Supplementary Information (SI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2025

Investigation of Arenes and Heteroarenes Nitration supported by High-Throughput Experimentation and Machine Learning

Taline Kerackian*[a][b], Clément Wespiser[b], Matthieu Daniel[b], Eric Pasquinet[b], Eugénie Romero*[a]

[a] Département Médicaments et Technologies pour la Santé (DMTS), SCBM, Université Paris Saclay, CEA, INRAE, 91191 Gif-sur-Yvette, France, E-mail: eugenie.romero@cea.fr

[b] CEA, DAM, Le Ripault, F-37260 Monts, France

SUPPORTING INFORMATION

General information

All chemical products commercially available were purchased from Sigma-Aldrich, Acros and Fluka and used without further purification.

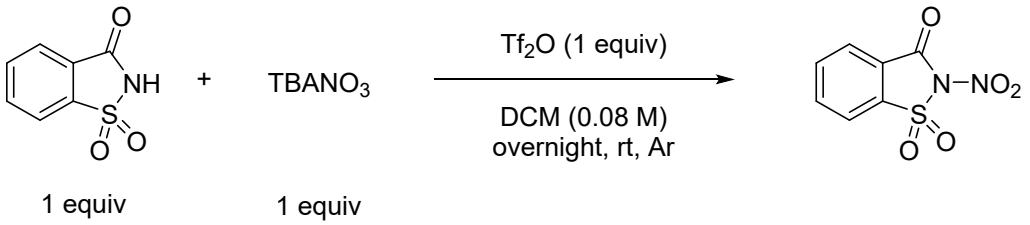
^{1}H NMR (400 MHz), ^{13}C NMR (101 MHz) were measured on a Bruker Avance 400 MHz spectrometer. Chemical shifts are reported in parts per million (ppm) downfield from residual solvents peaks and coupling constants are reported as Hertz (Hz). Splitting patterns are designated as singlet (s), broad singlet (br. s), doublet (d), triplet (t), quartet (q), quintet (quint), heptuplet (hept), multiplet (m). Splitting patterns that could not be interpreted or easily visualized are designated as multiplet (m).

Waters UPLC Acquity H-Class coupled to an Acquity diode array and to a Waters mass spectrometer SQD2 (ESI+ /ESI- source coupled to a quadripole) was used for HTE plate analysis.

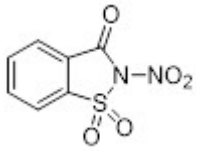
Flash chromatographies were performed using silica gel 60 Å (40-63 µm). Preparative chromatographies were performed using PLC plates of silica gel 60 F254 of 20x20 cm and 2 mm thickness. TLC analyses were carried out on pre-coated Glass TLC Silica gel plates 60 F254 from Merck. The plates were visualized using a 254 nm ultraviolet lamp.

Precision balance used: Sartorius Lab Quintix 125D-15 3

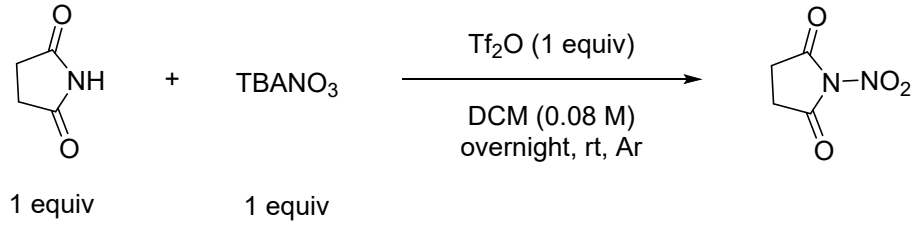
Preparation of nitration reagents



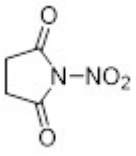
Procedure: A 250 mL round bottom flask equipped with a magnetic stir bar was charged with tetrabutylammonium nitrate (1 equiv, 10 mmol, 3.04 g) in DCM (80 mL), the vial was put at 0 °C under inert atmosphere. Trifluoromethanesulfonic anhydride (1 equiv, 10 mmol, 1.68 mL) was slowly added via syringe and the mixture was left to stir at 0 °C for 1 h. Then, saccharin (1 equiv, 10 mmol, 1.84 mg) was added and the reaction mixture was stirred at room temperature overnight. After which the solution is slightly yellow and a white precipitate is formed. The mixture was evaporated to give a white solid with a liquid. Flash chromatography was performed (120 g prepacked column, from 100:0 cyclohexane/ethyl acetate to 70/30 cyclohexane/ethyl acetate) to give the expected compound as a white solid (1.483 g, 65% yield).



N-Nitrosaccharin: Known compound [1] ^{1}H NMR (400 MHz, CD_3CN): δ 8.23-8.18 (m, 2H), 8.14 (dt, J = 7.2, 0.8 Hz, 1H), 8.06 (dt, J = 7.6, 1.2 Hz, 1H).



Procedure: A 250 mL round bottom flask equipped with a magnetic stir bar was charged with tetrabutylammonium nitrate (1 equiv, 10 mmol, 3.04 g) in DCM (80 mL), the vial was put at 0 °C under inert atmosphere. Trifluoromethanesulfonic anhydride (1 equiv, 10 mmol, 1.68 mL) was slowly added via syringe and the mixture was left to stir at 0 °C for 1 h. Then, succinimide (1 equiv, 10 mmol, 991 mg) was added and the reaction mixture was stirred at room temperature overnight. Flash chromatography was performed (120 g prepacked column, from 100:0 cyclohexane/ethyl acetate to 65/35 cyclohexane/ethyl acetate) to give the expected compound as a white solid (0.401 g, 28% yield).



N-Nitrosaccharin: Known compound [1] ^{1}H NMR (400 MHz, $CDCl_3$): δ 1H NMR (300 MHz, CDCl3) δ 2.93 (s, 4H).

Procedures for HTE experiments

General set-up information

The screening of conditions and scope evaluations were performed in a 96-well plate format using 1 mL vials (Figure S1.a., 8x30 mm vials number 884001 from Analytical Sales and Services) in a Paradox reactor (Figure S1.b., 96973 from Analytical Sales and Services). The homogeneous stirring was controlled thanks to stainless steel, parylene C coated stirring elements and a tumble stirrer (Figure S1.c and d., VP 711D-1 and VP 710 Series from V&P Scientific). The liquid dispensing was performed using calibrated manual pipettes and multipipettes (Fisher/Eppendorf). The heating control was performed thanks to a heating block from V&P scientific (Figure S1.d., VP 741DCE). The HTE experiments were designed using an in-house software called HTDesign®, and developed by the GIPSI team at CEA Paris-Saclay (DRF/Joliot).

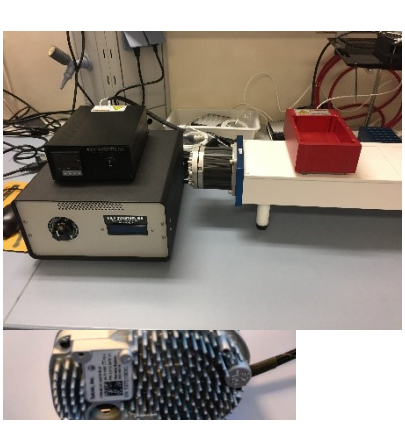


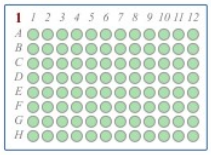
Figure S1: HTE set-up description

HTDesign® protocol

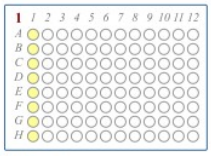
Solutions Mères

1	Solution «Naphtalene»:	ajouter **140,987 mg**		de **Naphtalene** dans	**11,00 ml** de solvant **acetone**
2	Solution «Bismuth(III) nitrate pentahydrate»:	ajouter **194,028 mg**		de **Bismuth(III) nitrate pentahydrate** dans	**2,00 ml** de solvant **acetone**
3	Solution «Silver nitrate»:	ajouter **67,948 mg**		de **Silver nitrate** dans	**2,00 ml** de solvant **water**
4	Solution «Nitronium tetrafluoroborate»:	ajouter **53,124 mg**		de **Nitronium tetrafluoroborate** dans	**2,00 ml** de solvant **acetone**
5	Solution « Acide nitrique 70%»:	ajouter **10,082 mg**	/ **7,13 µl**	de **Acide nitrique 70%** dans	**0,80 ml** de solvant **acetone**
6	Solution «Sodium nitrite»:	ajouter **27,600 mg**		de **Sodium nitrite** dans	**2,00 ml** de solvant **water**
7	Solution «Sodium nitrate»:	ajouter **33,996 mg**		de **Sodium nitrate** dans	**2,00 ml** de solvant **water**
8	Solution «Iron(III) nitrate nonahydrate»:	ajouter **161,600 mg**		de **Iron(III) nitrate nonahydrate** dans	**2,00 ml** de solvant **acetone**
9	Solution «Potassium nitrite»:	ajouter **34,040 mg**		de **Potassium nitrite** dans	**2,00 ml** de solvant **water**
10	Solution «Tetrabutylammonium nitrate»:	ajouter **121,788 mg**		de **Tetrabutylammonium nitrate** dans	**2,00 ml** de solvant **methylene chloride**
11	Solution «1-nitropyrrolidine-2,5-dione»:	ajouter **57,636 mg**		de **1-nitropyrrolidine-2,5-dione** dans	**2,00 ml** de solvant **methylene chloride**
12	Solution «2-nitrobenzo[d]isothiazol-3(2H)-one 1,1-dioxide»:	ajouter **91,272 mg**		de **2-nitrobenzo[d]isothiazol-3(2H)-one 1,1-dioxide** dans	**2,00 ml** de solvant **methylene chloride**
13	Solution «tert-butyl nitrite»:	ajouter **16,499 mg**	/ **19,03 µl**	de **tert-butyl nitrite** dans	**0,80 ml** de solvant **acetone**
14	Solution «Potassium persulfate»:	ajouter **162,192 mg**		de **Potassium persulfate** dans	**2,00 ml** de solvant **water**
15	Solution «Silver carbonate»:	ajouter **55,150 mg**		de **Silver carbonate** dans	**2,00 ml** de solvant **water**
16	Solution «2,2′-Azobis(2-methylpropionitrile)»:	ajouter **0,657 mg**		de **2,2′-Azobis(2-methylpropionitrile)** dans	**2,00 ml** de solvant **acetone**
17	Solution «Copper(II) trifluoromethanesulfonate»:	ajouter **18,084 mg**		de **Copper(II) trifluoromethanesulfonate** dans	**2,00 ml** de solvant **acetone**
18	Solution «Tris(dibenzylidèneacétone)dipalladium(0)»:	ajouter **1,831 mg**		de **Tris(dibenzylidèneacétone)dipalladium(0)** dans	**2,00 ml** de solvant **acetone**
19	Solution «tBuBrettPhos»:	ajouter **1,939 mg**		de **tBuBrettPhos** dans	**2,00 ml** de solvant **acetone**
20	Solution «Copper iodide»:	ajouter **5,714 mg**		de **Copper iodide** dans	**2,00 ml** de solvant **acetone**
21	Solution «N,N'-dimethylethylenediamine »:	ajouter **5,289 mg**	/ **6,46 µl**	de **N,N'-dimethylethylenediamine** dans	**2,00 ml** de solvant **acetone**
22	Solution «Magnesium perchlorate hexahydrate»:	ajouter **133,926 mg**		de **Magnesium perchlorate hexahydrate** dans	**2,00 ml** de solvant **water**
23	Solution «Acetonitrile»:	**SOLVANT**			

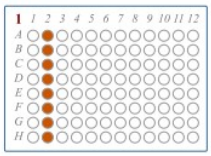
Etape 1: ajouter **100,00 µl** de la solution **«Naphtalene»** dans les puits:



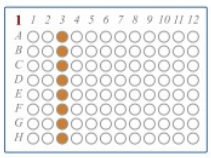
Etape 2: ajouter **100,00 µl** de la solution **«Bismuth(III) nitrate pentahydrate»** dans les puits:



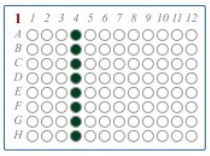
Etape 3: ajouter **100,00 µl** de la solution **«Silver nitrate»** dans les puits:



Etape 4: ajouter **100,00 µl** de la solution **«Nitronium tetrafluoroborate»** dans les puits:

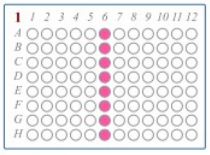


Etape 5: ajouter **100,00 µl** de la solution **« Acide nitrique 70%»** dans les puits:

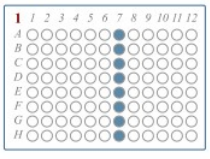


Etape 6: ajouter **100,00 µl** de la solution **«Sodium nitrite»** dans les puits:

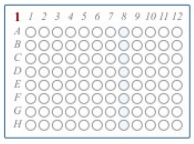
Etape 7: ajouter **100,00 µl** de la solution **«Sodium nitrate»** dans les puits:



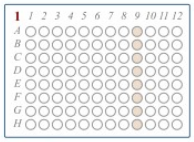
Etape 8: ajouter **100,00 µl** de la solution **«Iron(III) nitrate nonahydrate»** dans les puits:



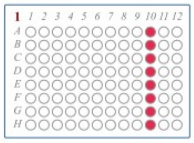
Etape 9: ajouter **100,00 µl** de la solution **«Potassium nitrite»** dans les puits:



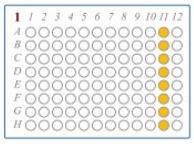
Etape 10: ajouter **100,00 µl** de la solution **«Tetrabutylammonium nitrate»** dans les puits:



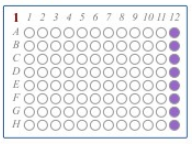
Etape 11: ajouter **100,00 µl** de la solution **«1-nitropyrrolidine-2,5-dione»** dans les puits:



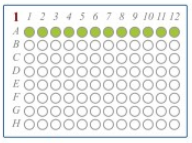
Etape 12: ajouter **100,00 µl** de la solution **«2-nitrobenzo[d]isothiazol-3(2H)-one 1,1-dioxide»** dans les puits:



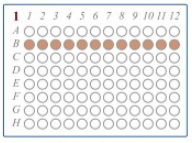
Etape 13: ajouter **100,00 µl** de la solution **«tert-butyl nitrite»** dans les puits:



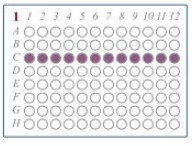
Etape 14: ajouter **100,00 µl** de la solution **«Potassium persulfate»** dans les puits:



Etape 15: ajouter **100,00 µl** de la solution **«Silver carbonate»** dans les puits:



Etape 16: ajouter **100,00 µl** de la solution **«2,2′-Azobis(2-methylpropionitrile)»** dans les puits:



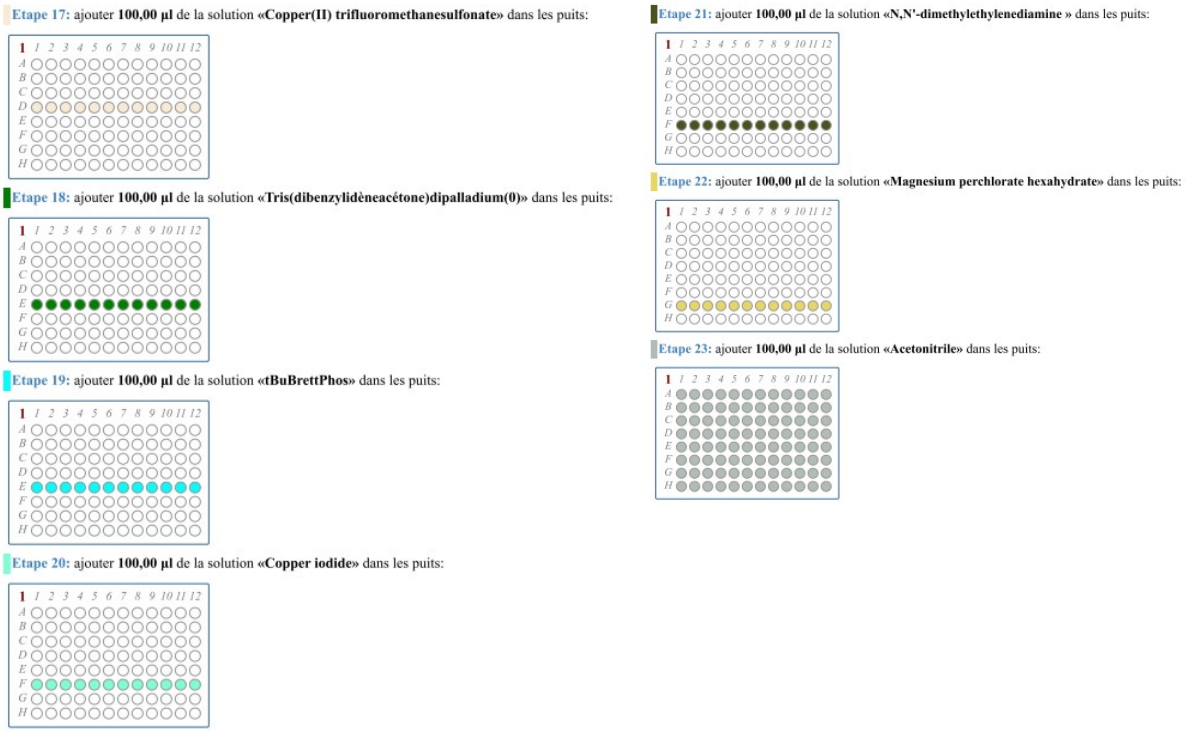


Figure S2: Example of HTE plate design to study naphthalene nitration conditions

Plate preparation procedure

Solution numbers correspond to the solution number assigned by HTDesign®, which can be found above.

