An Artificial Intelligence that Discovers Unpredictable Chemical Reactions

Dario Caramelli, Jarosław M. Granda, Dario Cambié, S. Hessam M. Mehr, Alon Henson and Leroy Cronin* School of Chemistry, the University of Glasgow, University Avenue, Glasgow G12 8QQ, UK. *Correspondence to: Lee.Cronin@Glasgow.ac.uk.

The robot-driven detection of novel organic chemical reactions[1-2] is difficult as there is no approach that guarantees discovery on demand[3]. Traditional approaches to find new chemical reactions often rely on human error, whilst theoretical[4,5] and artificial intelligence[6-8] approaches promise new targets, but these must be identified computationally[9,10]. However, it is very hard to turn these ideas into reality in a chemistry laboratory, and the targets are not often novel. Herein, we present an artificial intelligence, built to autonomously explore chemical reactions in the laboratory using deep learning[11]. The reactions are performed automatically, analysed online, and the data is processed using a convolutional neural network (CNN) trained on a small reaction dataset to assess the reactivity of reaction mixtures[12]. The network can be used to predict the reactivity of an unknown dataset, meaning that the system is able to abstract the reactivity assignment regardless the identity of the starting materials. The system was set up with 15 inputs that were combined in 1018 reactions, the analysis of which lead to the discovery of a 'multi-step, single-substrate' cascade reaction and a new mode of reactivity for methylene isocyanides. *p*-Toluenesulfonylmethyl isocyanide (TosMIC) in presence of an activator reacts consuming six equivalents of itself to yield a trimeric product in high (unoptimized) yield (47%) with formation of five new C-C bonds involving *sp*-sp^2 and *sp*-sp^3 carbon centres. A cheminformatics analysis reveals that this transformation is both highly unpredictable and able to generate an increase in complexity like a one-pot multicomponent reaction.

Organic synthesis is intrinsically target oriented[13] which means that the discovery of new reactions is a chance event, or results from the need to access a new transformation[14]. The development of new transformations and methodologies[15] however is a complex problem requiring a high degree of expert knowledge[16,17]. Furthermore, the current approaches to reaction or method discovery are generally constrained to known heuristics and the discovery of novel reactions is a rare event that often relies on serendipitous results. The search for unexpected results can be accelerated with automated systems and in the last decade high-throughput experimentation[18] has shown its potential in speeding up reaction preparation and analysis (typically applied in reaction optimization and combinatorial chemistry)[19-21]. However, an increase of reaction throughput does not automatically lead to the serendipitous discovery of entirely new transformations while, on the other hand, the discovery of new reaction pathways from first principle (i.e. *in silico*, based on quantum mechanics) is hard both due to the combinatorial explosion of possible reaction pathways and the computational cost of accurate modelling of the energy hypersurface. To overcome these limitations, an increasing number of approaches are starting to involve a feedback loop from the on-line analytics and a decision-making algorithm to perform only a fraction of the possible combinations, considered interesting[20]. In such a "closed-loop"[21-25] approach, the system automatically explores a chemical space in a trial-and-error way similar to how a human experimenter would do it. The system requires three main parts: a chemical robot to perform and analyse the reactions, a program for the analytical data interpretation, and an algorithm that correlates the outcome of the reaction with the input and process parameters. This last part closes the loop by suggesting the predicted optimal parameters for the next reactions (Figure 1). Although closed-loop approaches have proved effective in reaction optimization, their application towards the discovery of new reactions remains underexplored. This is because assessing the reactivity of an unknown reaction with unpredictable products is harder than using metrics meant for

optimization of a known target compound, such as yield or selectivity. Whilst binary reactive/non-reactive classification of reactions has led to progress[12,26], we hypothesised that a continuous measure would allow the system to abstract the notion of reactivity, rather than restrict it to a given fixed set of reagents. To exploit this, we designed a system to explore reactivity of an experimental space using a liquid handling platform to prepare and analyse the reactions, a convolutional neural network for the reactivity assessment and a linear regressor to correlate the starting materials to the reactivity.

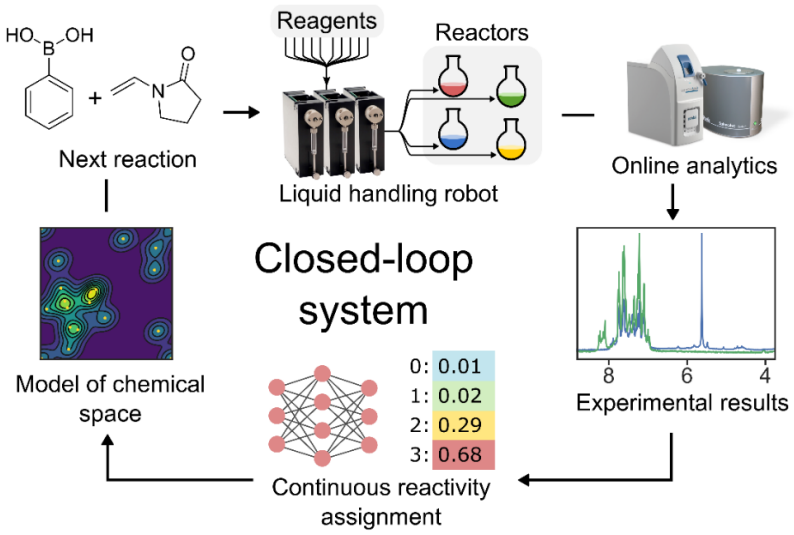


Figure 1: **Closed-loop framework for chemical space exploration.** A liquid handling robot performs an experiment and collects NMR and MS spectra. This data is processed to extract information about the reactivity and used to create a model of the chemical space to suggest the next experimental parameters for the robot.

The chemical robot

To emulate a human chemist, experiments were performed in conventional round bottom flasks that were automatically cleaned after each reaction by flushing them with clean solvent. Starting materials were stored as 1M stock solutions in dimethyl sulfoxide (DMSO) and the platform used 30 syringe pumps with integrated valves to mix them into six parallel reactors. As optional expansion of the chemical space, three of the reactors were also equipped with

visible-light Light Emitting Diods (LED) to promote photochemical reactions. During the exploration of Chemical space 2 (Figure 3c) the reactions performed in these reactors were prepared adding 2.5% mol of a molecule known to act as photocatalysts, associated to the LED wavelength: 2,4,6-triphenylpyrylium tetrafluoroborate (**PC1**, 405 nm), tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate (**PC2**, 450 nm) and rose bengal (**PC3**, 565 nm). After three hours the mixtures were analysed automatically with a benchtop NMR and MS. The software managed the preparation and analysis of the reactions. It was designed to run them in parallel by shifting each experiment starting time to efficiently share the on-line analytics and cleaning cycles. Through this optimized schedule it was possible to perform up to 36 reactions per day, each with a reaction time of three hours giving a total of over 100 reaction hours per day (Figure 2).

