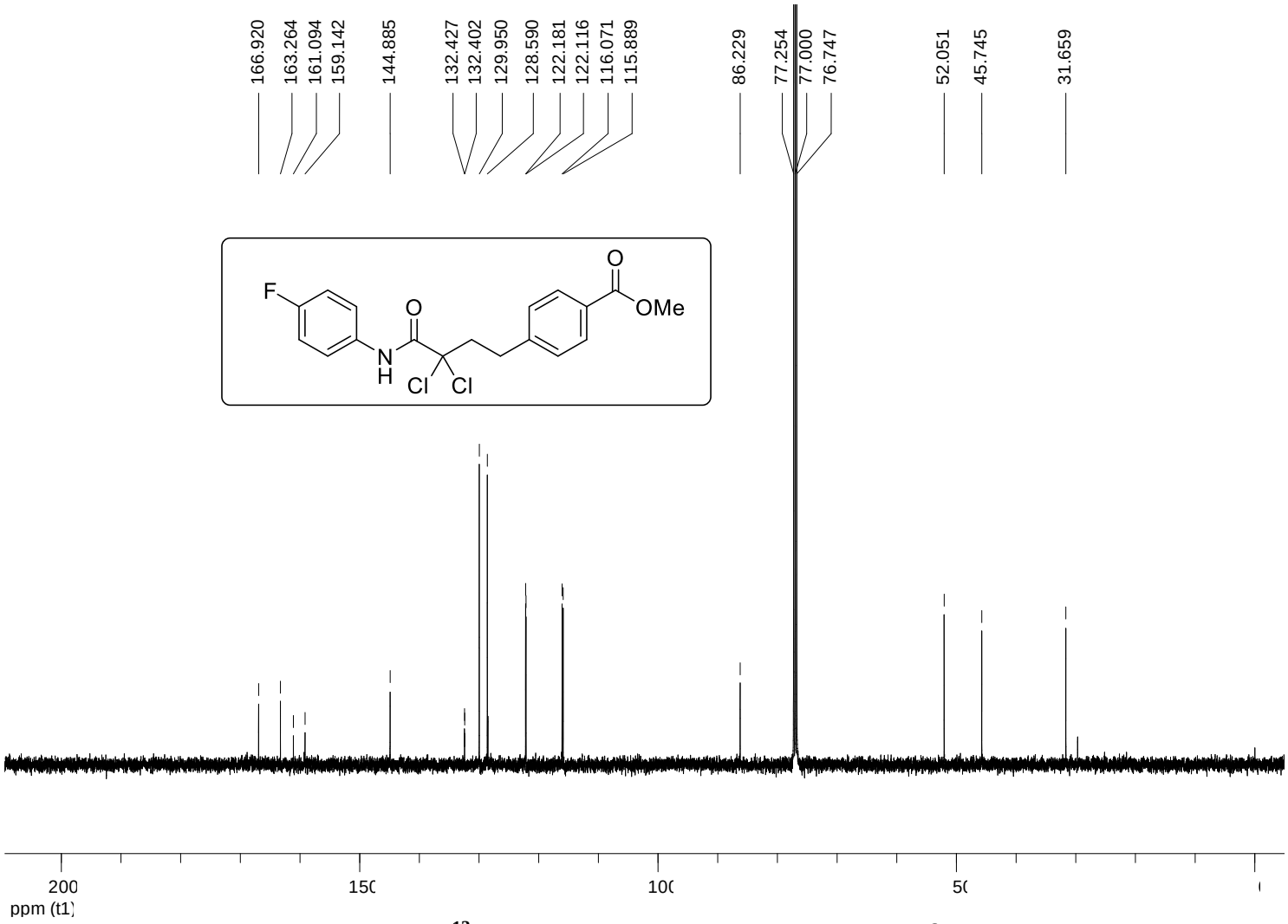


Fig. S18. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of I

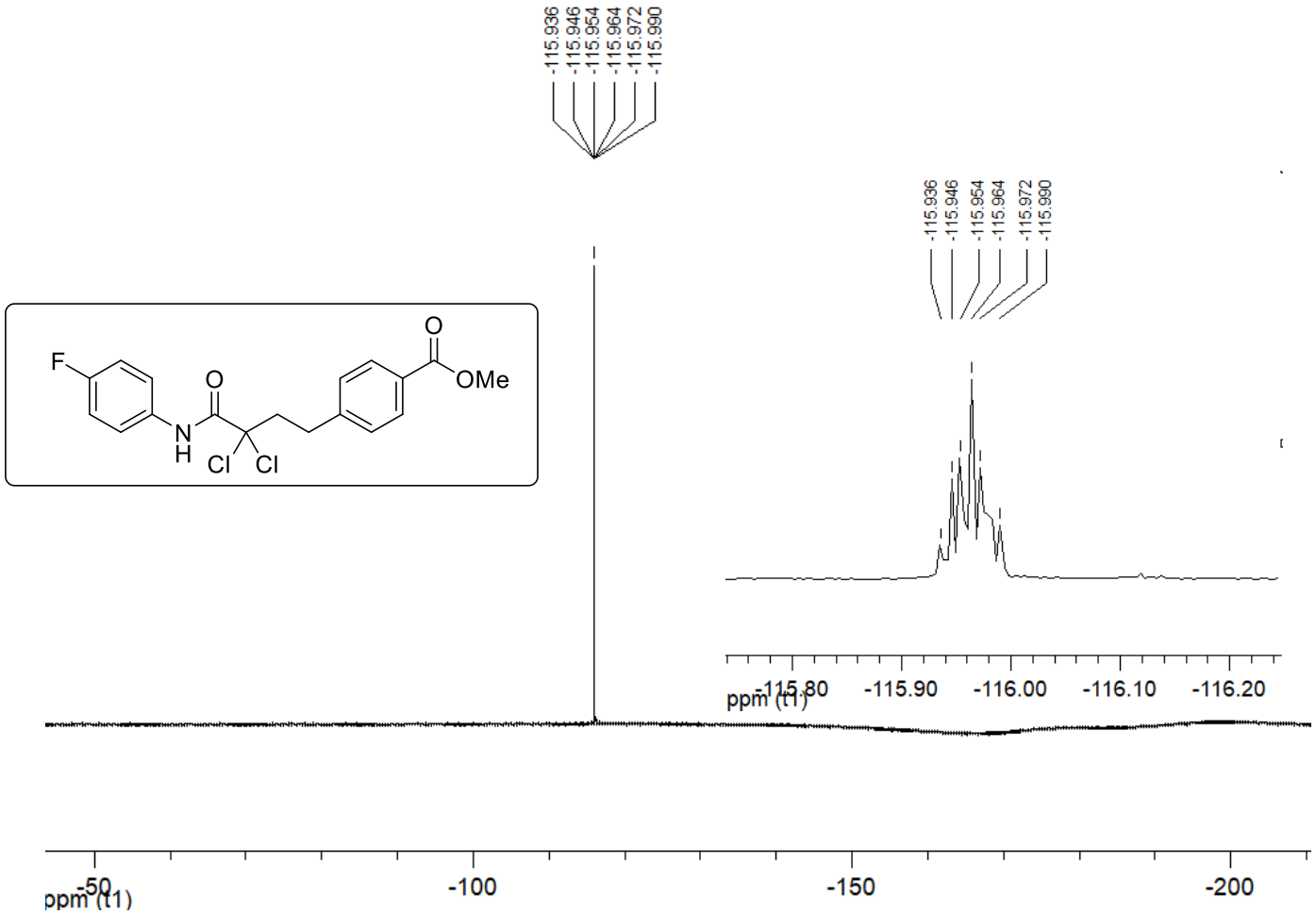


Fig. S19. ^{19}F NMR (376 MHz, $CDCl_3$) spectrum of I

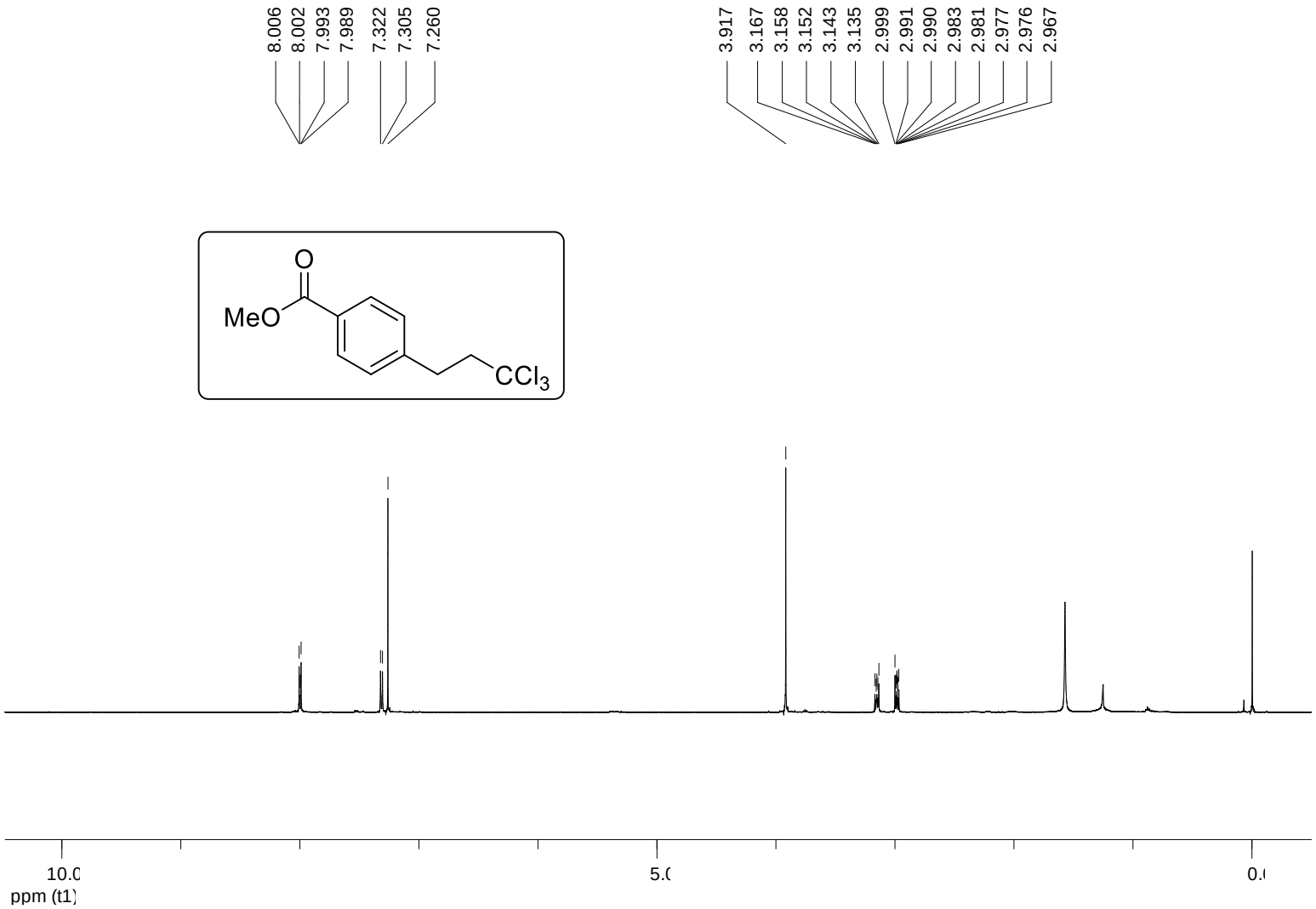


Fig. S20. ^{1}H NMR (500 MHz, $CDCl_3$) spectrum of J

5

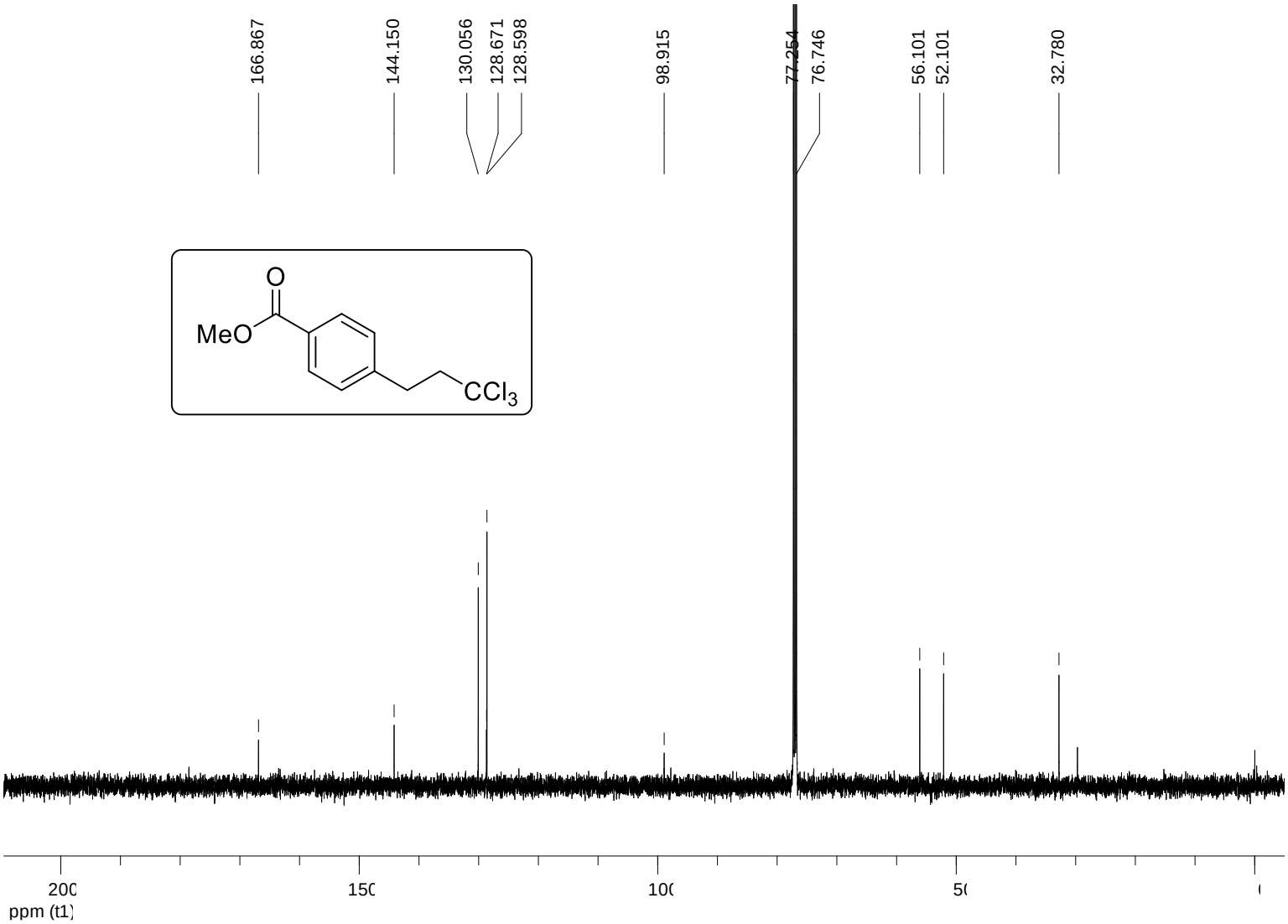


Fig. S21. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of J