Day before the reaction:

- S2, S3, S4, S6, S7, S8, S9, S10, dispensed in the plate (liquid dispensing), then evaporated using nitrogen blowing system (Equavap Analytical Sales and Services)
- S14, S15, S17, S22 dispensed in the plate (liquid dispensing), then evaporated

On the day of the reaction:

- S16, S18, S19, S20, S21 dispensed in the plate (liquid dispensing), then evaporated
- S11, S12 dispensed in the plate (liquid dispensing), then evaporated
- **If solid** S1 dispensed in the plate (liquid dispensing), then evaporated
- Addition of stirrers
- Transfer into Paradox reactor
- Addition of solvent
- S5, S13 dispensed directly with micropipette
- **If liquid** S1 dispensed directly with micropipette
- Plate sealed and placed under stirring at 100°C for 24 hours (300 rpm).

General observations: No issues were encountered with the preparation of HTE. A special attention was paid to the solubility of compounds prepared as stock solutions. Most of them were soluble in the most adequate solvent chosen (either acetone, DCM or water). When acceptable, some compounds could be dispensed as slurry solutions. No impact of this dispensing methodology was noted as it is a widely used methodology for HTE plate preparation and a validated method in-house. The HTE plate has to be well sealed so that the solvent doesn't dry out during reaction. It was checked that no solvent loss had happened during reaction.

HTE analytical method

After the reaction each sample was diluted with a MeCN solution containing 1 µmol of biphenyl (500 µL, 0.002 M). The vials were centrifuged. 50 µL aliquots of the supernatant were sampled into a 1 mL deep 96 well plate containing 500 µL MeCN.

Ratios of Area Under Curve of starting material (when relevant), products and side products were tabulated as well as areas for peaks of interest.

Analytical assay

Assay with Waters UPLC-UV-MS:

Plate n°1 (Naphthalene): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 80% H_2O to 20% in 7.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°2 (1-Naphthoic acid): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 80% H_2O to 20% in 7.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°3 (1-Bromonaphthalene): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 70% H_2O to 30% in 7.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°4 (Pyridine): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 95% H_2O to 0% in 4.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°5 (Picolinic acid): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 95% H_2O to 0% in 4.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°6 (2-Bromopyridine): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 95% H_2O to 0% in 4.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°7 (2,3-Benzofuran): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 70% H_2O to 50% in 9.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°8 (Benzofuran-2-carboxylic acid): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 95% H_2O to 0% in 4.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Plate n°9 (2-Bromobenzofuran): C18 50*2,1 in 1,7 µm; 4 µL injection; 50 mm x 2.1 mm; 0.4 mL/min; H_2O/MeCN; gradient: 80% H_2O to 20% in 7.00 min; QDA ESI 110-1200 pos/neg; nebulizer: 700 L/hr; cone gas: 30 L/hr; source 150 °C; desolvation 450 °C.

Retention times

Plate n°1 (Naphthalene):

Peak Label	Retention Time (in min)
Starting Material (SM)	2.7
Internal Standard (IS)	4.9
Final Product (FP)	3.9

Plate n°2 (1-Naphthoic acid):

Peak Label	Retention Time (in min)
Starting Material (SM)	4.2
Internal Standard (IS)	4.8
Final Product (FP)	3.8

Plate n°3 (1-Bromonaphthalene):

Peak Label	Retention Time (in min)
Starting Material (SM)	5.2
Internal Standard (IS)	4.6
Final Product (FP)	3.3

Plate n°4 (Pyridine):

Peak Label	Retention Time (in min)
Starting Material (SM)	0.5[a]
Internal Standard (IS)	1.1
Final Product (FP)	2.1

[a] not quantified because the signal is located in the injection peak

Plate n°5 (Picolinic acid):

Peak Label	Retention Time (in min)
Starting Material (SM)	0.5[a]
Internal Standard (IS)	3.2
Final Product (FP)	1.7

[a] not quantified because the signal is located in the injection peak

Plate n°6 (2-Bromopyridine):

Peak Label	Retention Time (in min)
Starting Material (SM)	2.0
Internal Standard (IS)	3.1
Final Product (FP)	1.5

Plate n°7 (2,3-Benzofuran):

Peak Label	Retention Time (in min)
Starting Material (SM)	3.0
Internal Standard (IS)	6.6
Final Product (FP)	2.6

Plate n°8 (Benzofuran-2-carboxylic acid):

Peak Label	Retention Time (in min)
Starting Material (SM)	2.2
Internal Standard (IS)	3.2
Final Product (FP)	2.8

Plate n°9 (2-Bromobenzofuran):

Peak Label	Retention Time (in min)
Starting Material (SM)	4.3
Internal Standard (IS)	4.6
Final Product (FP)	3.0