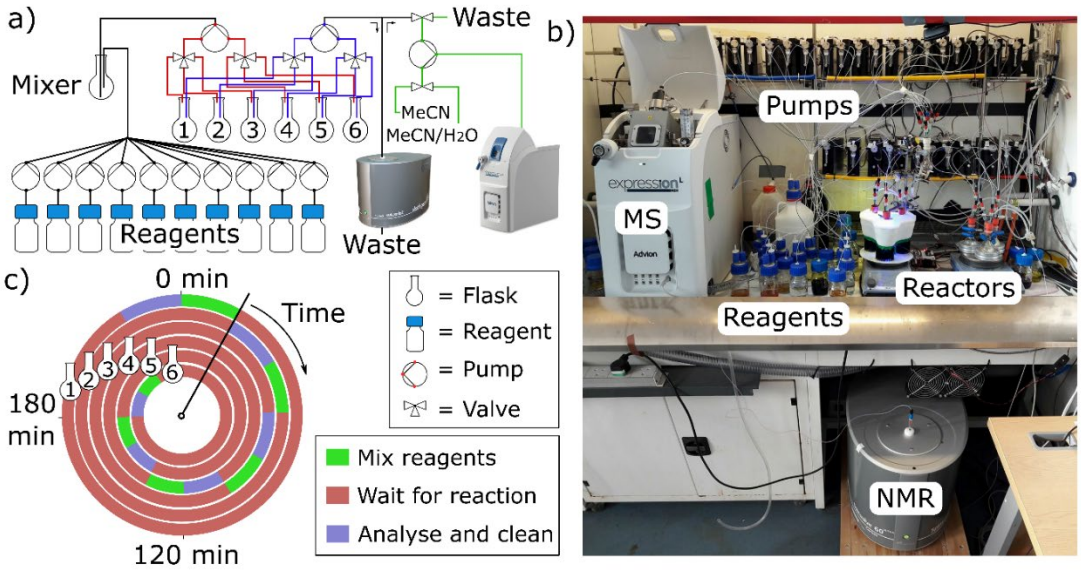


Figure 2: **Liquid handling platform** (**a**) Scheme of the platform, a series of reagents are added by dedicated pumps to a mixer flask. A pump expanded with two extra valves (obtained by removing the syringe from a normal pump) is used to transfer the reaction mixture in one of the six reactors (1-6) (red). Similarly, another pump with the same set-up (blue) is used to connect the reactors with the benchtop NMR. Finally, a third expanded pump is used for an in-line dilution prior to injection in the MS (green). (**b**) Picture of the platform. The pumps are visible on the shelves, on two lines. At the bottom there is the NMR instrument equipped with a flow probe, The MS is on the left, the reactors are in the centre while the reagents, the solvent drum and the waste container are on the left. (**c**) Six parallel reactions were started with a time-offset to allow the platform to continuously perform physical operations.

The reactivity assignment algorithm

Initially, we explored the chemical space of six simple molecules (Chemical space 1, Figure 3b) mixed in binary and tertiary combinations yielding 440 reaction. To assess the reactivity of the reaction mixtures by ^{1}H NMR we decided to use a convolutional neural network (CNN) and try to train it to mimic the assignments of a human experimenter. As any other supervised learning algorithm, we needed to provide the software with the true values associated to the inputs. The reactions were therefore manually scored by an expert organic chemist using four classes of reactivity (0 – non reactive, 3 – very reactive mixture) to describe the difference between the reaction mixture and the superimposed ^{1}H-NMR spectra of the starting materials. High values were assigned to mixtures presenting several new peaks and disappearing of the starting materials' signals. Experiments showing little or no changes in the spectra were assigned a low reactivity class (Figure 3e). The reactivity assignment network was designed purposely without any information about the chemical structures of the materials and trained to detect reactive reactions by correlating the raw spectroscopic data to the values assigned by a chemist. The network architecture involved a combination of convolutional and dense layers (Figure 3a) and was designed to accept as input a pair of NMR signals corresponding to the mixture and a reconstructed spectrum obtained by superimposing the starting materials' spectra. The output corresponded to one of the four reactivity classes. After training the network on the data of the 440 initial reactions we decided to test its performance by predicting the reactivity values of the 1018 reactions set obtained from the combinations of fifteen different materials (Chemical space 2, Figure 3c). The network showed 56% accuracy for predicting classes of reactivity and 79% top-2 accuracy (having four possible classes the random baseline is at 25% and 50% for top-2). The results are showed in the confusion matrix in Figure 3d where the class assignment made by the neural network (predicted) is compared with the manual assignment (true values).

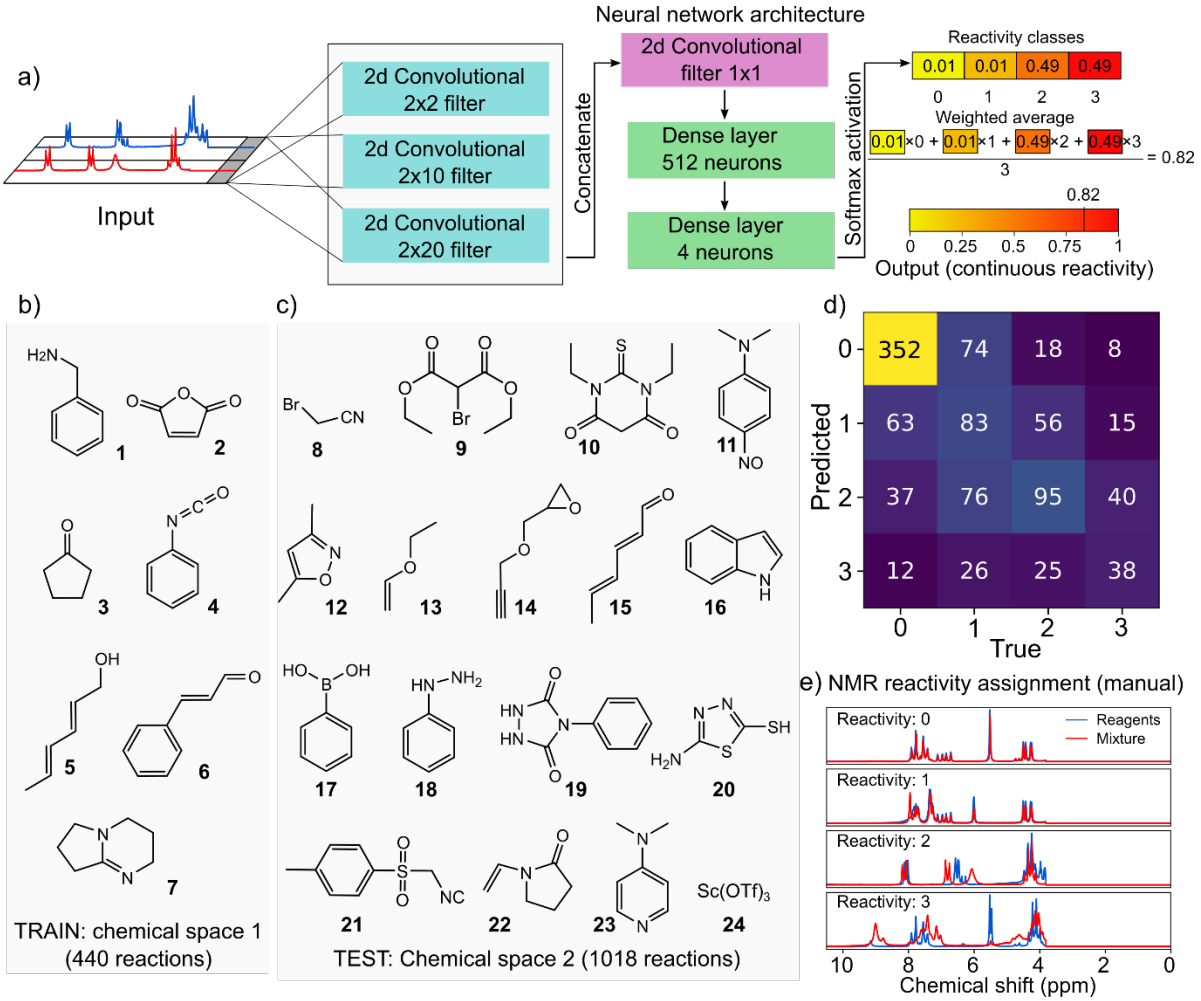


Figure 3. **A CNN has been used for the reactivity assessment using two different chemical spaces (a)**. NMR data of the mixture and the sum of starting materials are used as input. The network is trained using 440 reactions from a chemical space (**b**) and tested on 1018 reactions performed from combinations of 15 different molecules (**c**). (**d**) The accuracy on the Test set plotted as a confusion matrix shows that the network successfully learned to generalize the reactivity beyond reagents in the training set. (**e**) All data are manually classified into four classes.

The ability of the network in successfully classify the reactivity of data obtained from an unseen set of molecules suggests that the CNN was capable of abstracting the reactivity from NMR analysis regardless of the reagents used. The output classes were then converted into continuous values using a linear combination of the probabilities assigned by the CNN to the 4 values and then normalised to unity. For example, if the network output of an experiment

was [0.01, 0.01, 0.49, 0.49], meaning that there is a 49% chance of the assignment being 2 and 3 then the overall value is 2.46 which corresponds to a normalised result of 0.82.