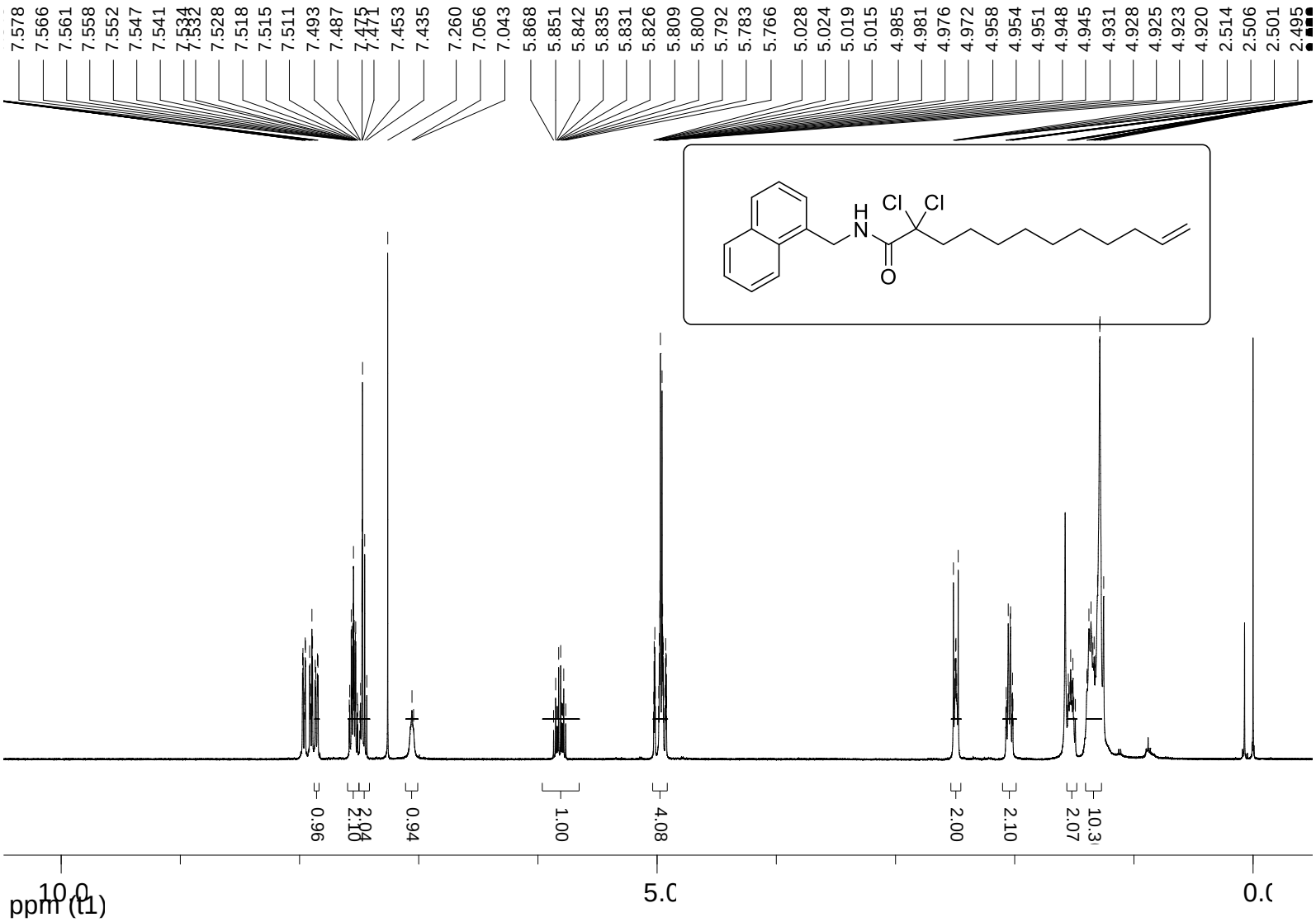


Fig. S22. ^{1}H NMR (500 MHz, $CDCl_3$) spectrum of K

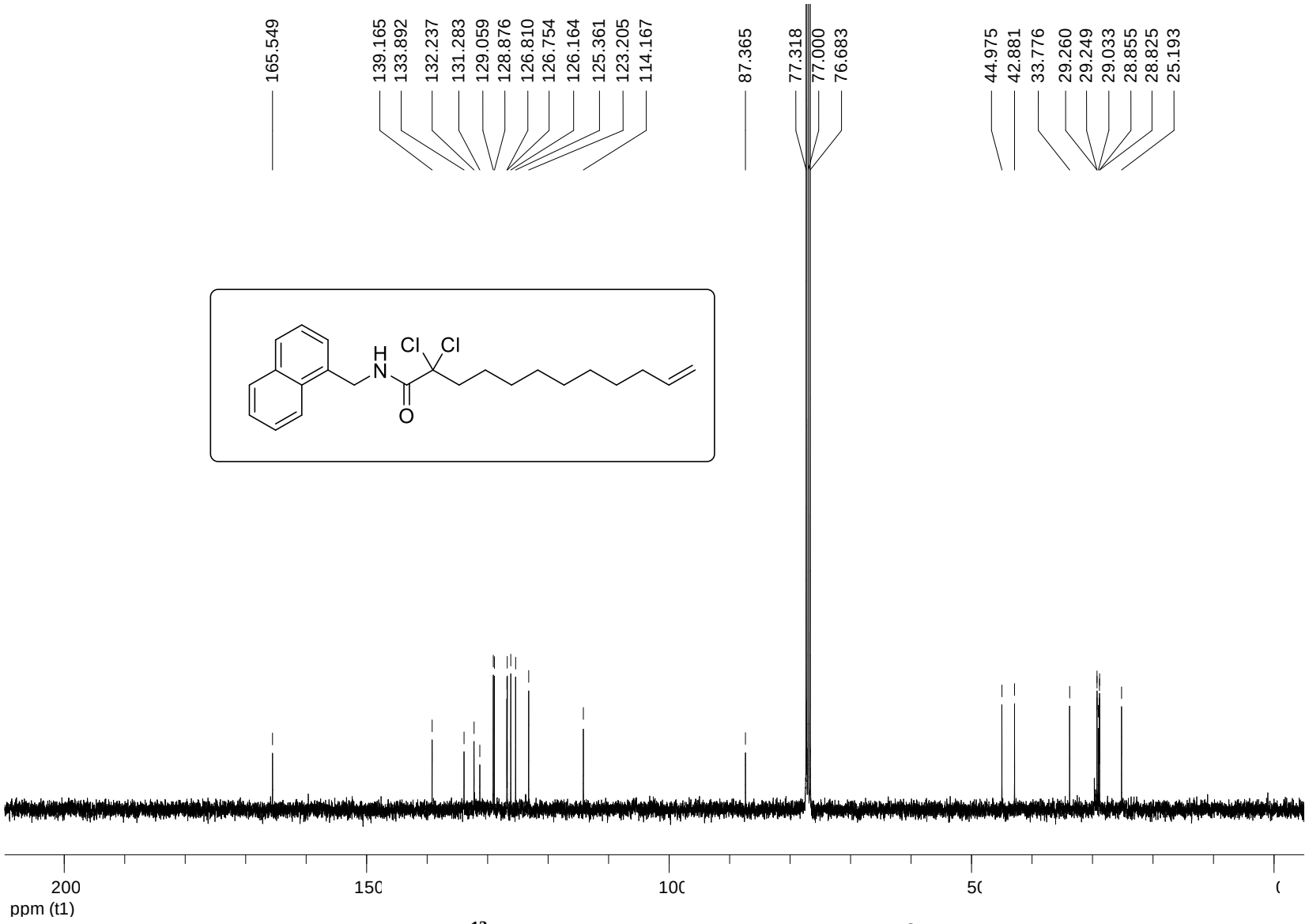


Fig. S23. ^{13}C NMR (125 MHz, $CDCl_3$) spectrum of K

3. Supplementary Tables

Table S1. Performance on the Buchwald–Hartwig coupling HTE Dataset (the best results are in bold).

Split Mode	Measure	RS-coreset	YieldBERT	YieldBERT-DA	Uncertainty-Aware
2.5/97.5	R^2 ↑	**0.656±0.033**	0.436±0.034	0.604±0.031·	0.628±0.062