TWC Chromatogram example

Plate n°1 (Naphthalene): vial H7

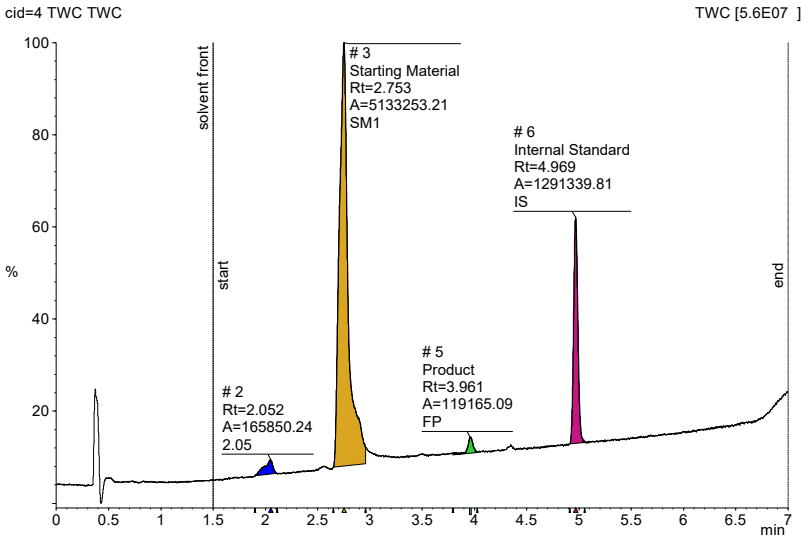


Plate n°2 (1-Naphthoic acid): vial D11

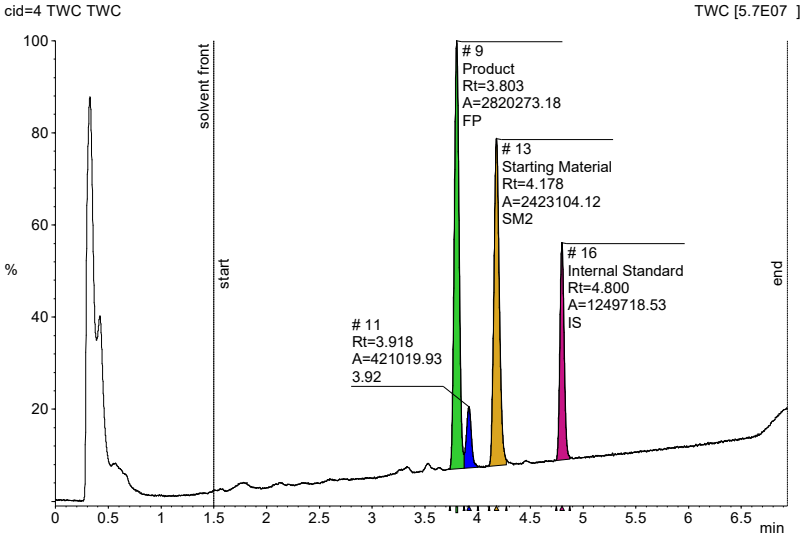


Plate n°3 (1-Bromonaphthalene): vial A2

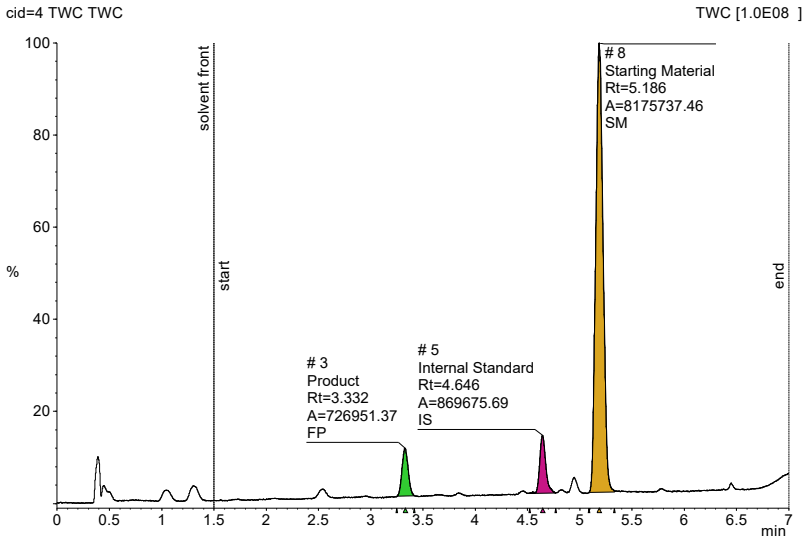


Plate n°4 (Pyridine): vial G11

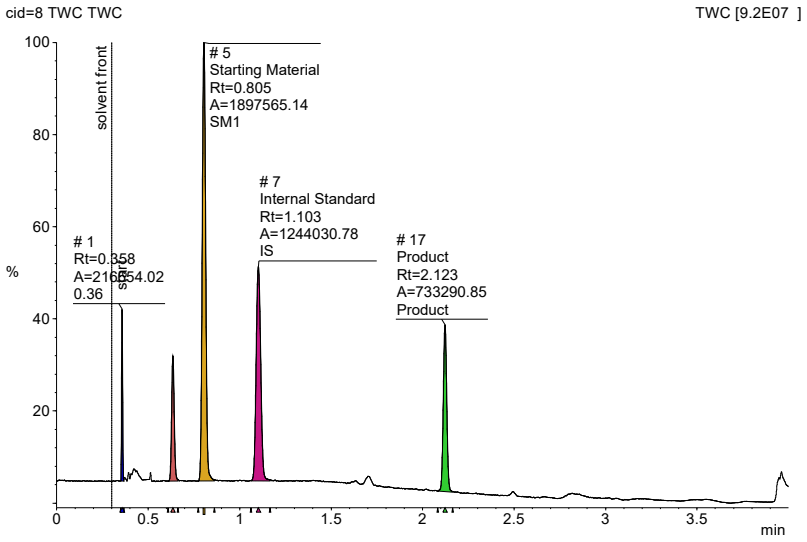


Plate n°5 (Picolinic acid): vial C4

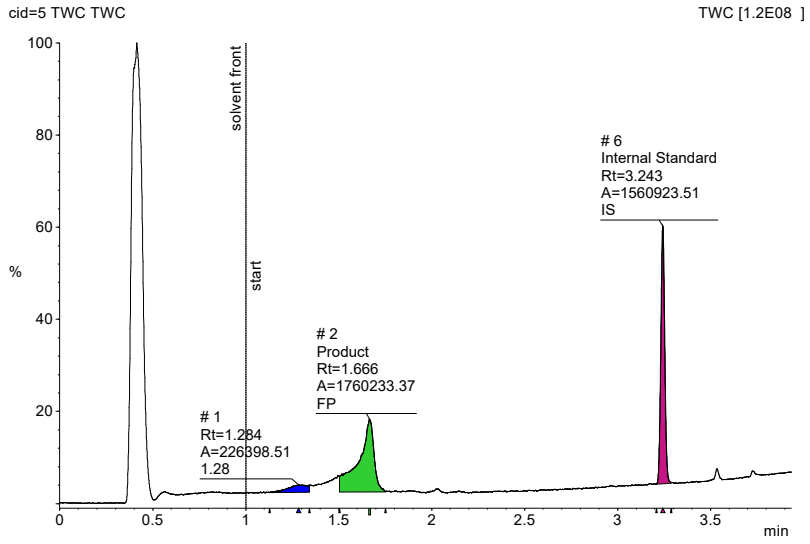


Plate n°6 (2-Bromopyridine): vial C7

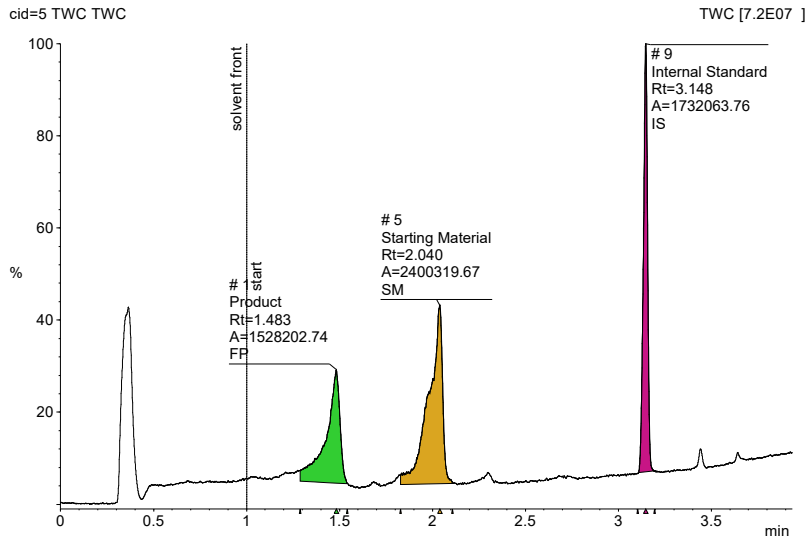


Plate n°7 (2,3-Benzofuran): vial H11

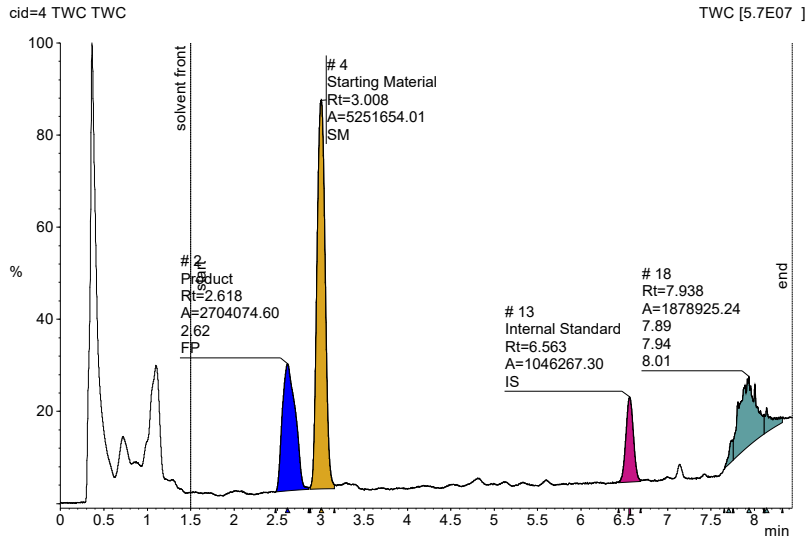


Plate n°8 (Benzofuran-2-carboxylic acid): vial B2

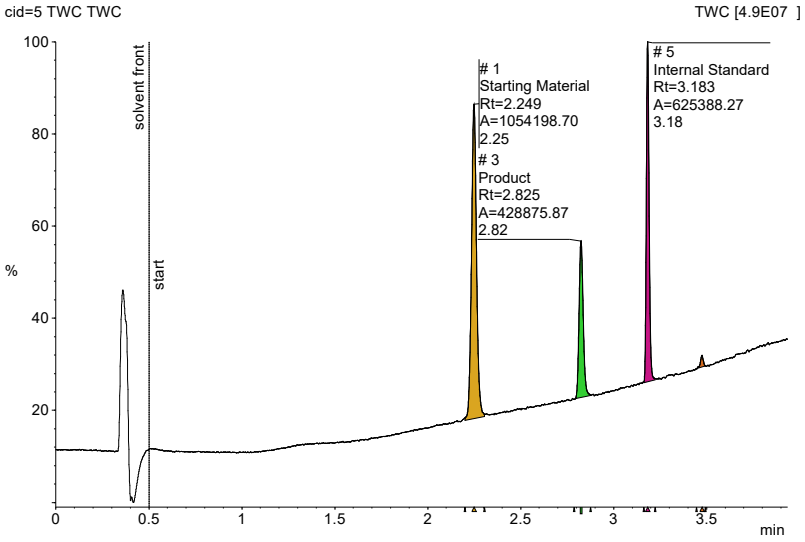
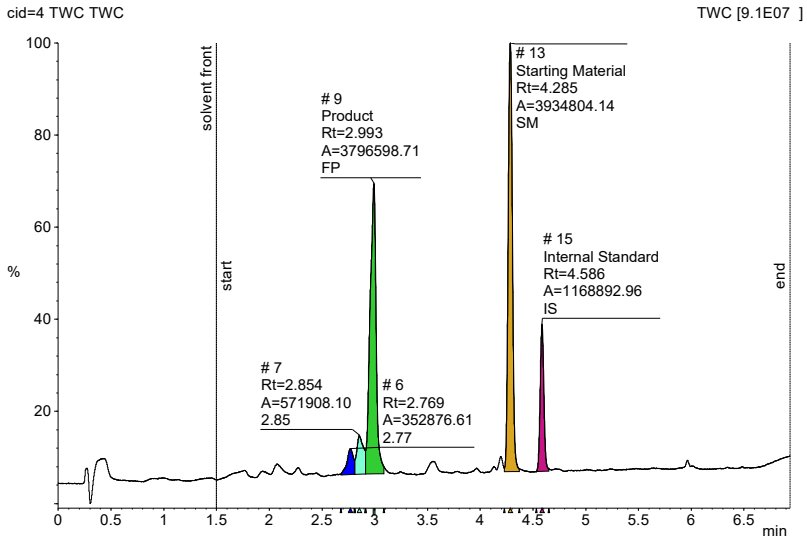


Plate n°9 (2-Bromobenzofuran): vial A9



Results

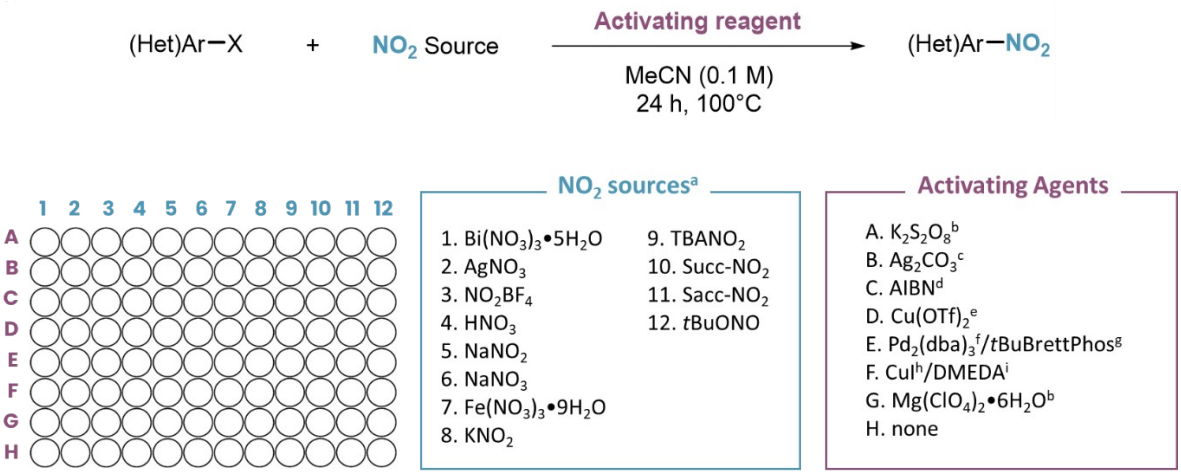


Figure S3: Reminder of the general plate design for the HTE campaign. [a] 2 equivalents, [b] 3 equivalents, [c] 1 equivalents, [d] 2 mol%, [e] 25 mol%, [f] 0.5 mol%, [g] 1.2 mol%, [h] 15 mol%, [i] 30 mol%.

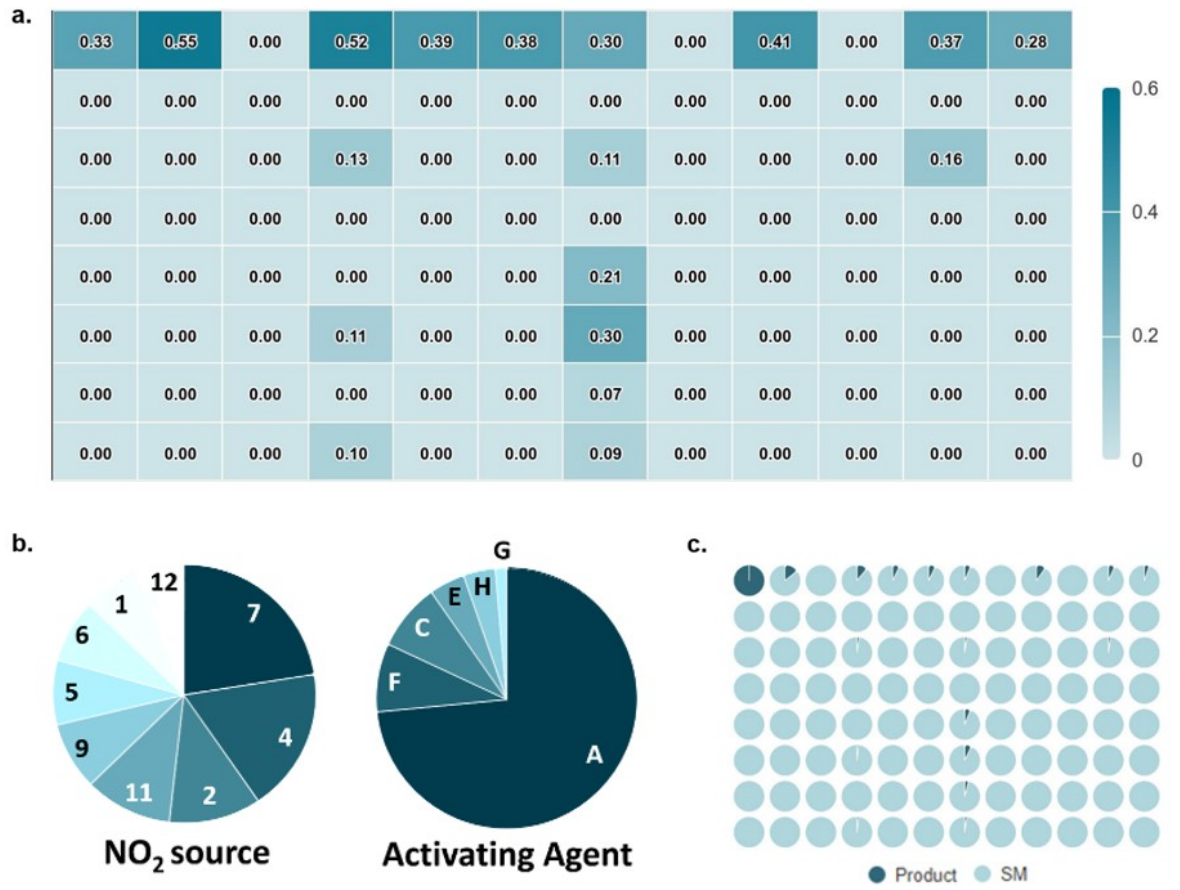


Figure S4: Results obtained for the HTE campaign on naphthalene: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

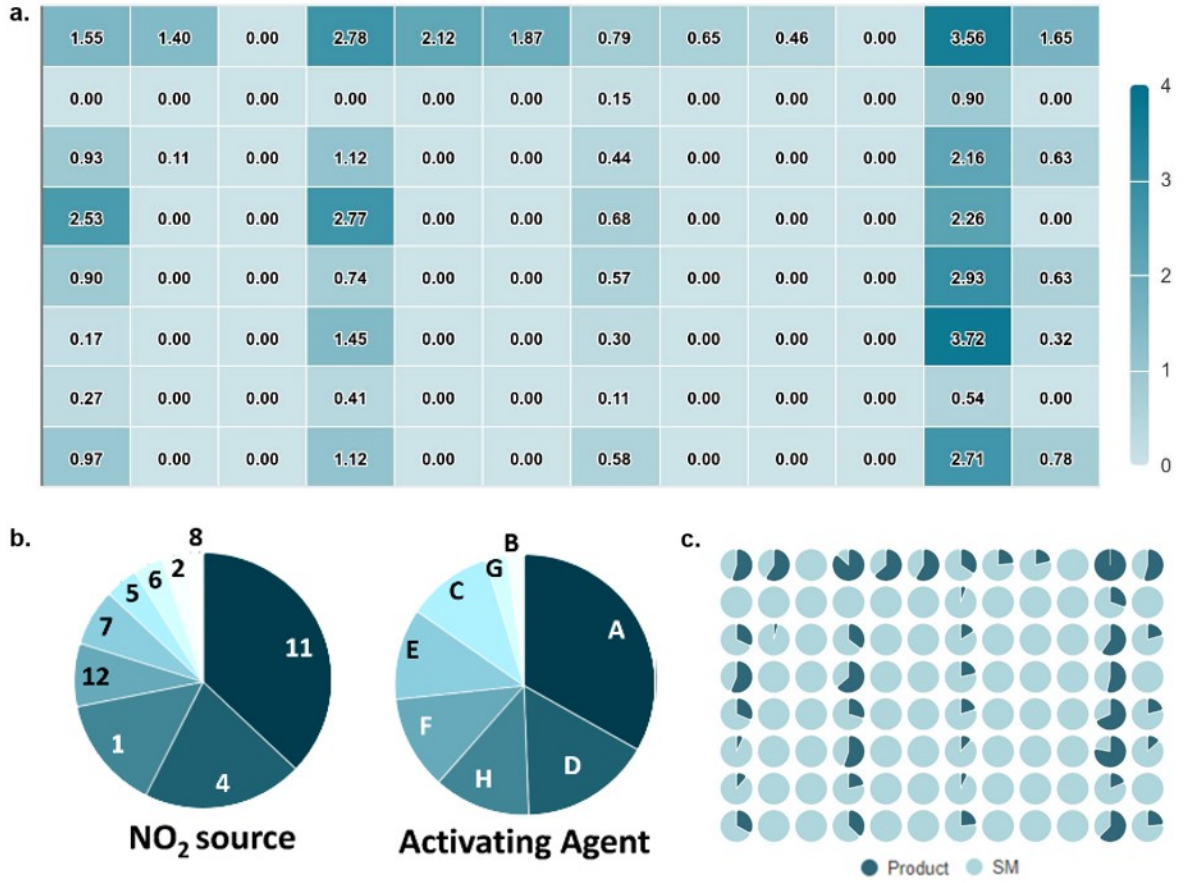


Figure S5: Results obtained for the HTE campaign on 1-naphthoic acid: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

Figure S6: Results obtained for the HTE campaign on 1-bromonaphthalene: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

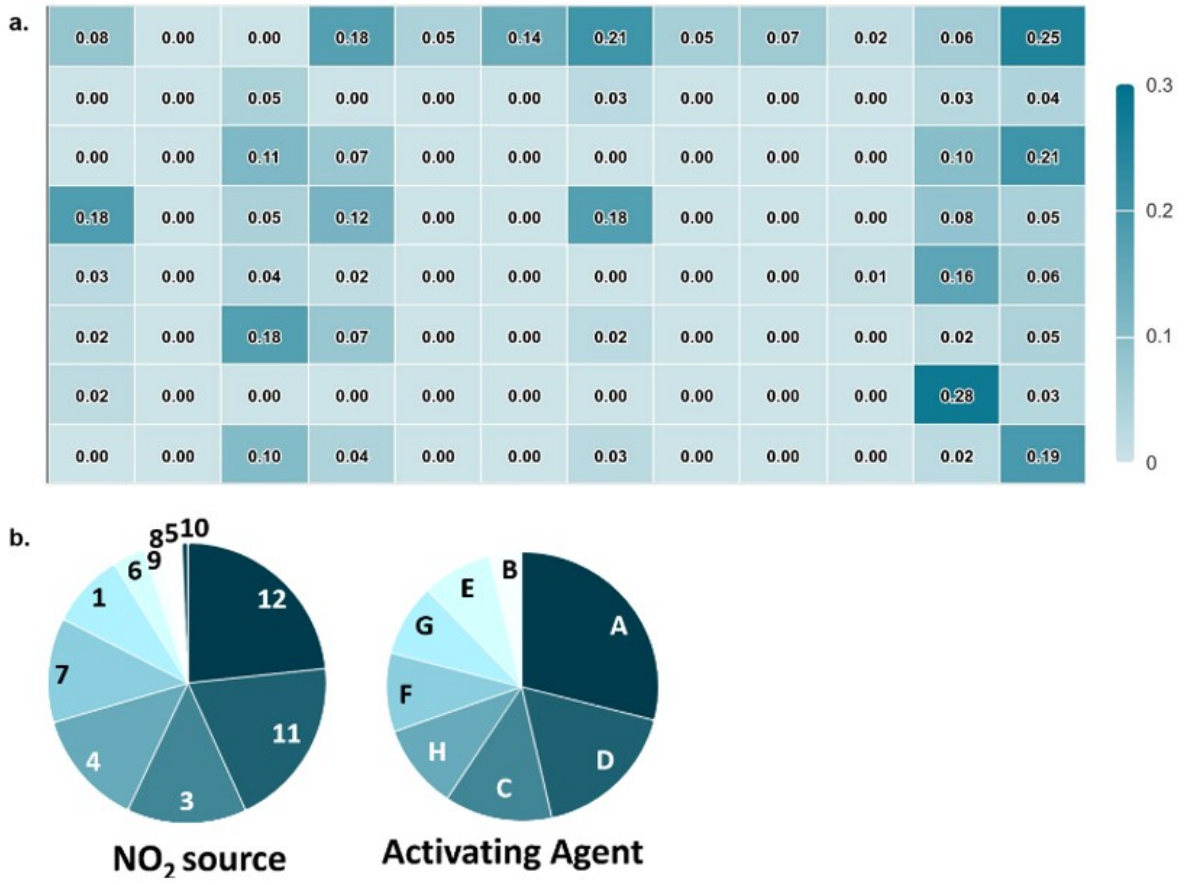


Figure S7: Results obtained for the HTE campaign on pyridine: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances.

Figure S8: Results obtained for the HTE campaign on picolinic acid: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances.