Algorithm for the exploration of the chemical space

In order to close the loop and drive the exploration of the chemical space we implemented a linear regressor model to correlate reagents (*X*) with their reactivity (*Y*), defined as described above. The idea behind the algorithm is to train the model on a small fraction of the chemical space, explored at random, and use the knowledge acquired to predict the reactivity of the remaining possible combinations. The reactant combinations are then sorted by predicted reactivity and the best candidates are reacted in the platform. After each reaction the model is re-trained using the newly obtained reactivity information. By guiding the robot with such a reactivity-driven algorithm it will be possible to perform the reactive combinations first, meaning that only a fraction of the chemical space will need to be explored. This method is easily scalable to vast chemical spaces involving several variables that would normally be too hard to search using a brute force alone. In order to feed the model, we represented the reactions' parameters into a vector of 20 elements corresponding to reagents and additives, see Figure 4a. In a similar way to the one-hot encoding pre-processing, the presence of a specific reagent was indicated as a "1" while the rest of the values were set to "0". This representation is unrelated to the chemical structure of the reagents as it only gives information about their presence in the reaction, and hence is applicable to any chemical space including any type of binary variable. While this simple model is able to predict unseen reactions, it can also provide meaningful information about the dataset. For example, by plotting the weights of the reagents after the regressor fitting we obtained a visualization of the reactivity of single reagents. Bromoacetonitrile **8** is shown to be the most reactive reagent while indole **16** the most inert (Figure 4b), in agreement with basic chemical intuition. After the first run of 1018 reactions,

the dataset was used to simulate further explorations in order to test the ability of the system to navigate the chemical space.

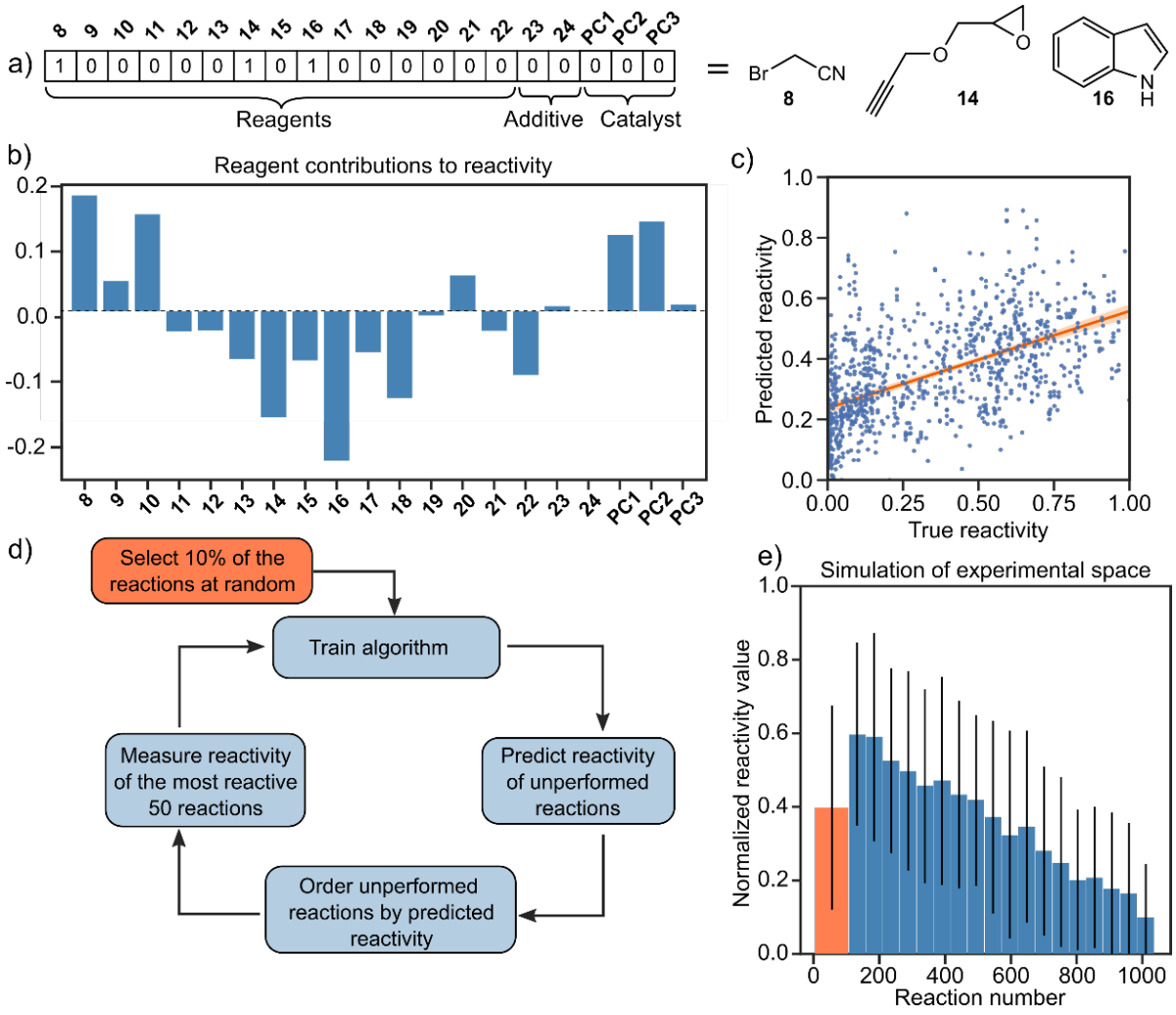


Figure 4. **The parts of the algorithm used for the chemical space exploration** (**a**) One hot encoding of the reaction parameters. The presence of each reagent is indicated by 1 while the absence is 0. The resulting vector was used as input for the linear regressor. PC1-2-3 represent the photocatalysts (**b**) The weights assigned by the linear regressor after fitting are plotted to visualize an absolute value of reactivity (**c**) Comparison of predicted versus observed reactivity for the test set. The correlation is demonstrated by fitting a linear regression model with the shaded area representing the 99% confidence interval obtained using bootstrap. (**d**) Scheme of the algorithm used to simulate the chemical space exploration. (**e**) Results of the chemical space exploration simulation. After the initial selection of 100 random reactions (orange), the algorithms start to create a model that correlates parameters to reactivity. By prioritizing combinations that are predicted to be reactive, the space is explored in a more efficient way. The error bars show standard deviation.

Initially, we tested the linear regressor on the full dataset (train/test ratio: 0.8), obtaining a mean squared error in the prediction of 0.24 (Figure 4c). Then we proceeded to simulate the exploration. The regressor was initially trained on a random 10% (ca. 100 reactions) of the dataset and used to assign the reactivity values to the remaining ca. 900 combinations. The most reactive 50 predictions were added to the "known" group of reactions. In a real closed-loop system these combinations would have been physically performed and analysed. In our simulation we restricted the predictions to the combinations in the rest of the database and when a reaction was selected to be "performed" its data was simply added to the system. The regressor was then retrained on the expanded dataset of 150 reactions and the process was repeated until all the 1018 reactions are explored (Figure 4d). The results of the simulation are showed in Figure 4e. The initial random acquisition has an average reactivity around 0.4 with a standard deviation of 0.55. After the first training, the model starts to suggest the reactions to be performed and the first batch of 50 reactions obtained an average reactivity over 0.6 with a standard deviation of 0.4. Over time the algorithm is trained on more data but at the same time the reactive combinations are taken out of the dataset leaving only the unreactive ones to be chosen (Figure 4e). A simulated exploration repeated multiple times with different random starting points showed that the new reaction discovered (reported below) was found on average after 3.8 iterations out of 20, indicating the effectiveness of this method.