	MAE(%) ↓	11.976±0.718	15.979±0.817	12.243±0.631	**11.747±1.005**
	RMSE(%) ↓	**15.971±0.805**	20.463±0.623	17.151±0.677	16.586±1.364
5/95	R^2 ↑	**0.773±0.021**	0.622±0.042	0.733±0.027	0.734±0.019
	MAE(%) ↓	**9.534±0.420**	12.117±0.789	9.651±0.338	9.677±0.408
	RMSE(%) ↓	**12.989±0.599**	16.740±0.950	14.073±0.687	14.041±0.492

Table S2. Performance on the Suzuki-Miyaura Cross Coupling HTE Dataset (the best results are in bold).

Split Mode	Measure	RS-corest	YieldBERT	YieldBERT-DA	Uncertainty-Aware	DeepReac+
2.5/97.5	R^2 ↑	**0.470±0.047**	0.330±0.047	0.282±0.047	0.375±0.072	0.125±0.141
	MAE(%) ↓	**15.357±0.616**	17.666±0.496	17.587±0.690	16.705±1.090	18.568±1.668
	RMSE(%) ↓	**20.427±0.902**	22.967±0.804	23.780±0.793	22.156±1.273	24.589±2.200
5/95	R^2 ↑	**0.604±0.019**	0.430±0.040	0.491±0.034	0.567±0.014	0.379±0.063
	MAE(%) ↓	**12.885±0.279**	15.695±0.618	14.294±0.507	13.364±0.223	14.960±0.497
	RMSE(%) ↓	**17.666±0.423**	21.181±0.724	20.016±0.661	18.463±0.308	20.090±0.816

Table S3. Performance on the B-H Dataset Provided by Denmark's Group.

Product Group	Reported result		Our model	
	MAE(%)	R^2	MAE(%)	R^2
a	26.7	0.124	**17.186**	**0.691**
b	14.364	0.777	**12.550**	**0.841**
c	**25.182**	0.257	25.920	**0.455**
d	**9.083**	**0.560**	15.351	-
e	**30.889**	**0.192**	37.436	-
f	**4.222**	**0.781**	6.545	0.510
g	**12.458**	**0.715**	17.007	0.506
h	**14**	-	22.010	-
i	30.958	-	**20.154**	**0.512**
j	**19.875**	-	20.655	-
k	20.375	0.380	**15.214**	**0.620**
l	56.5	-	**7.365**	-
m	30.625	-	**28.580**	-
Average	20.241	0.381	**18.611**	**0.506**

4. Supplementary References

1. Zhao, Q. et al. Boryl Radicals Enabled a Three-Step Sequence to Assemble All-Carbon Quaternary Centers from Activated Trichloromethyl Groups. *J. Am. Chem. Soc.* (2022).