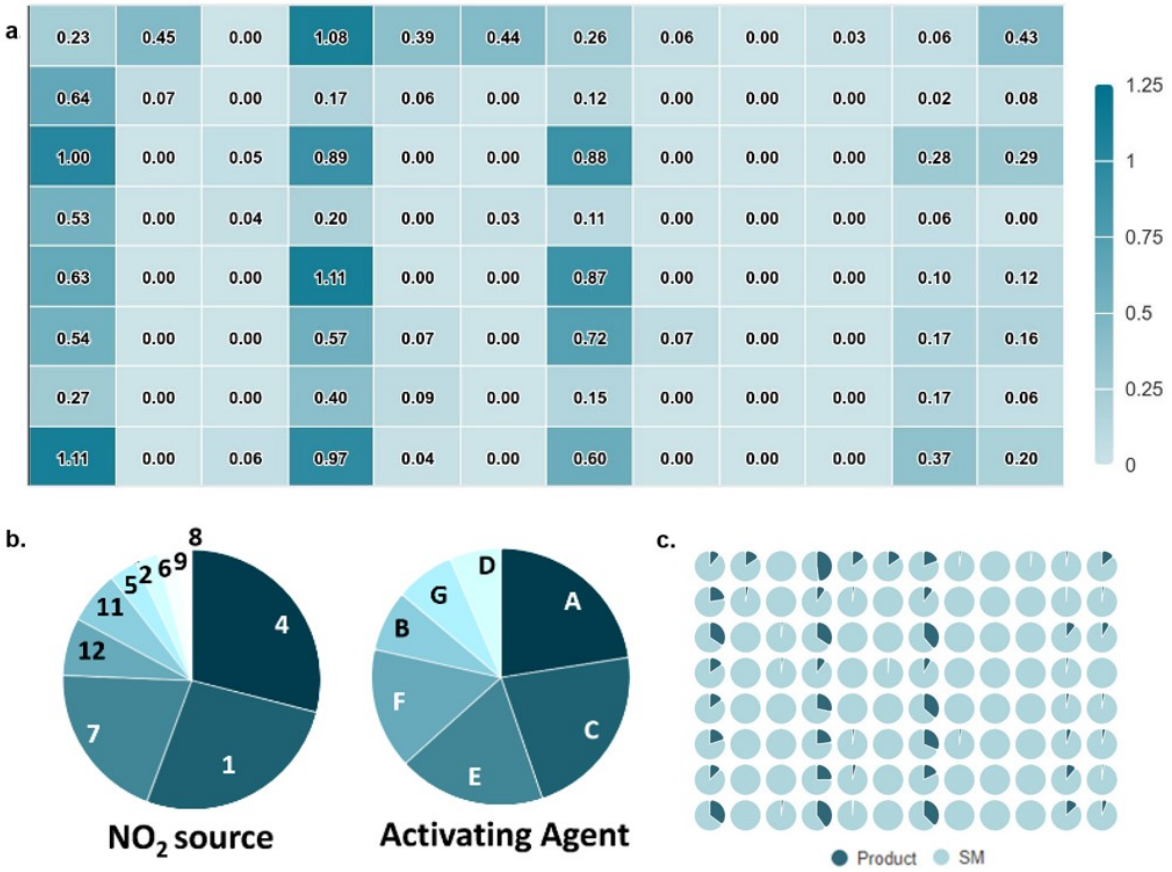


Figure S9: Results obtained for the HTE campaign on 2-bromopyridine: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

Figure S10: Results obtained for the HTE campaign on 2,3-benzofuran: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

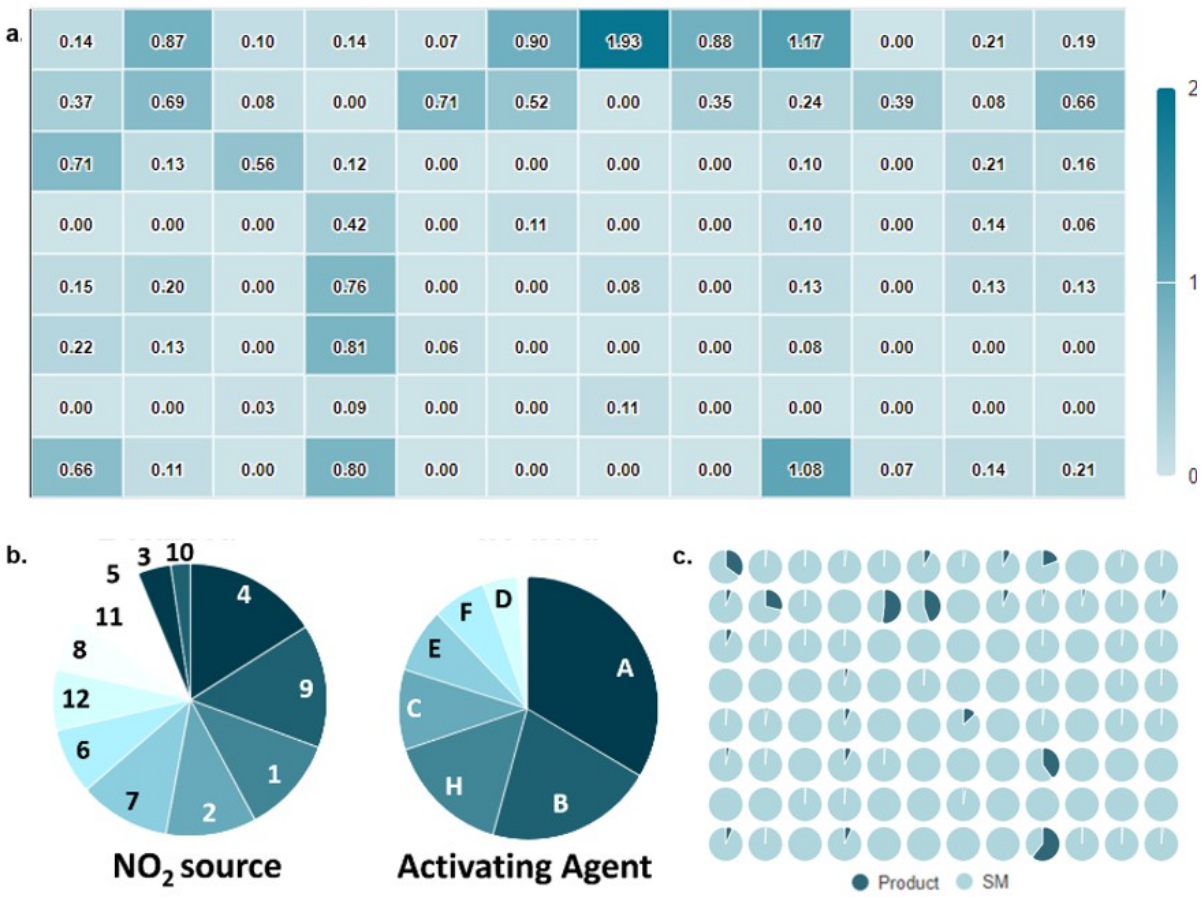


Figure S11: Results obtained for the HTE campaign on benzofuran-2-carboxylic acid: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

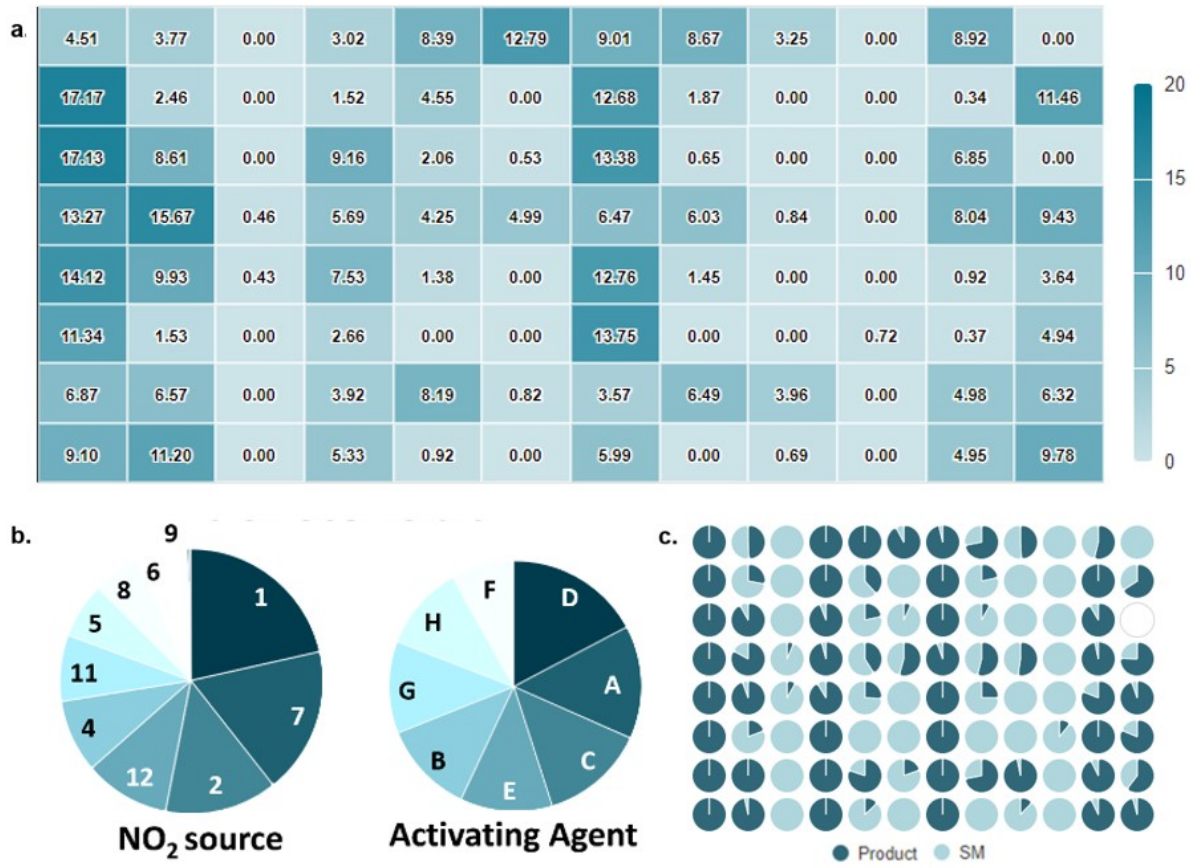


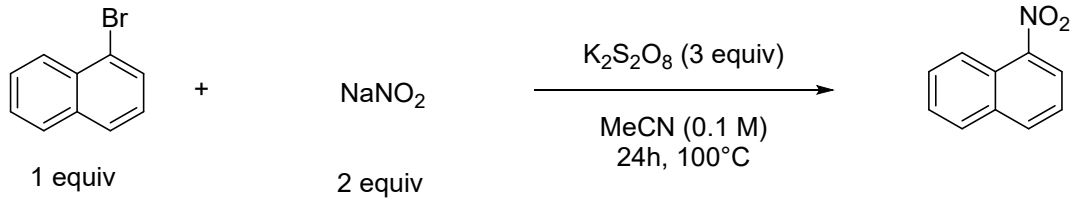
Figure S12: Results obtained for the HTE campaign on 2-bromobenzofuran: a. Heat map plate results; b. Pie chart representation of nitrating and activating agents' performances; c. Pie chart representation of product and starting material presence.

Isolation of nitrated compounds

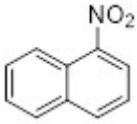
General Procedure for the isolation of nitrated compounds

A 20 mL Schlenck tube equipped with a magnetic stirbar was charged with nitrating agent (2 equiv, 1 mmol), activating agent and selected substrate (1 equiv, 0.5 mmol) in MeCN (5 mL). The vial was sealed (under air atmosphere) and the reaction mixture was stirred at 100°C for 24h. The mixture was then analyzed by GC-MS and purified by flash chromatography.

Isolation of 1-nitronaphthalene

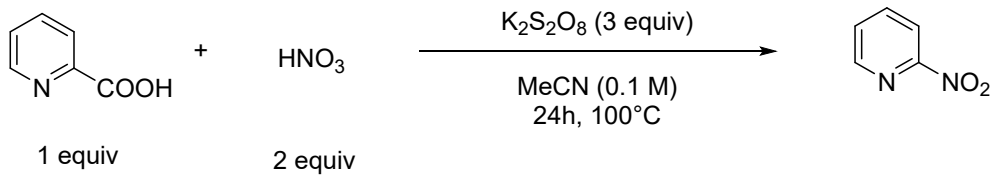


Procedure: A 20 mL Schlenck tube equipped with a magnetic stir bar was charged with sodium nitrite (2 equiv; 1 mmol; 69 mg), potassium persulfate (3 equiv, 1.5 mmol, 405 mg) and 1-bromonaphtalene (1 equiv; 0.5 mmol, 70 µL) in MeCN (5 mL). The vial was sealed (under air atmosphere) and the reaction mixture was stirred at 100°C for 24 h. The mixture was then analyzed by GC-MS.

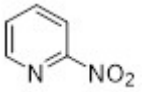


1-nitronaphtalene: Known compound [2] prepared according to general conditions and purified by flash chromatography (cyclohexane/ethyl acetate, gradient 100:0 to 95:5) to give an orange solid (0.029 g; 36%). **^{1}H NMR** (400 MHz, $CDCl_3$) δ 8.57 (d, *J* = 8.8 Hz, 1H), 8.24 (d, *J* = 7.6 Hz, 1H), 8.13 (d, *J* = 8.4 Hz, 1H), 7.97 (d, *J* = 8.0 Hz, 16 1H), 7.73 (t, *J* = 7.2 Hz, 1H), 7.63 (t, *J* = 7.6 Hz, 1H), 7.55 (t, *J* = 8.0 Hz, 1H); **^{13}C NMR** (101 MHz, $CDCl_3$) δ 146.7, 134.8, 134.5, 129.6, 128.7, 127.5, 125.2, 124.3, 124.1, 123.2.

Isolation of 2-nitropyridine

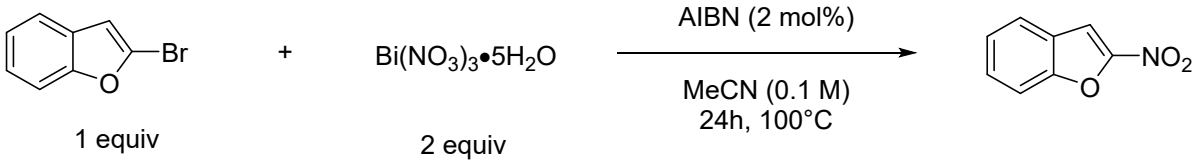


Procedure: A 20 mL Schlenck tube equipped with a magnetic stir bar was charged with potassium persulfate (3 equiv, 1.5 mmol, 405 mg), picolinic acid (1 equiv, 0.5 mmol, 62 mg), and nitric acid 69% (2 equiv, 1 mmol, 45 µL) in MeCN (5 mL). The vial was sealed (under air atmosphere) and the reaction mixture was stirred at 100°C for 24 h. The mixture was then analyzed by GC-MS.

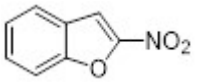


2-nitropyridine: Known compound [3] prepared according to general conditions and purified by flash chromatography (cyclohexane/ethyl acetate, gradient 100:0 to 7:3) to give an orange solid (0.005 g; 8%). Due to low isolated mass, significant amount of grease are present in the isolated compound. **^{1}H NMR** (400 MHz, $CDCl_3$) δ 8.68 (ddd, *J* = 4.8, 2.0, 0.8 Hz, 1H), 8.28 (dt, *J* = 8.0, 0.8 Hz, 1H), 8.07 (td, *J* = 7.6, 2.0 Hz, 1H), 7.70 (ddd, *J* = 7.6, 4.8, 1.2 Hz, 1H); **^{13}C NMR** (101 MHz, $CDCl_3$) δ 156.9, 149.2, 140.0, 129.3, 118.2.