Discovery of a multi-step-one-substrate reaction cascade

Based on the data obtained from the benchtop instruments, a small selection of seven combinations showing high reactivity were repeated in the platform and the products manually isolated. Three reactions (one of which photochemical) already known in literature were rediscovered and two unknown reactions were successfully discovered. One of them is a photochemical reaction involving the addition of phenylhydrazine and bromoacetonitrile in

presence of tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate and 450 nm irradiation. Details are reported in supplementary information (SI-6). The other reaction does not use light and is described in depth below. During the analysis of the mixture of *p*-toluensulfonylmethyl isocyanide (TosMIC) and diethyl bromomalonate we were intrigued by the ^{1}H-NMR spectrum of a precipitated product as it resembled the one of TosMIC, but with an additional signal in the NMR at 2.96 ppm. Further ESI-MS analysis showed a molecular peak at *m/z* = 693, suggesting that the trimerization or tetramerization product of TosMIC had occurred. Due to the high symmetry of the product we were unable to solve its structure via NMR alone, and since it was already obtained pure by precipitation during the work-up we decided to grow a single crystal for X-ray analysis. Single crystals were obtained by the slow diffusion of water into DMSO/ether 1:1 solution and X-ray analysis confirmed that the trimeric product **25** was formed. The crystal structure of the product showed a tubular supramolecular assembly composed of six molecules packed as a ring and multiple rings stacked together leaving void space with an average diameter of 12 Å (Figure 5a). To explore this transformation further we decided to perform the reaction with a range of isocyanides to elucidate a possible mechanism. From the seven isocyanides (SI-3.6) tested, a similar product was isolated in the reaction with the 1*H*-benzotriazol-1-ylmethyl isocyanide **26** while traces of the reaction with (trimethylsilyl)methyl isocyanide **27** were detected by LC-MS (Figure 5b). Variations of diethyl bromomalonate have also been explored finding working alternatives in several similar molecules including trifluoroacetic acid (TFA, **34**) (Figure 5c). All these variations yielded the same product, suggesting that the second reagent is not directly involved in the product formation but rather acts as some kind of a promoter; this is because the reaction does not give the product in any detectable amount in the absence of it. In order to confirm the involvement of the hypothesised intermediates presented in the mechanism below the reaction has also been carried out in the presence of various amines (Figure 5d) and the presence of the relative

products **35** and **39** has been established by LC-MS in all cases. These correspond to an asymmetric version of the product **25,** where one or two branches have been replaced with the amine R-group. The possibility of tuning the branches in this way gave us precious information about the mechanism and increased the flexibility and the possible applications of this reaction.

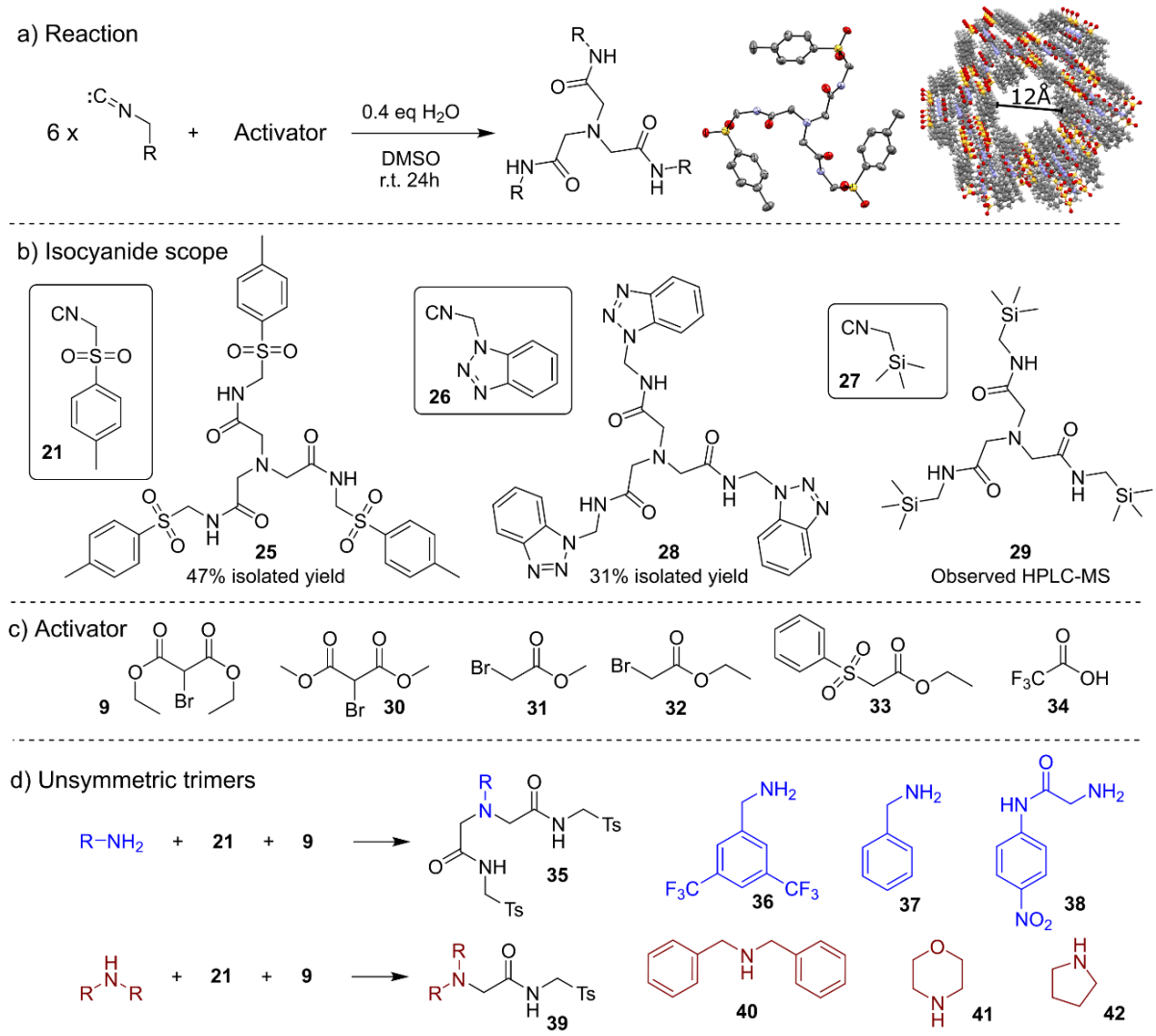


Figure. 5. **The reaction of diethyl bromomalonate and TosMIC discovered with the automated platform.** (**a**) General scheme of the reaction. Six equivalents of isocyanide are consumed in presence of an activator, water and DMSO **25,** on the right-hand side the X-ray structure of **25** and its tube-shaped supramolecular structure**.** (**b**) Analogous products obtained with variations of the isocyanide. (**c**) The reaction has been carried out using variations of diethyl bromomalonate, yielding the same product. (**d**) By performing the reaction in the presence of an amine we observed variations of the product, suggesting the mechanism reported in the next figure.

In order to understand the formation of the core of the molecule we prepared two isotopically enriched versions of the TosMIC substrate, labelling the isocyanide carbon and the methylene carbon. To confirm the source of the three oxygen atoms found in the product we performed the reaction in both synthesised ^{18}O-DMSO and anhydrous DMSO with small amounts of $H_2{}^{18}O$. These experiments revealed that all the three central methylene carbons come from the CH_2 carbons of the TosMIC, while the oxygen atoms come from DMSO (Figure 6b). This means that the product is obtained using at least six equivalents of isocyanide and that DMSO is also taking part in the reaction, in fact the reaction has been repeated in DMF and MeCN without success (SI-4.4). While testing the reaction in different conditions we also noticed its poor reproducibility, with significant yield fluctuations even under apparently identical conditions. Upon further investigation, it was found that the amount of trace water present in the solvent has a marked effect on the reaction profile. By testing different amounts of water, we found that the reaction does not yield any product under strictly anhydrous conditions and in presence of more than 2 equivalents of water. The best yields were obtained with 0.4 equivalents (SI-4.3). The reaction kinetics have been investigated with online HPLC, showing the formation of an intermediate after 2 hours that eventually disappears after 34 hours with the simultaneous formation of product **25** (SI-4.2). The MS analysis of the corresponding peak showed a mass consistent with compound **51**: the two-branches imine analogous to the product **25**. The chromatogram also showed the presence of high amounts of the single-branched amine **48**. This was in accordance with the data reported in Figure 5d and supported the hypothesis of a mechanism involving the formation of a central amine group that undergoes two identical semi-reactions to build the other two branches. Given this information we propose the mechanism reported in Figure 6a. The role of diethyl bromomalonate (and the other activators) is to promote the oxidation of the isocyanide group to isocyanate[27]. The formation of the central methylene carbons can be explained with the formation of imine **47** by oxidation of the amine