Isolation of 2-nitrobenzofuran

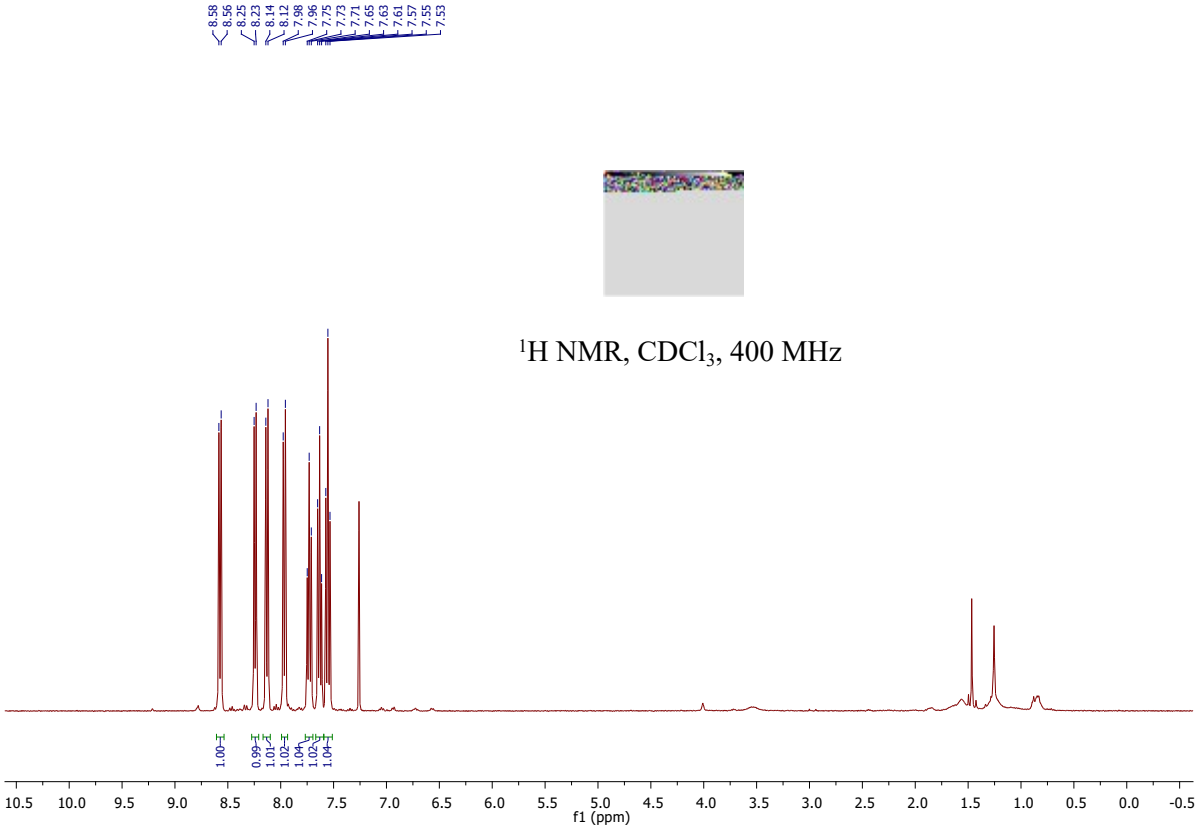


Procedure: A 20 mL Schlenck tube equipped with a magnetic stirbar was charged with bismuth(III) nitrate pentahydrate (2 equiv, 1 mmol, 485 mg), AIBN (2 mol%, 0.01 mmol, 1.6 mg) and 2-bromobenzofuran (1 equiv, 0.5 mmol, 99 mg, 60 µL) in MeCN (5 mL). The vial was sealed (under air atmosphere) and the reaction mixture was stirred at 100°C for 24 h. The mixture was then analyzed by GC-MS.

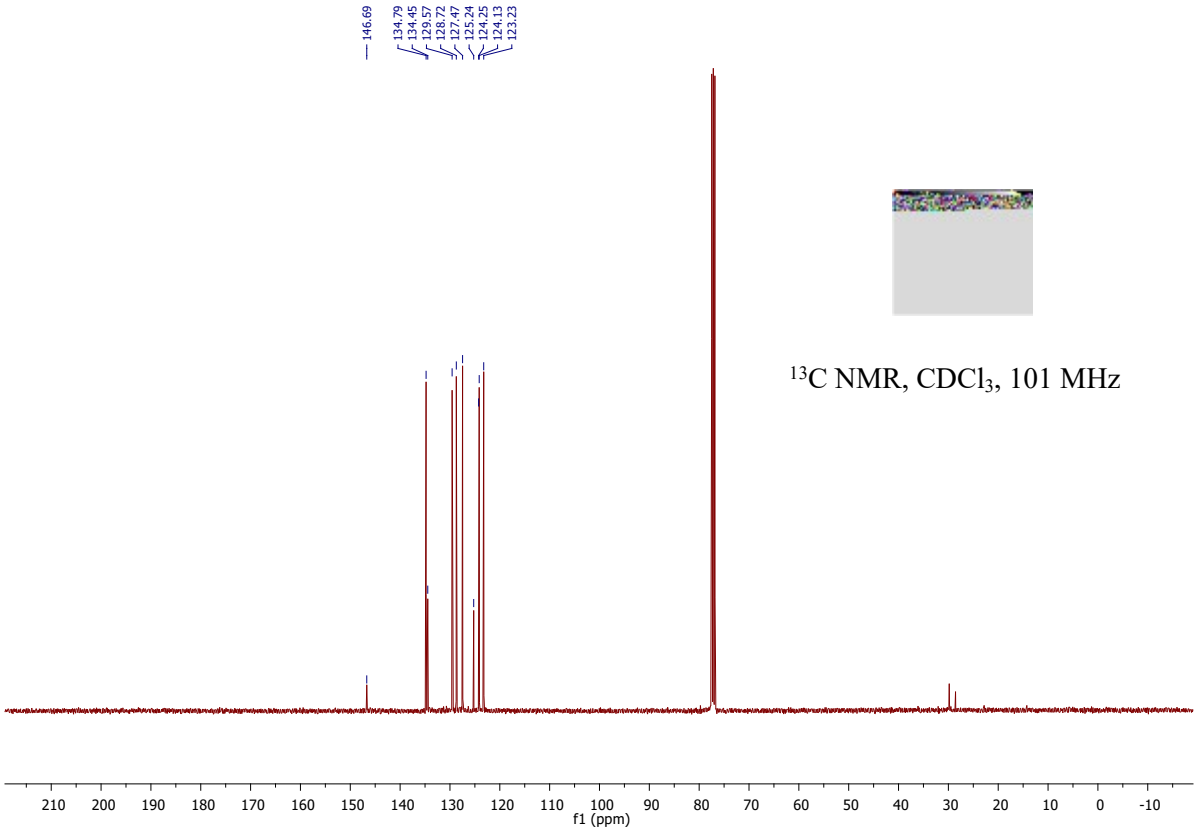


2-nitrobenzofuran: Known compound [4] prepared according to general conditions and purified by flash chromatography (cyclohexane/ethyl acetate, gradient 100:0 to 9:1) to give an orange solid (0.029 g; 36%). **^{1}H NMR** (400 MHz, $CDCl_3$) δ 7.78 (dt, *J* = 8.0, 0.8f Hz, 1H), 7.69 (d, *J* = 0.8 Hz, 1H), 7.65 – 7.58 (m, 2H), 7.45 – 7.41 (m, 1H); **^{13}C NMR** (101 MHz, $CDCl_3$) δ 153.5, 153.2, 130.1, 126.0, 125.5, 124.2, 112.9, 107.4.