46[28]. **47** is then reduced to form the single branched amine **48** and attacked by it to form **49**, which undergoes an elimination of ammonia followed by a reduction to produce **51**[29]**.** The mechanism is then repeated with the consumption of another equivalent of **47** to yield the final product **25**. In addition, this mechanistic hypothesis is supported by an online IR experiment showing the presence of an isocyanate group (**43** and **45**) and CO_2 as well as the observation of dimethyl sulphide and ammonium signals in the NMR analysis of the mixture (SI-4.6-4.7). The reaction discovered has been compared with 1656 known reactions found in the *Reaxys* database[30] involving TosMIC. The RDkit[31] python library was used to extract a fingerprint difference indicating how much this new reaction transforms the structure of the input reagents compared to the output product. In order to compare fingerprint differences among various reactions the ℓ^2-norm of the fingerprint-difference vectors was calculated. Low values of the ℓ^2-norm indicate a high structural similarity between reagents and products while high values correspond to complex reactions involving several transformations. The results are shown in Figure 6c, where the ℓ^2-norm values of literature reactions are grouped into bins on a logarithmic x-axis. The ℓ^2-norm of the reaction discovered is indicated with the red line showing an unusual degree of structural change compared to known reactions. To gauge the serendipitous nature of our discovery quantitatively, we carried out a simulation to assess the size of the chemical reaction network. This network was generated by repeatedly applying a set of common reaction templates, including the ones invoked in our proposed mechanism, to generate an expanding pool of chemicals. Comparing the total size of the resulting network to the subset leading to the product gives an indication of the unpredictability of product formation. The results are shown in Figure 6d, where blue lines indicate the reaction network relevant to product formation. The overall network is much larger in size that the product derived subset (10^{10} chemicals vs. 10^5 chemicals), indicating that the observed pathway is highly improbable to guess *a priori*.

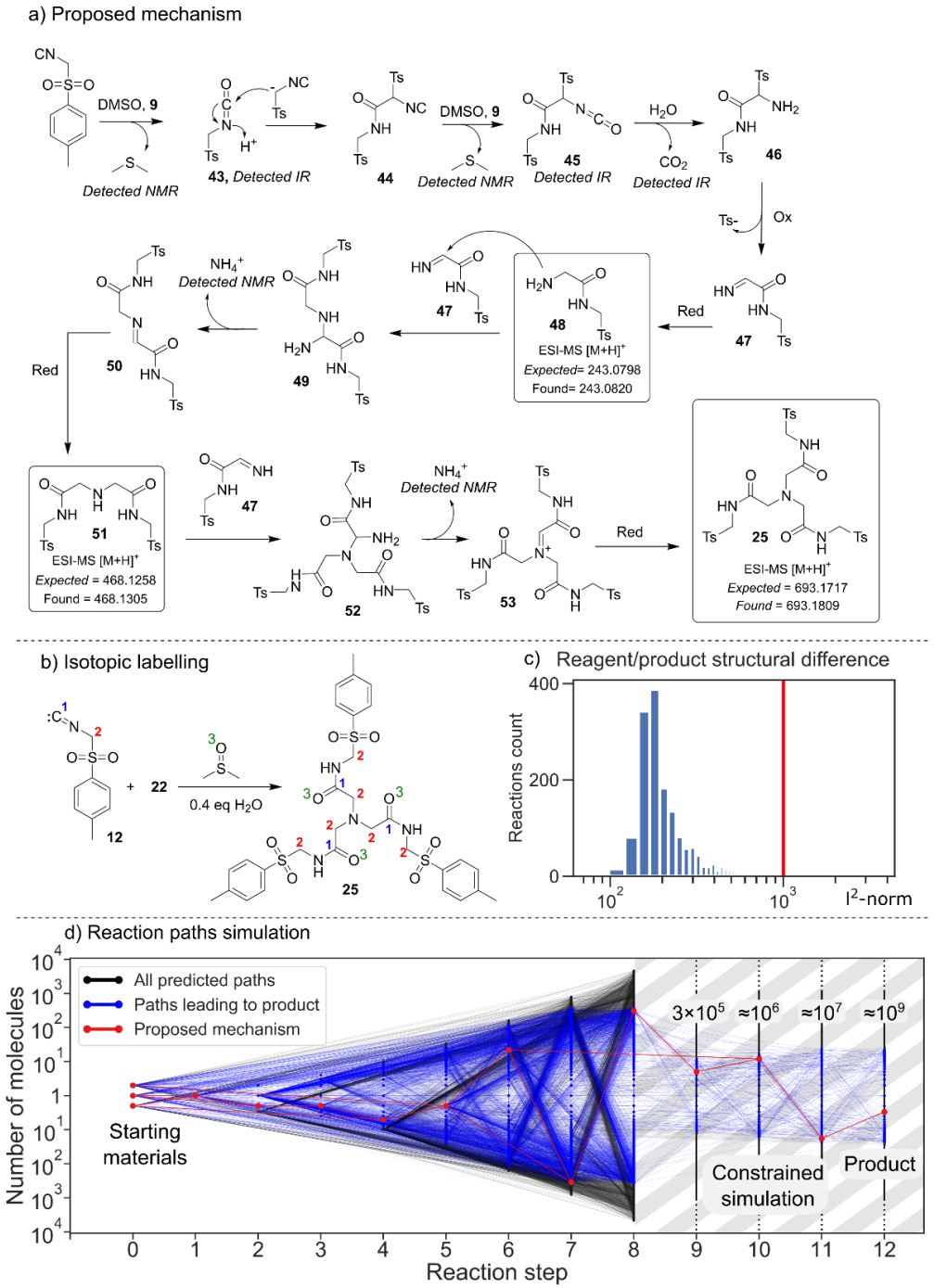


Figure 6: **Proposed mechanism of the reaction and cheminformatics plots.** (**a**) Scheme of the mechanism. Two of the intermediates have been found on the HPLC-MS analysis while isocyanate group, CO_2, DMS and ammonium have been detected at IR and NMR, respectively. (**b**) The isotopic labelling of both the interested carbons of TosMIC and the DMSO oxygen helped determine the source of product core atoms. (**c**) Comparison between the structural change between reagents and products among known reactions of TosMIC, the discovered reaction is indicated with the red line. (**d**) Estimation of reaction unpredictability by estimating the size of the relevant reaction network. A full simulation could only be carried out for the first eight steps of the simulation as the combinatorial explosion produces a vast number of molecules that could not be analysed given our computational resources.

In summary we showed that closed-loop approaches involving automatic reaction execution and assessing reactivity using machine learning can play a crucial role in the discovery of novel reactions in an unknown chemical space. Our neural network model can abstract the reactivity from the identity of the reagents, and we expect that this type of algorithm will also progressively improve in accuracy if trained on multiple chemical spaces. The continuous reactivities provided by the CNN are correlated with the reagents showing that it is possible to explore the chemical space and perform only a fraction of the total known reaction combinations. This opens the way to exploring complex chemical systems and reaction spaces using advanced machine learning techniques and reaction metrics.

Methods

General experimental remarks. Chemicals and solvents were supplied by *Fisher Chemicals*, *Sigma Aldrich, Lancaster Chemicals Ltd* and *Tokyo chemicals industry*, used as received. Deuterated solvents were obtained from *Goss Scientific Instruments Ltd.* and *Cambridge Isotope Laboratories Inc*. All commercial starting materials were used as supplied, without further purification. Off-line NMR data was recorded on a Bruker Advance 600 MHz or a Bruker Advance 400 MHz, in deuterated solvent, at T = 298 K, using TMS as the scale reference. Chemical shifts are reported using the δ-scale, referenced to the residual solvent protons in the deuterated solvent for ^{1}H and ^{13}C NMR (i. e. ^{1}H: δ (CDCl3) = 7.26; ^{13}C: δ (CDCl3) = 77.16). All chemical shifts are given in ppm and all coupling constants (*J*) are given in Hz (J) as absolute values. Characterization of spin multiplicities: s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, dd = double doublet, dt = double triplet, dq = double quartet, and ddt = double doublet of triplets. Chromatographic separation of the reaction mixture was achieved with a reverse phase column by Agilent (Poroshell 120 HPH C18, 3.0 x 100 mm, 2.7

μm) on a Thermo Fisher UltiMate 3000 HPLC. The MS apparatus was a Bruker MaXis Impact instrument, acquisition range at 50–2000 *m/z*.

Liquid handling platform. The control over the fluids was performed using C3000 model, TriContinent™ pumps (Tricontinent Ltd, CA, USA). They were equipped with distribution (3-way) and 90°/120° (2-way) valves. 5 ml syringes (TriContinent™) were used for all functions except the pumps connected to the MS instrument and the photocatalysts which used a 0.5 ml syringe. The pumps were connected to the computer and each other by a daisy chain with a RS232 serial communication cable and DA-15 connectors. The liquid connectivity was assured using PTFE tubing (1/16" 1.6mm OD x 0.8mm ID) cut to the desired length and PEEK/PTFE flangeless fittings. To perform a reaction the robot mixed 2 ml of the selected starting materials (from 1 M stock solutions) into the mixer flask and then moved the mixture into one of the six round bottom flasks (25 ml). They were placed on top of two hotplates for magnetic stirring and three of them were irradiated with visible-light LED (Thorlabs, USA). After three hours the mixture was analysed and the flask was washed with 5 ml of DMSO for three times. The software to control the platform was written in Python and was optimised to continuously run six reactions in parallel.

Benchtop NMR spectroscopy. The online NMR spectra were recorded using a Spinsolve benchtop NMR from Magritek (60 MHz). Shimming was performed before each experiment directly on the sample. The instrument was equipped with a flow-cell to allow online analysis. The cell was designed to go through the instrument and its location placed the thicker part (5 mm diameter) at the centre of the magnets. Both inlet and outlet were connected to normal PTFE tubing with screw caps (Figure S1). The flow cell allowed automatic reaction monitoring in real time by pumping 3 ml of solution from the reaction mixture. The instrument was controlled with Python through a TCP connection with the API exposed by Spinsolve software.

Benchtop MS spectroscopy. The spectra were recorded using an Advion Expression CMS equipped with an ESI (electrospray ionization) module. The mass spectrometer was controlled using a Python library created to wrap around the binary libraries supplied by Advion. Before injection the mixture was diluted by taking 0.1 ml of reaction mixture into the syringe and adding with 0.4 ml of acetonitrile. 0.4 ml of the diluted solution was pumped into waste and the process repeated five times to obtain a 10^{-4}M solution. After each injection the instrument was cleaned by flushing it with acetonitrile and a water/acetonitrile 1:1 solution.

Automatic reactivity assignment. NMR data was checked manually and a reactivity value was assigned between 0 and 3. The mixture spectrum was compared with the superimposition of the starting materials spectra and the criteria for the assignment are the appearance of new peaks, their intensity, peaks shifting and reagents peaks disappearing. Although there were borderline cases between two values some general guidelines were followed: a) absolutely no difference or a slight shift = 0, b) one peak appearing or a big shift, medium intensities = 1, c) two or three peaks appearing in high intensity = 2 and d) more than three peaks appearing with a high intensity = 3. Before the training of the neural network the NMR spectra were resampled to rescale them from 4878 to 271 points. They were then normalized to 1 and the solvent peak was removed by cutting the spectrum at 3 ppm. In order to avoid overfitting a random scaling (y-axis) and shifting (x-axis) was applied on both the mixture spectrum and the reagents superimposition during training. A 2x271 matrix obtained by the processed spectra of the mixture and the superimposition of the starting materials was used as input for the neural network. Details about the neural network architecture can be found in the SI-2.3. The network was trained on 440 reactions obtained from combinations of Chemical space 1 (Figure 3b) and used to assess its accuracy on 1018 reactions from Chemical space 2 (figure 3c).

Reactivity predictions and automatic space exploration. The reactivity data assigned by the neural network was correlated with the starting materials represented as one-hot encoded vectors using a linear regressor. The software for fitting and predicting was written using scikit-learn Python library. The algorithm was used to run a simulation of the chemical space exploration where data from the full dataset was progressively (in batches of 50 reactions) accessed following the reactivity predictions generated by the linear regressor. The simulation was written in Python.

General procedure for synthesis of product 25. Diethyl 2-bromomalonate (2 mmol, 0.41 ml), *p*-toluenesulfonylmethyl isocyanide (2 mmol, 0.39 g) and water (0.8 mmol, 15 μl) are mixed in 4 ml of anhydrous DMSO and stirred for 24 hours at 30° C. The reaction mixture is diluted with water (20:1) and extracted with ethyl acetate. The organic phase is separated and washed with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under reduced pressure. During the evaporation of ethyl acetate, the product precipitates as a white solid, it is isolated by filtration and washed with ethyl acetate.

Code availability. The code for the neural network testing, the chemical space exploration and the cheminformatics plots can be found online at https://github.com/croningp/OrganicFinder and the code is provided under the GPLv3 license.

Supplementary information is available in the online version of the paper.

Acknowledgments. The authors gratefully acknowledge financial support from the EPSRC (Grant Nos EP/S030603/1, EP/S019472/1, EP/S017046/1, EP/L015668/1, EP/L023652/1), the ERC (project 670467 SMART-POM). JMG acknowledge financial support from the Polish Ministry of Science and Higher Education grant no. 1295/MOB/IV/2015/0.

Author Contributions LC conceived the concept, the abstraction, algorithm, and the project and coordinated the efforts of the research team. DCaramelli built and coded the platform with help from AH, developed the algorithms for data analysis and manually characterised the

reactions with help from JMG and gathered data for the proposed mechanism with help from DCambie. HM and JMG did the cheminformatics analysis. LC, DCaramelli, JMG, and HM wrote the manuscript with input from all the authors.

References

1. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 doi: 10.1038/nchem.136 (2009)
2. Oeschger, R. et al. Diverse functionalization of strong alkyl C – H bonds by undirected borylation, *Science*, **3**, 736–741 doi: 10.1126/science.aba6146 (2020).
3. Reymond, J. L., Ruddigkeit, L., Blum, L. & van Deursen, R. The enumeration of chemical space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 717–733 doi: 10.1002/wcms.1104 (2012).
4. Herges, R. Reaction planning: prediction of new organic reactions. *J. Chem. Inf. Comput. Sci.* **30**, 377–383 doi: 10.1021/ci00068a006 (1990)
5. Suleimanov, Y. V. & Green, W. H. Automated Discovery of Elementary Chemical Reaction Steps Using Freezing String and Berny Optimization Methods. *J. Chem. Theory Comput.* **11**, 4248–4259 doi: 10.1021/acs.jctc.5b00407 (2015)
6. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 doi: 10.1021/acscentsci.7b00512 (2018)
7. Baskin, I. I., Madzhidov, T. I., Antipin, I. S. & Varnek, A. A. Artificial intelligence in synthetic chemistry: achievements and prospects. *Russ. Chem. Rev.* **86**, 1127–1156 doi: 10.1070/rcr4746 (2017)

8. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **51**, 1281–1289 doi: 10.1021/acs.accounts.8b00087 (2018)
9. Katsila, T., Spyroulias, G. A., Patrinos, G. P. & Matsoukas, M. T. Computational approaches in target identification and drug discovery. *Comput. Struct. Biotechnol. J.* **14**, 177–184 doi: 10.1016/j.csbj.2016.04.004 (2016).
10. Tabor, D. P. *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 doi: 10.1038/s41578-018-0005-z (2018).
11. Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L., How to explore chemical space using algorithms and automation, *Nat. Rev. Chem.*, **3**, 119–128, doi: 10.1038/s41570-018-0066-y (2019)
12. Granda, J.M.; Donina, L.; Dragone, V.; Long, D.L.; Cronin, L., Controlling an organic synthesis robot with machine learning to search for new reactivity, *Nature*, **559** (7714), 377-381, doi: 10.1038/s41586-018-0307-8 (2018)
13. Burke, M. D. & Lalic, G. Teaching target-oriented and diversity-oriented organic synthesis at Harvard University. *Chem. Biol.* **9**, 535–541 doi: 10.1016/S1074-5521(02)00143-6 (2002).
14. Mcnally, A.; Prier, C. K.; Macmillan, D. W. C. Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity, *Science*, **334**, 6059, 1114, doi: 10.1126/science.1213920 (2011)
15. Houben, C.; Lapkin, A. A. Automatic discovery and optimization of chemical processes. *Curr. Opin. Chem. Eng.*, **9**, 1, doi: 10.1016/j.coche.2015.07.001 (2015)
16. Gajewska, E. P. *et al.* Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem* **6**, 280–293 doi: 10.1016/j.chempr.2019.11.016 (2020).

17. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 doi: 10.1039/c9sc05704h (2020).

18. Mennen, S. M. *et al.* The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **23**, 1213–1242 doi: 10.1021/acs.oprd.9b00140 (2019).

19. Poscharny, K. *et al.* Machine assisted reaction optimization: A self-optimizing reactor system for continuous-flow photochemical reactions. *Tetrahedron* **74**, 3171–3175 doi: 10.1016/j.tet.2018.04.019 (2018)

20. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **4**, 1134–1145 doi: 10.1021/acscentsci.8b00307 (2018).

21. Montgomery, J. High-Throughput Discovery of New Chemical Reactions, *Science*, **333**, 1387–1389, doi: 10.1126/science.1210735 (2011)

22. Perera, D. *et al.* A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science.* **359**, 429–434 doi: 10.1126/science.aap9112 (2018)

23. Richmond, C. J. *et al.* A flow-system array for the discovery and scale up of inorganic clusters. *Nat. Chem.* **4**, 1037–1043 doi: 10.1038/nchem.1489 (2012)

24. Schweidtmann, A. M. et al. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chem. Eng. J.* **352**, 277–282, doi: 10.1016/j.cej.2018.07.031 (2018)

25. Rebrov, E.V; Expósito, A.J., OpenFlowChem – a platform for quick, robust and flexible automation and self-optimisation of flow chemistry, *React. Chem. Eng*, **3**, 769–780, doi: 10.1039/c8re00046h (2018)

26. Dragone, V.; Sans, V.; Henson, A. B.; Granda, J. M.; Cronin, L., An autonomous organic reaction search engine for chemical reactivity, *Nat. Commun*., **8**, 15733, doi: 10.1038/ncomms15733 (2017)
27. Le, H. V.; Ganem, B., Trifluoroacetic anhydride-catalyzed oxidation of isonitriles by DMSO: A rapid, convenient synthesis of isocyanates, *Org. Lett.*, **13** (10), 2584-2585, doi: 10.1021/ol200695y (2011).
28. van Leusen, A. M., Wildeman, J. & Oldenziel, O. H. Base-Induced Cycloaddition of Sulfonylmethyl Isocyanides to C, N Double Bonds. Synthesis of 1, 5-Disubstituted and 1, 4, 5-Trisubstituted Imidazoles from Aldimines and Imidoyl Chlorides. *J. Org. Chem.* **42**, 1153–1159 doi: 10.1021/jo00427a012 (1977).
29. Guérin, C.; Bellosta, V.; Guillamot, G.; Cossy, Mild nonepimerizing N -alkylation of amines by alcohols without transition metals, *J. Org. Lett*., **13** (13), 3534-3537, doi: 10.1021/ol201351a (2011)
30. Reaxys. https://new.reaxys.com/ (accessed on May 18, 2020)
31. www.rdkit.org (accessed on May 28, 2020)

Supplementary information

An Artificial Intelligence that Discovers Unpredictable Chemical Reactions

Dario Caramelli, Jaroslaw M. Granda, Dario Cambie, S. Hessam M. Mehr, Alon Henson, Leroy Cronin*

School of Chemistry, The University of Glasgow, Glasgow G12 8QQ (UK)

Web: www.croninlab.com Email: Lee.Cronin@Glasgow.ac.uk

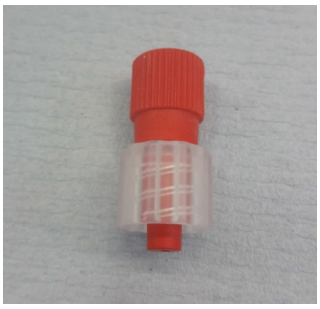
Code availability. The code for the neural network testing, the chemical space exploration and the cheminformatics plots can be found online at https://github.com/croningp/OrganicFinder and the code is provided under the GPLv3 license.

CONTENTS

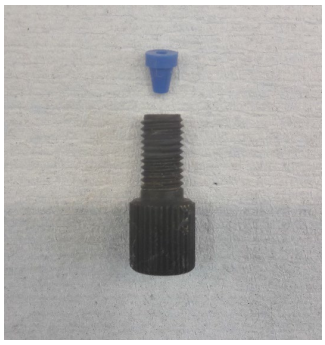
1 Hardware specification 4
1.1 Benchtop MS 6
1.2 Benchtop NMR 7
1.3 Syringe pumps 8
1.4 Building the platform 9
1.5 The chemical space 11
1.6 The light shield 11
2 Neural network for reactivity assignment 13
2.1 Manual reactivity assignment 13
2.2 NMR spectra pre-processing 15
2.3 Neural network structure 16
3 Reaction discovered with the platform and variations 17
3.1 Diethyl 2-bromomalonate and p-toluenesulfonylmethyl isocyanide. 17
3.2 1H-benzotriazol-1-ylmethyl isocyanide and diethyl 2-bromomalonate 18
3.3 Silyl isocyanide and diethyl 2-bromomalonate 19
3.4 Variations of activators 20
3.5 Asymmetric variations of product 25 22
3.6 Other Isocyanide variations 24
4 Reaction mechanism data 25
4.1 HPLC-analysis 25
4.2 Time resolved HPLC analysis 28
4.3 Water influence 31
4.4 Reaction performed in different solvents 32
4.5 Synthesis of isotopically substituted starting materials 34
4.5.1 Synthesis of (^{13}C)TosMIC 34
4.5.2 Isocyanide Substitution 34

4.5.3 Methylene Labelling 37
4.5.4 DMSO Labelling 40
4.5.5 Reaction in $^{18}OH_2$ 42
4.6 IR reaction monitoring. 45
4.7 Online NMR monitoring 47
5 Cheminformatics simulation 49
5.1 Reaction network 49
5.2 Similarity index with known TosMIC reactions 50
6 Other reactions discovered and re-discovered 52
6.1 Phenylhydrazine and bromoacetonitrile under 450 nm irradiation 53
6.2 Diethyl 2-bromomalonate and 1-vinyl-2-pyrrolidinone under 450 nm irradiation 54
6.3 *N,N*-dimethyl-4-nitrosoaniline, bromoacetonitrile and diethyl 2-bromomalonate 55
6.4 Diethyl 2-bromomalonate, 4-phenylurazole and 1-vinyl-2-pyrrolidinone 56
6.5 trans,trans-2,4-Hexadienal dimerization reaction and reduction **Error! Bookmark not defined.**
7 Crystal structure details 57
8 ^{1}H and ^{13}C NMR spectra 60

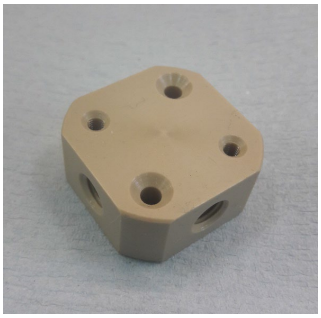
1 Hardware specification



Adapter, Luer (Male) to 1/4"-28 Flat Bottom (Female), ETFE/Polypropylene P-675 (IDEX-HS P-675)



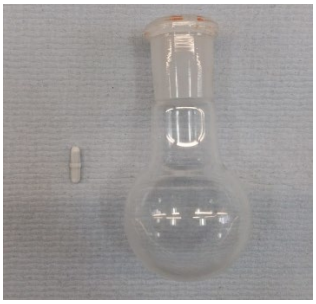
Flangeless Fitting, for 1/16" OD Tubing, 1/4"-28 Flat Bottom, PEEK/ETFE (IDEX-HS XP-218BLK)



Block Connector, 3-Way, 1/4"-28 (Flat Bottom), PEEK, with mounting holes (Diba Dibafit 001057)