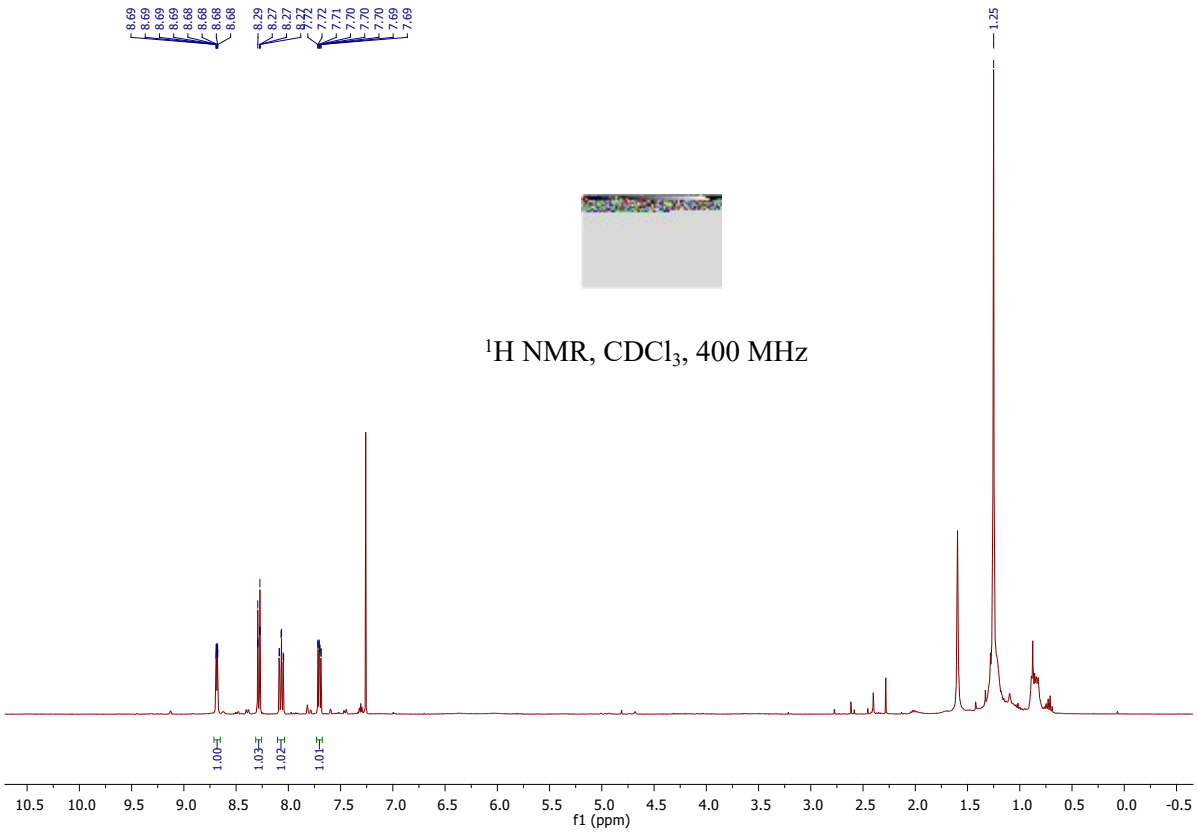
NMR Spectra



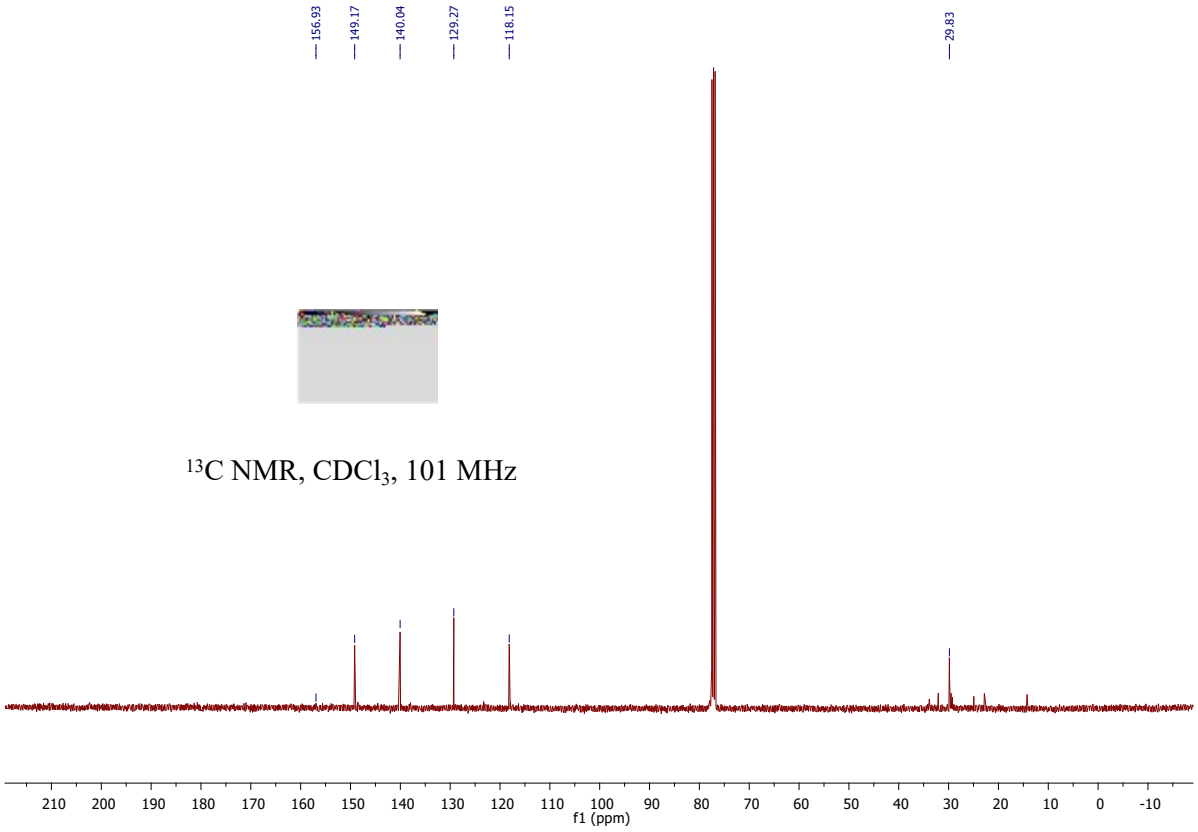
^{1}H NMR, $CDCl_3$, 400 MHz



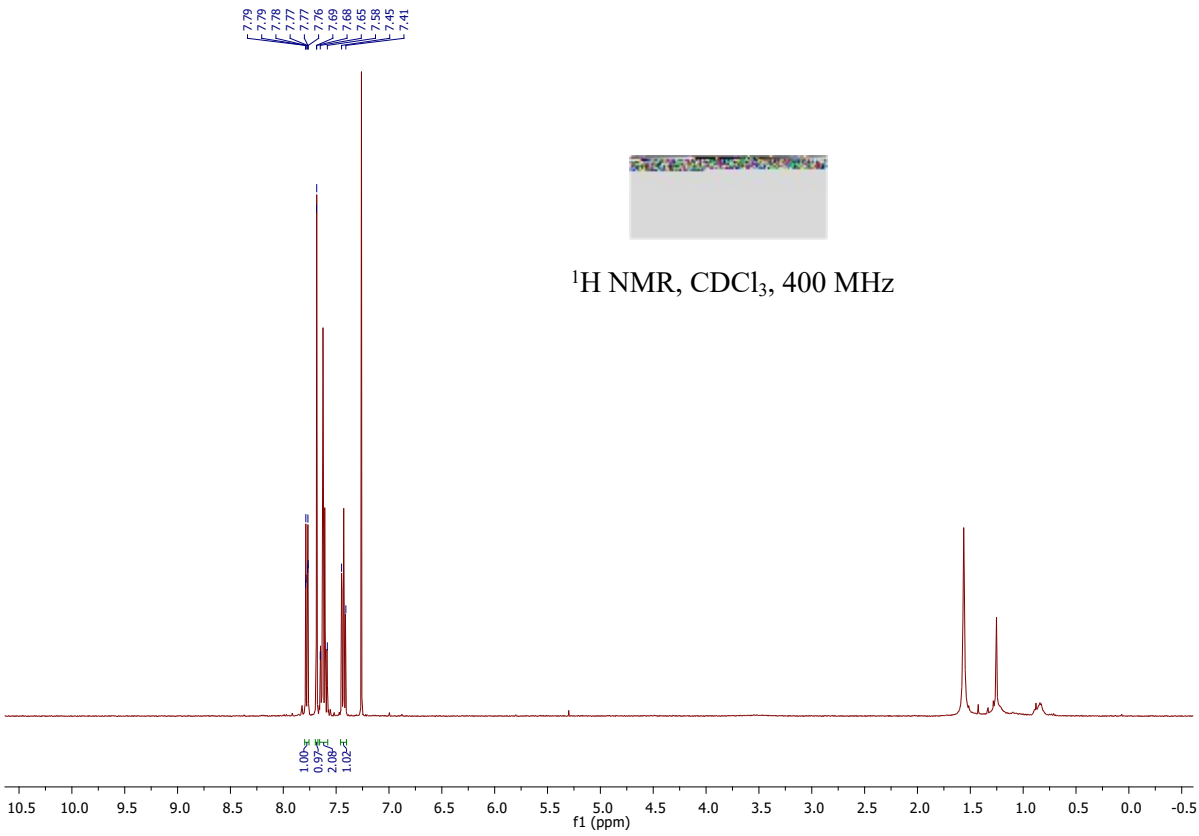
^{13}C NMR, $CDCl_3$, 101 MHz



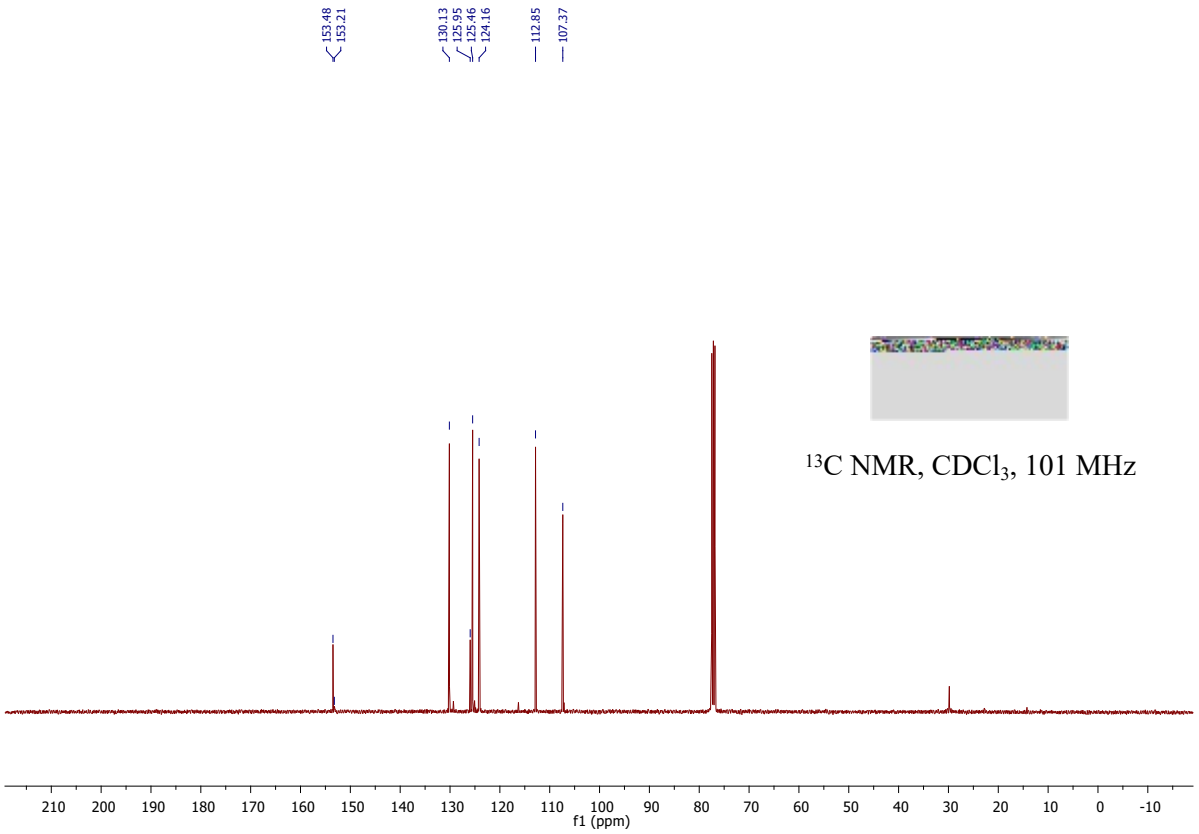
^{1}H NMR, $CDCl_3$, 400 MHz



^{13}C NMR, $CDCl_3$, 101 MHz



^{1}H NMR, $CDCl_3$, 400 MHz



^{13}C NMR, $CDCl_3$, 101 MHz

Machine learning

Data repartition, binary classification and metrics definition

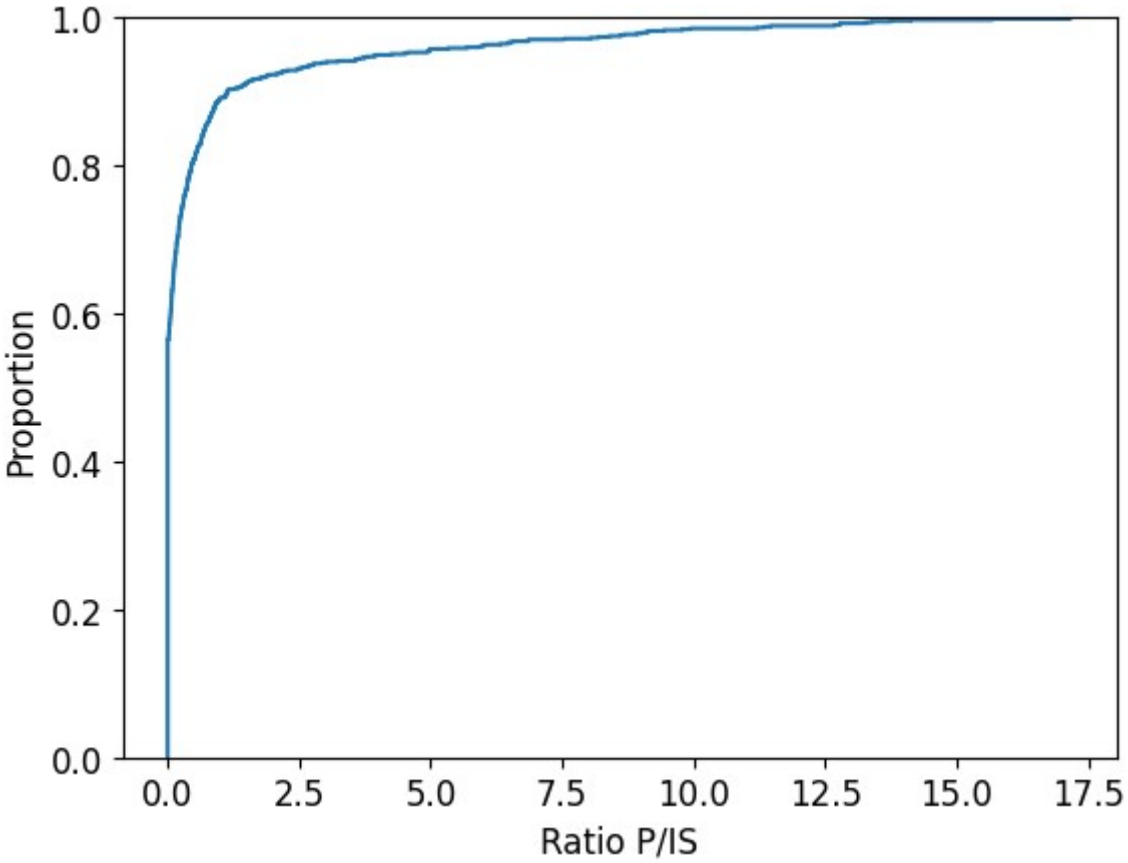


Figure S13: Cumulative distribution of the ratio between product and internal standard signals

The cumulative distribution of the ratio between product and internal standard signals reveals that 56 % of experiments did not yield any detectable product. Given this data distribution, reaction output prediction is approached as a binary classification problem where the failure (P/IS=0) or success (P/IS>0) of the reaction is the prediction target, depending on the substrate, nitrating reagent and activation agent.

Descriptors encoding: Nitration and activation agents were encoded as one-hot vectors of 20 components (12 possible nitrating agents + 8 activation agents). Substrates are either encoded using Morgan fingerprints (explained in greater details below) or rdkit descriptors.

Target encoding: For each experiment, the target is set to 1 if P/IS>0 or to 0 if p/IS=0.

Classification metrics : Binary classification models can either correctly predict reaction success (true positive, TP) or reaction failure (true negative, TN). If not correct, models can be wrong on two ways : incorrectly predict that a failed reaction succeeded (false positive, FP) or incorrectly predict that a successful reaction failed (false negative, FN). The accuracy, sensitivity, precision and specificity metrics are defined using these values:

Accuracy : (TP+ TN) / (TP+TN+FP+FN)

Sensitivity : TP / (TP+FN)

Precision : TP / (TP+FP)

Specificity : TN / (TN+FP)

The accuracy reflects a model's capacity to correctly predict the experiment success or failure. Sensitivity measures its capacity to detect all successful experiments, Precision measures its capacity not to label a failed experiment as successful and Specificity quantifies its capacity to correctly identify failed experiments. For imbalanced datasets, where one class is largely predominant over the other one, the accuracy may not be appropriate because correctly classifying experiments belonging to the underrepresented class becomes harder, and the accuracy doe not reflect this added difficulty. In these cases, the balanced accuracy is more appropriate because it gives equal weight to correctly predicting 4

Balanced Accuracy : (Sensitivity + Specificity) / 2

Molecular Fingerprints

Fingerprints capture the presence of structural features on molecules (atom- and bond-wise). There are typically encoded as binary vectors with a user-defined size. If a specific atomic environment is present on a molecule, a bit at a specific fingerprint position is set to 1. However, a bit set to 1 at a specific position in a fixed-sized fingerprint does not necessarily correspond to a single specific environment, as several different atomic environments may be encoded on the same bit depending on the fingerprint size. This is referred to as "bit collision" and is discussed below.

We used the *GetFingerprint()* function implemented in rdkit's *Chem.rdFingerprintGenerator.GetMorganGenerator()* module to compute circular fingerprints with the Morgan algorithm up to a radius of 2 and with different fingerprint sizes.

For each substrate considered in this study, atomic environments corresponding to bits that are set to 1 in their fingerprints are represented below. The central atom of the environment is highlighted in blue, atoms up to a radius 2 of this central atom are highlighted in yellow, and atoms/bonds that are drawn in light gray indicate pieces of the structure that influence the atoms' connectivity but that are not directly part of the fingerprint.

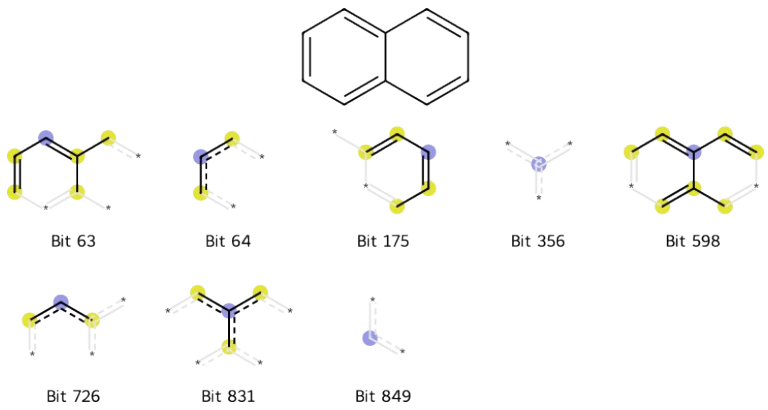


Figure S14: Bits set to 1 in naphthalene's fingerprint (Morgan fingerprints of radius 2 and size 1024)

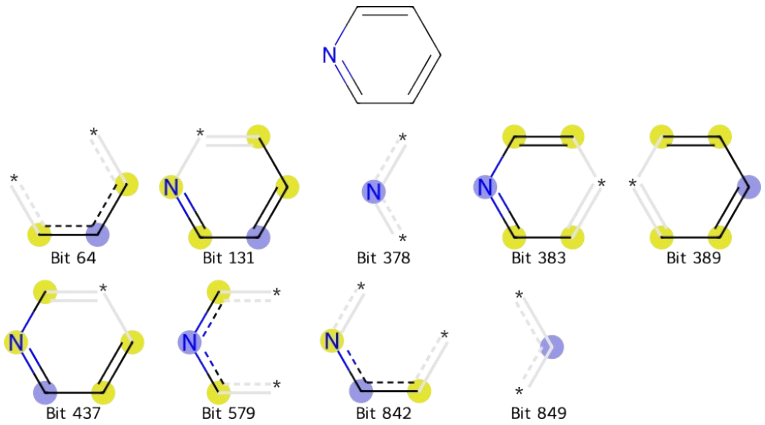


Figure S15: Bits set to 1 in pyridine's fingerprint (Morgan fingerprints of radius 2 and size 1024)

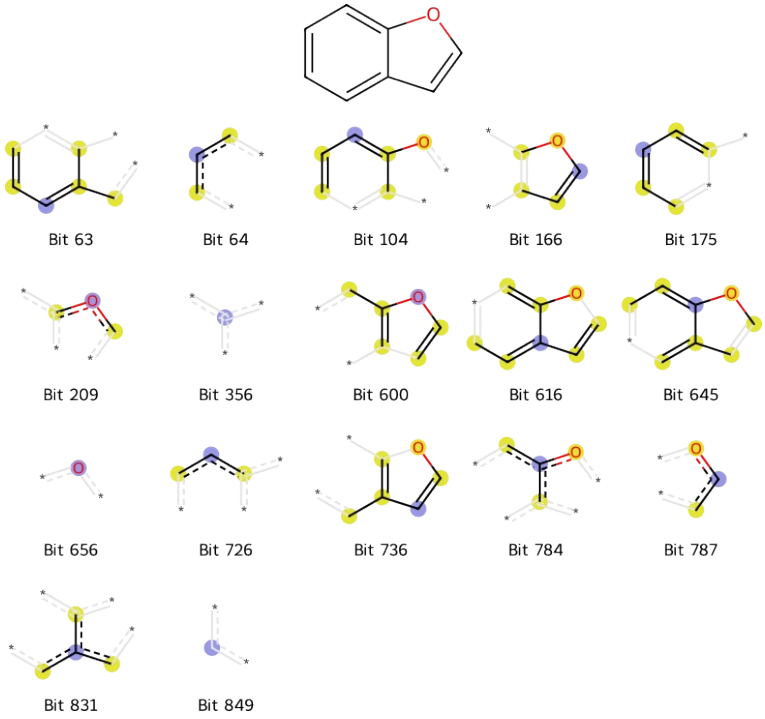


Figure S16: Bits set to 1 in benzofuran's fingerprint (Morgan fingerprints of radius 2 and size 1024)

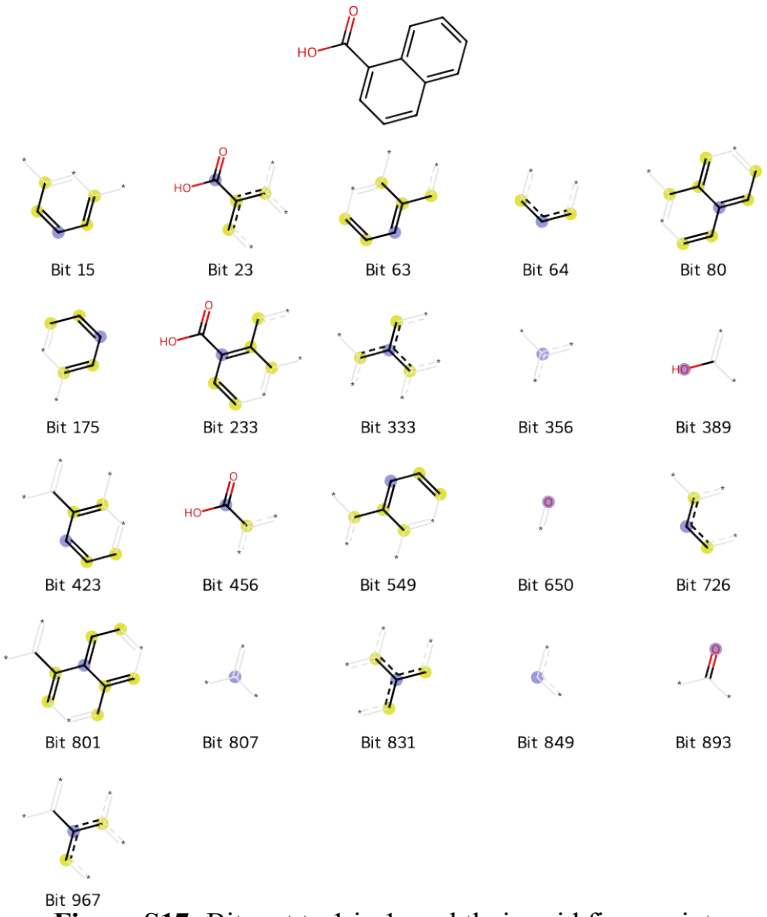


Figure S17: Bits set to 1 in 1-naphthoic acid fingerprint (Morgan fingerprints of radius 2 and size 1024)

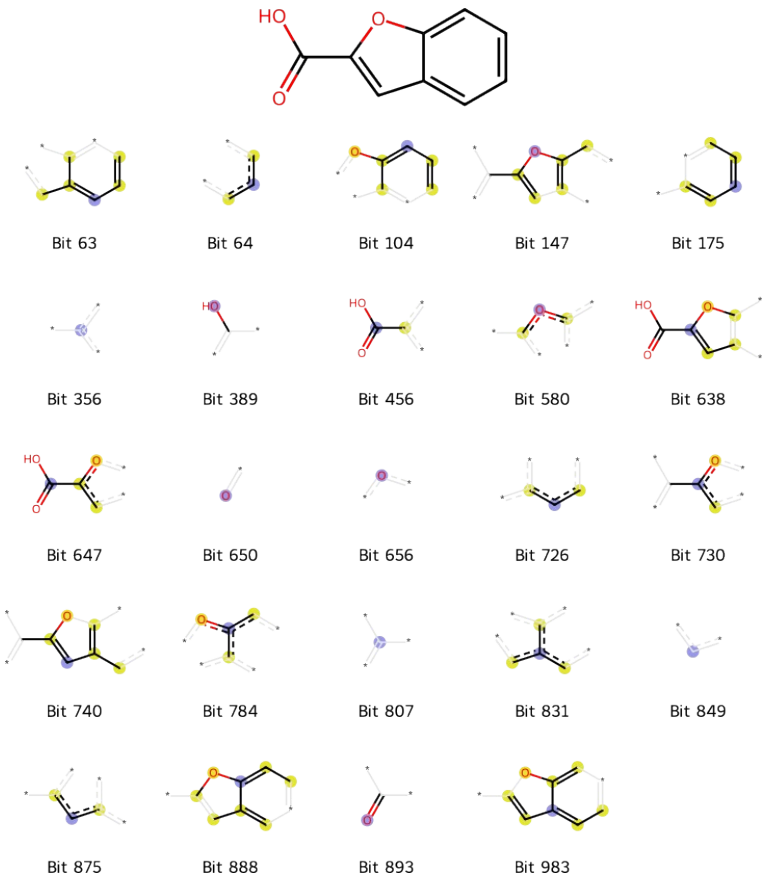


Figure S18: Bits set to 1 in benzofuran-2-carboxylic acid fingerprint (Morgan fingerprints of radius 2 and size 1024)

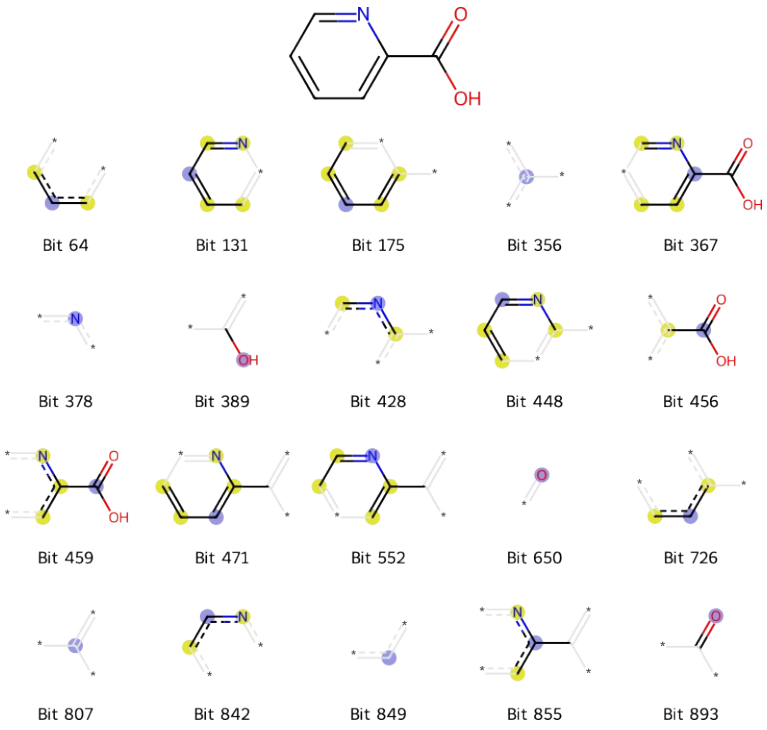


Figure S19: Bits set to 1 in picolinic acid fingerprint (Morgan fingerprints of radius 2 and size 1024)

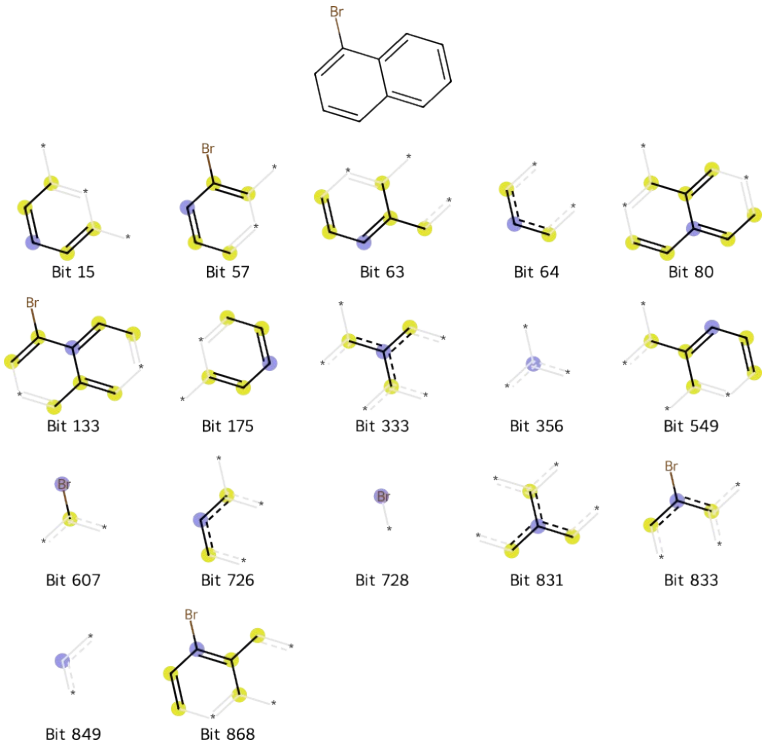


Figure S20: Bits set to 1 in 1-bromonaphthalene fingerprint (Morgan fingerprints of radius 2 and size 1024)

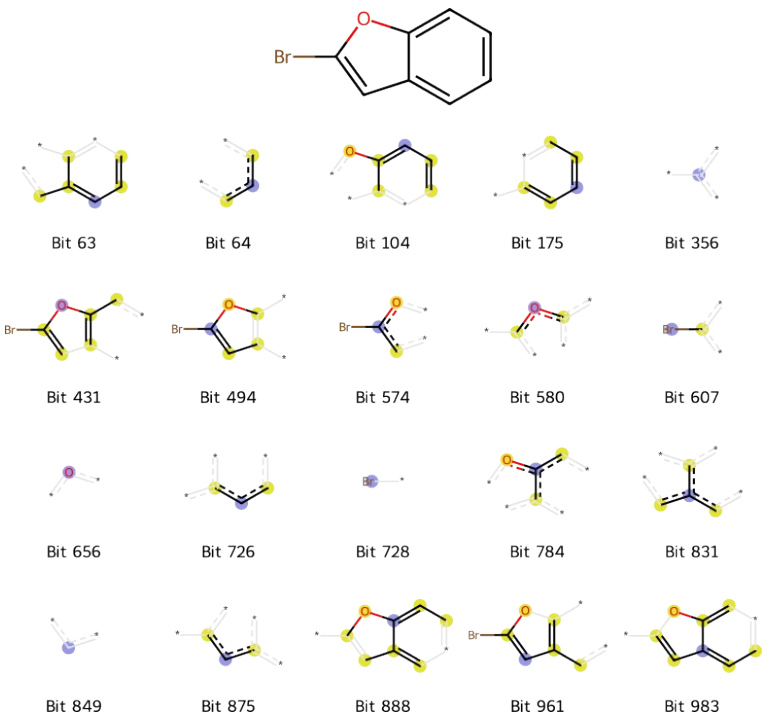


Figure S21: Bits set to 1 in 2-bromobenzofuran fingerprint (Morgan fingerprints of radius 2 and size 1024)

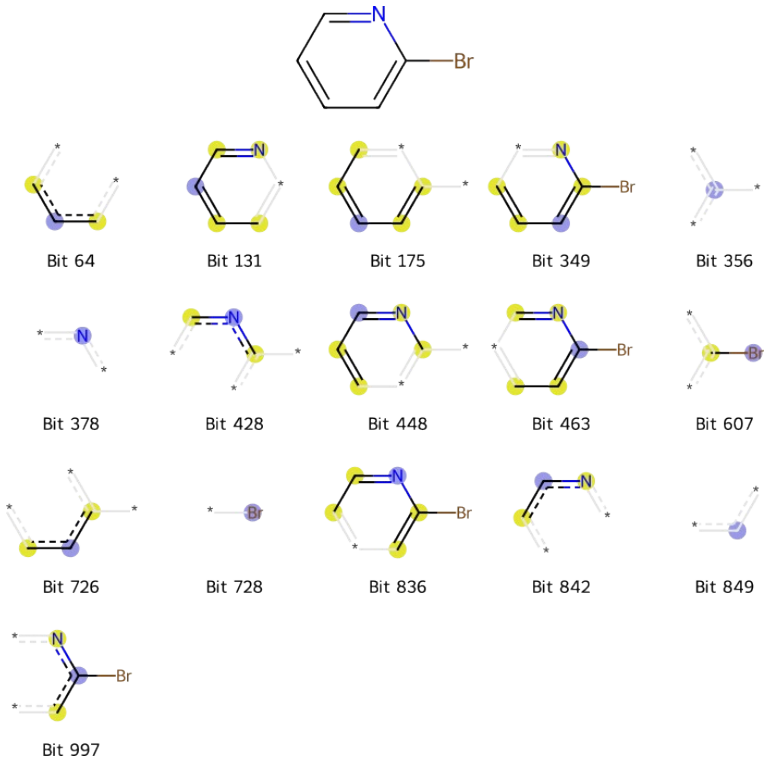
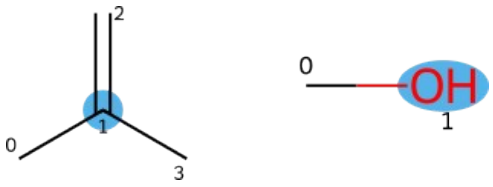


Figure S22: Bits set to 1 in 2-bromopyridine fingerprint (Morgan fingerprints of radius 2 and size 1024)

Atomic environments and bit collisions in fingerprints



Consider the two groups highlighted in blue in the above molecules. Following rdkit's implementation of the fingerprinting algorithm used here, these carbon atom and OH group are represented respectively by the integers 864662311 and 2246699815. With a fingerprint of size 1024, they both correspond to the bit at position 807 because 864662311%1024 = 2246699815%1024 = 807, where "%" denotes the modulo operation. For each of these molecules' fingerprints, a 1 is thus found at position 807, although this bit does not reflect the presence of the same environment. This mathematical artifact is known as bit collision. A fingerprint of different size (not necessarily larger) would be needed for these two particular environments not to be encoded on the same bit. In general, increasing the fingerprint size decreases the overall probability of bit collision, but this is not necessarily recommended when fingerprints are used as input to machine learning algorithm, for reasons discussed above. Detailed discussion about fingerprinting molecular structures can be found in [5].

Feature importance

AdaBoost, Random Forest, Decision Trees, Extra Trees and Gradient Boosting algorithms are able to automatically generate descriptor importance, i.e., scores reflecting the usefulness of each descriptor in their classification process. When these descriptors can be interpreted on a molecular basis, their scores can help rationalize the input-output relationship. For each of these five algorithms, the four following descriptors always appear in the top 10 most important descriptors:

- MolWt: the molecular weight,
- NumValenceElectrons: number of valence electrons,
- MinEStateIndex: the minimum electrotopological state index [6], combining both the electronic and topological environments of atoms in a molecule. These indices have been found to correlate with NMR ^{17}O chemical shift for a series of carbonyl compounds.
- BCUT2D_MWLOW: for a molecule containing N atoms, BCUT2D_MWLOW is the lowest eigenvalue of a NxN matrix containing atomic masses on the diagonal elements, 1/sqrt(bond order) on matrix elements corresponding to atoms that are bonded together, and 0.001 elsewhere. The bond orders are respectively set to 1, 2, 3 and 1.5 for single, double, triple and aromatic bonds. More details about the BCUT (Burden-CAS-University of Texas) descriptors can be found in [7].

The three following descriptors appear in the top 10 descriptors for four of these algorithms:

- MaxAbsEStateIndex: the maximum absolute value of electrotopological state index [6]
- MinAbsEStateIndex: the minimum absolute value of electrotopological state index [6]
- QED: the quantitative estimation of drug-likeness [8], a metric comprised between 0 and 1 defined in term of the molecular weight, logP, the topological polar surface area, the number of hydrogen bond donors and acceptors, the number of aromatic rings and rotatable bonds, and the presence of unwanted chemical groups.

From a machine learning perspective, the most useful descriptors are the ones which are highly correlated to the prediction target, but not correlated to one another. Figure S23 shows that this ideal situation is far from being met, as most of the seven aforementioned descriptors are highly correlated to one another. This precludes the identification of intuitive molecular descriptors that could be used to infer the reactivity of an unseen substrate. The fact that descriptors somehow related to the molecular weight (MolWt, BCUT2D_MWLOW and QED) are useful for the classification algorithms is not surprising though, because the nature of the leaving groups studied here (H, COOH and Br) greatly influence both reactivity and total molecular weight. The same goes for the number of valence electrons, which systematically increases by 6 from H to Br and by 10 from Br to COOH. Given that electrotopological state indices capture some of the electronic character of the substrates, it comes as no surprise that descriptors related to these indices also play an important role in the reactivity classification process. Again, this retrospective analysis of the machine learning results should be nuanced by the fact that these descriptors are highly correlated to one another, computed solely on the basis of 2D topological considerations, and completely ignore the nature of activation and nitration agents. Finding descriptors that are both chemically intuitive and efficient for machine learning algorithms would need further research.

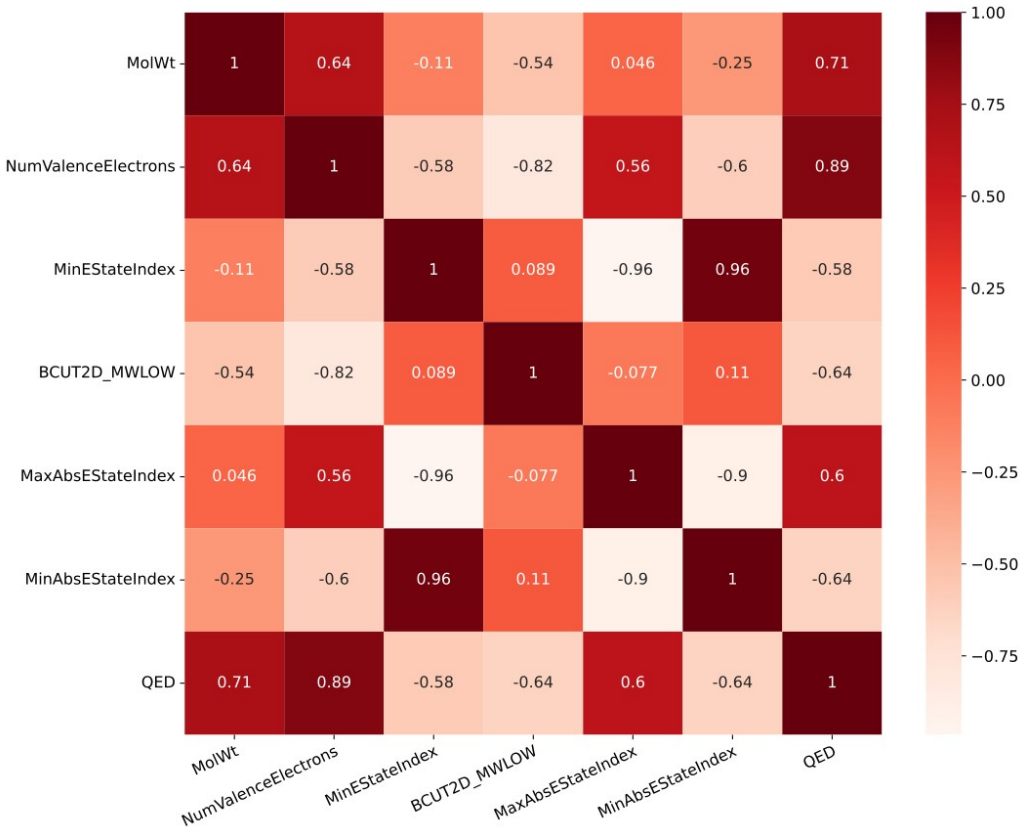


Figure S23: Pearson correlation matrix between the seven most important rdkit descriptors. The Pearson correlation coefficient is computed among each possible pairs of descriptors. A coefficient close to 1 (-1) indicates a positive (negative) linear correlation between two pairs, whereas a coefficient close to 0 indicates low linear correlation.

Studies in terms of Balanced accuracies

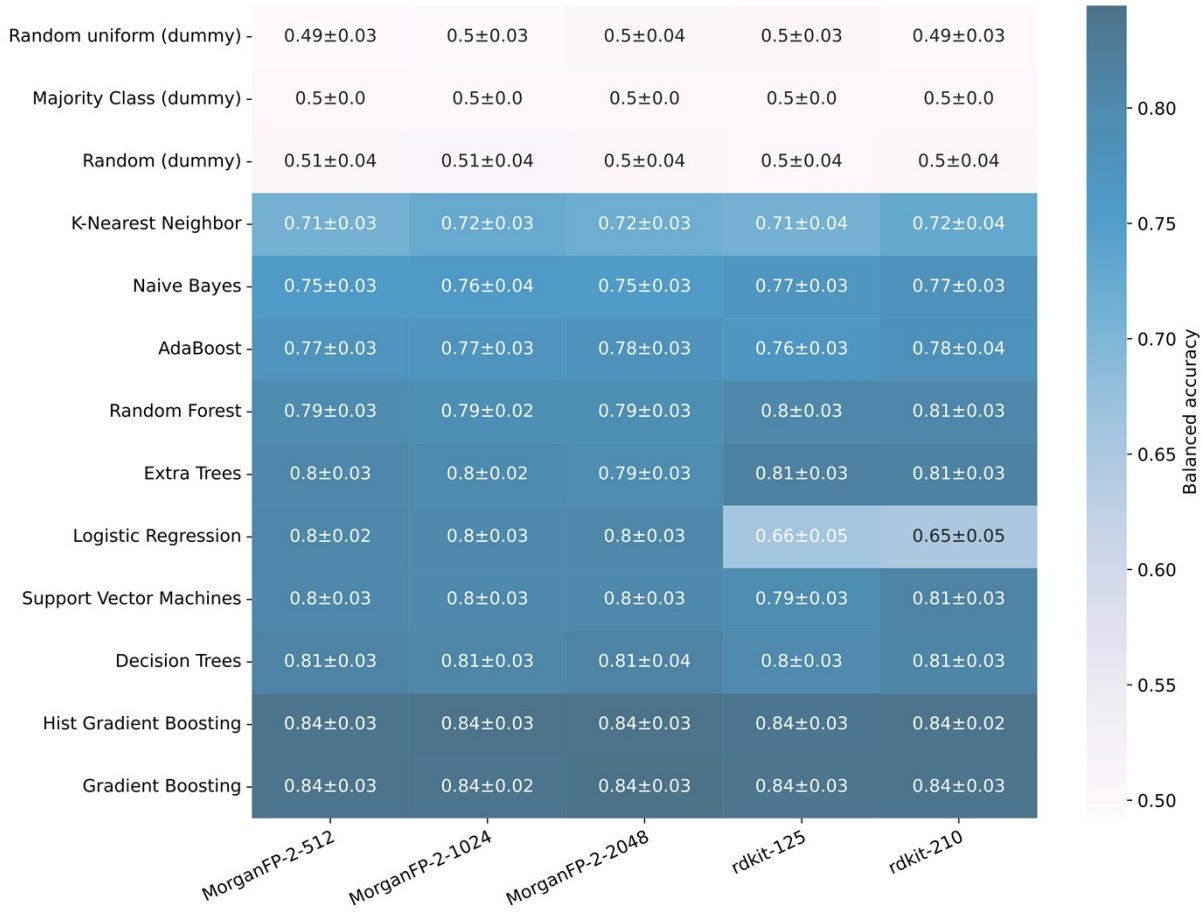


Figure 24. Balanced accuracy for the entire dataset depending on the predicting model and the chemical descriptors set.

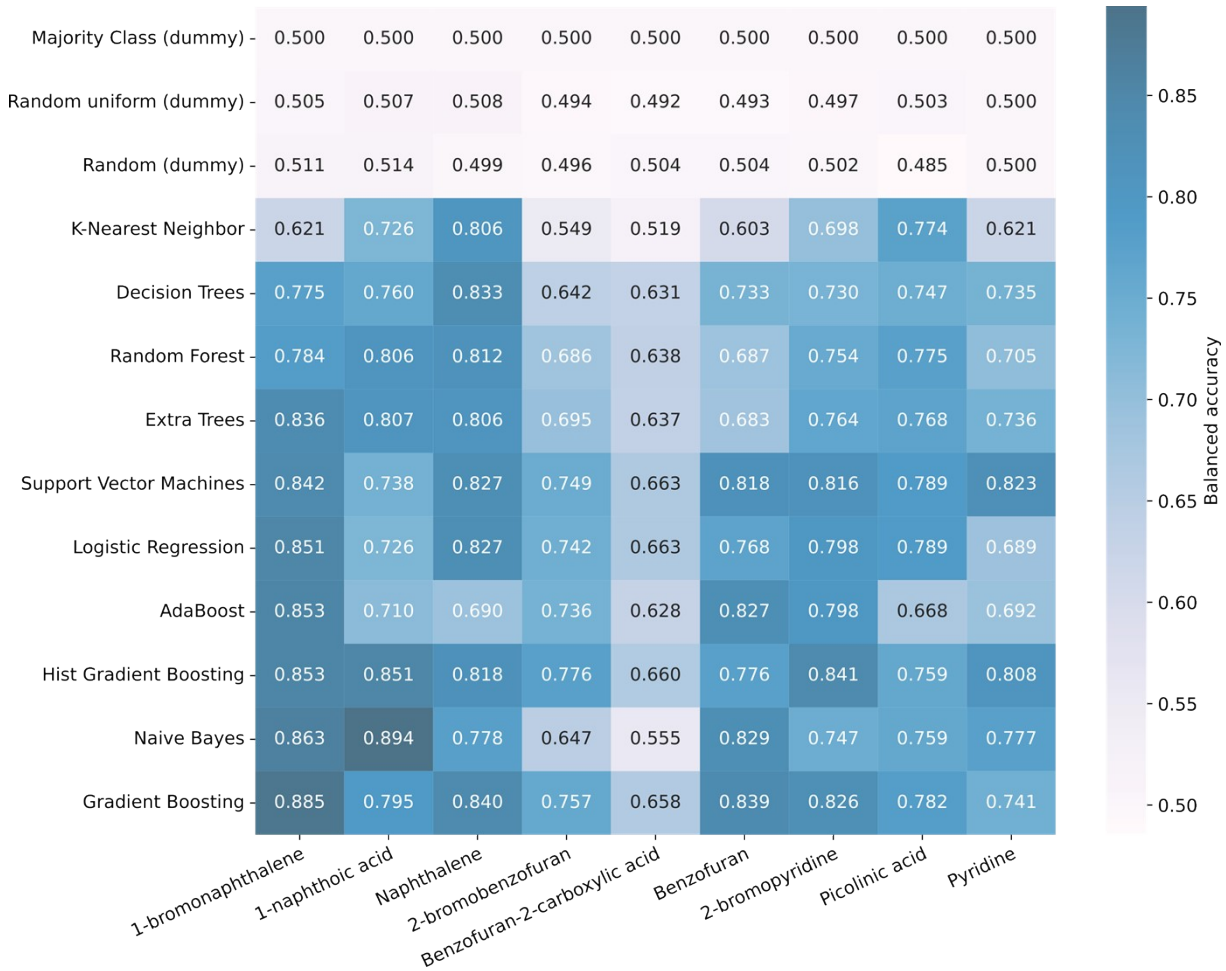


Figure 25. Balanced accuracy depending on the left-out substrate and the predicting model.

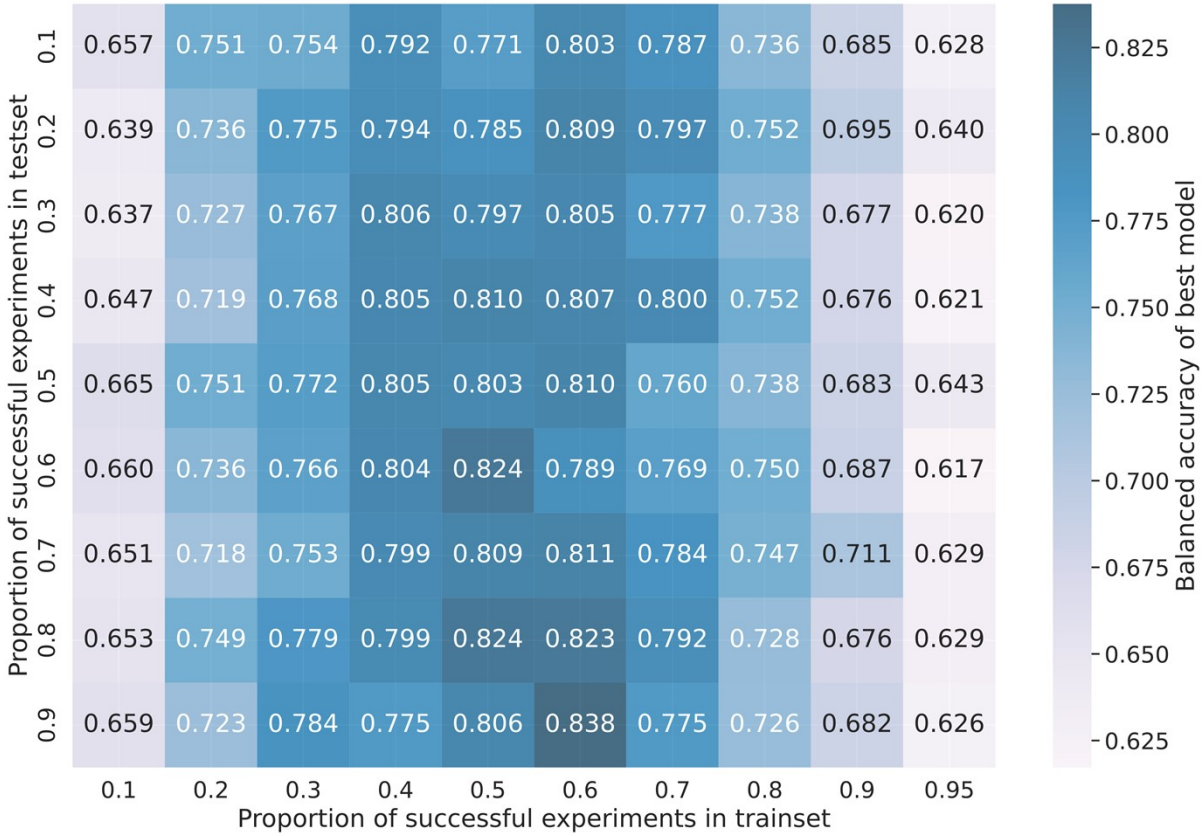


Figure 26. Balanced accuracy of best model obtained for different splits of the train and test sets.

According to the balanced accuracy metric, the best models are obtained when the trainset is well balanced, containing between around 40 and 60% of successful experiments, irrespective of the proportion of successful experiments in the testset. On the other hand, the worst-performing models correspond to the one trained on highly unbalanced trainsets.

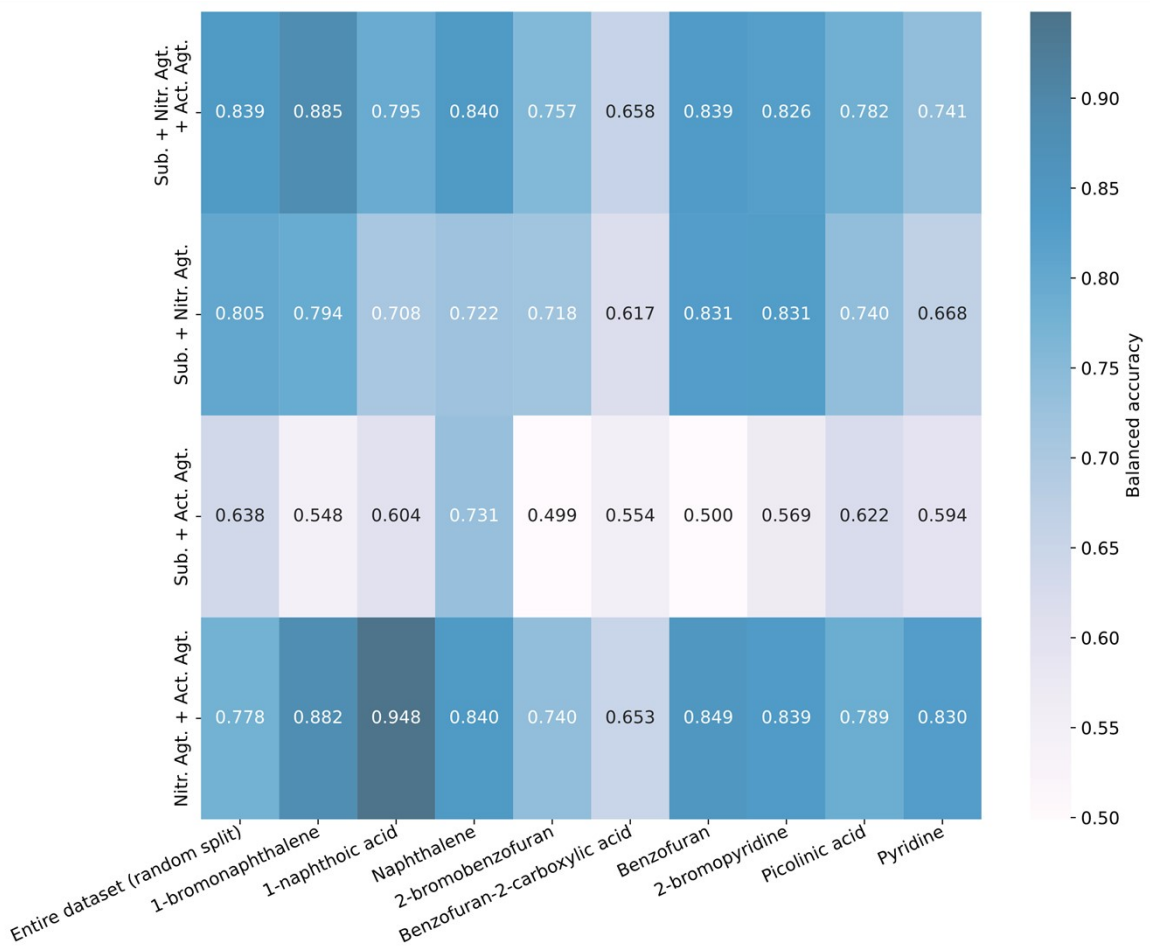


Figure 27. Balanced accuracy of Gradient boost algorithm trained with only two descriptors out of three.

References

[1] K. Zhang, A. Budinská, A. Passera, D. Katayev, *Org. Lett.* **2020**, *22*, 2714–2719.

[2] R. Calvo, K. Zhang, A. Passera, D. Katayev, *Nat. Comm.* **2019**, *10*.

[3] E. Voutyritsa, A. Theodorou, M. G. Kokotou, C. G. Kokotos, *Green Chem.* **2017**, *19*, 1291-1298.

[4] J. Ling, M. Laugeois, V. Michelet, V. Ratovelomanana-Vidal, M. R. Vitale, *Synlett* **2018**, *29*, 928-932.

[5] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* 2010, *50*, 742–754.

[6] L. H. Hall, B. Mohney, L. B. Kier, *Quant. Struct.-Act. Relat.* **1991**, *10*, 43-51.

[7] R. S. Pearlman, K. M. Smith, *Perspectives in Drug Discovery and Design*, **1998**, 339–353.

[8] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, A. L. Hopkins, *Nat. Chem.* **2012**, *4*, 90–98.