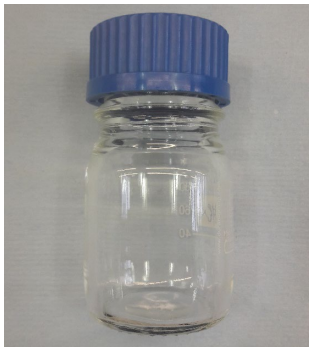
Needles, Sterican, Long Length 120mm

Tubing, PTFE, 1/16" (1.6mm) OD x 0.8mm ID, 20m Part No. 008T16-080-20 (Kinesis)



Round bottom flasks, 25 ml. (Fisher)

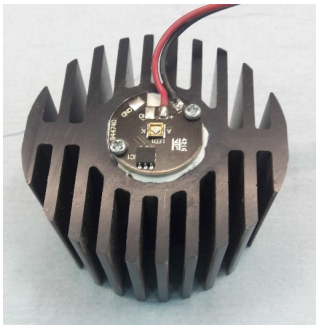
Stirrer bars, PTFE coated, 13x3 mm (Fisher)



Duran® laboratory bottles, with caps capacity 100 mL, blue PP screw cap and pouring ring Sigma-Aldrich Z305170-10EA and Z305200-10EA 18 x 50-250 mL for reagents 2 x 1 L for solvent and waste

C-Series Syringe Pumps Tricontinent C3000 with 5.0 and 0.5 mL syringes, 4-way non-distribution and 4-way distribution valves.

7 x Suba-Seal® septa red rubber, Suba Seal, 33, neck I.D., 17.5 mm, Z124613-100EA (Sigma-Aldrich)



Visible light LED:
Thorlabs
-M565D2, 565nm, 880mW
-M450D3, 450nm, 1850mW
-M405D2, 405nm, 1500mW
Heatsink: 1.8K/W, 60x37.5mm, RS components, 722-6795

1.1 Benchtop MS

Advion expression-CMS

- Dimensions: 66 x 28 x 56 cm
- Weight: 32 kg
- Gas requirements: Nitrogen 98% pure, 4.1 Bar, 8 L/min
- Flow rate range: 10 µL/min to 2 mL/min
- Polarity: Positive & Negative ion switching in single analysis
- *m/z* Range: 10 to 2,000 *m/z*
- Resolution: 0.5 - 0.7 *m/z* units (FWHM) at 1000 *m/z* units sec^{-1} over entire acquisition range
- Accuracy: ± 0.1 *m/z* units over entire acquisition range
- Linear dynamic range of 5 x 10^3

ESI parameters

- Capillary temperature(V) 250.0
- Capillary Voltage(V) 180.0
- ESI Gas Temperature: 250 °C
- ESI Voltage: 3500 V
- Calibration: Agilent ESI Tuning mix G2421A

1.2 Benchtop NMR

Spinsolve 60 Carbon from Magritek

- Frequency: 60 MHz Proton
- Resolution: 50% linewidth < 0.5 Hz
- Lineshape: 0.55% linewidth < 20 Hz
- ^{1}H Sensitivity: >120:1 for 1% ethyl benzene
- Dimensions: 58 x 43 x 40 cm
- Weight: 60 kg
- Magnet: Permanent and cryogen free
- Stray field: < 2 G all around system

The instrument is equipped with a flow-cell to allow online analysis. The cell goes through the instrument and its location places the NMR tube part at the centre of the magnets. Both inlet and outlet are connected to normal PTFE tubing with screw caps (**Figure S1**). The flow cell allows automatic reaction monitoring in real time by pumping 3 ml of solution from the reaction mixture.

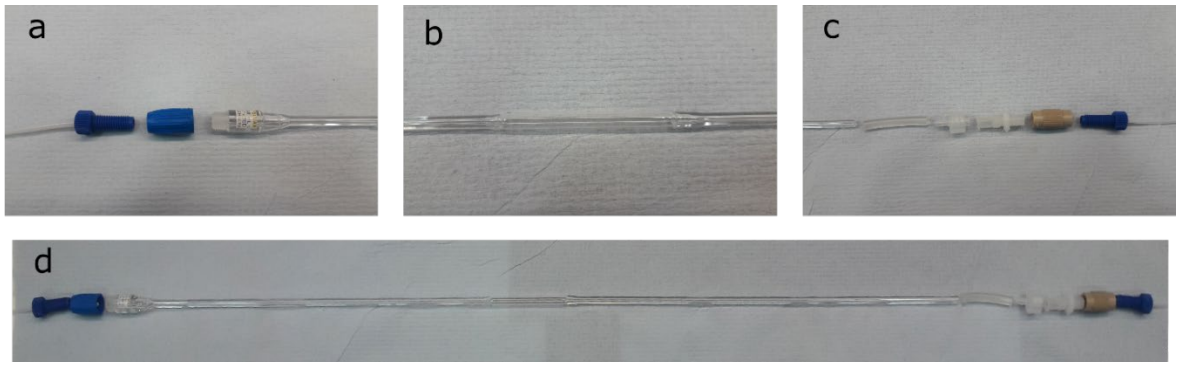


Figure S1: a) top connector, the screw connector has been welded to the glass cell by the glassblower. b) Central part, it has been made by welding a normal 5 mm NMR tube to the rest of the cell. c) bottom connection, the glass could not have the threading because the whole cell needs to fit through the instrument. The connection is made with a chemical resistant tubing followed to a plug to syringe, a syringe to male connector, female to female and finally a normal screwed connector. d) entire NMR flow cell

1.3 Syringe pumps

The control over the fluids was performed using C3000 model, TriContinent™ pumps (Tricontinent Ltd, CA, USA). 5 ml syringes (TriContinent™) were used for all functions except the pumps connected to the MS instrument and the photocatalysts which used a 0.5 ml syringe. The pumps were connected to the computer and each other by a daisy chain with a RS232 serial communication cable and DA-15 connectors. Up to 15 pumps can be connected on the same line and addresses are selected with a physical switch on the back (positions 0 to E, F is used for debugging) (**Figure S2**). Our project used two serial lines for a total of 30 pumps.

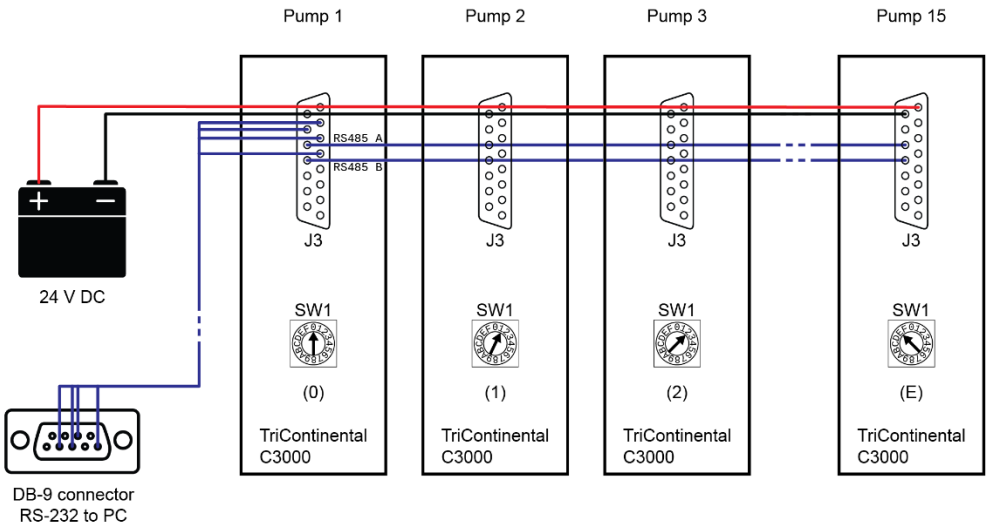


Figure S2: Schematic representation of the daisy chain connection used to control and power the pumps. On a single line up to 15 pumps can be connected, the individual addressed are defined with a physical switch on their back with positions between 0 and E. F position is used for debugging.

The pumps were equipped with distribution (4-port), 90° and 120° (3-port) valves. The distribution valve allows a 3-way connectivity (is it possible to pump a solution from the syringe through one of three ports). 90° and 120° valves allow 2-way connectivity (**Figure S3**).

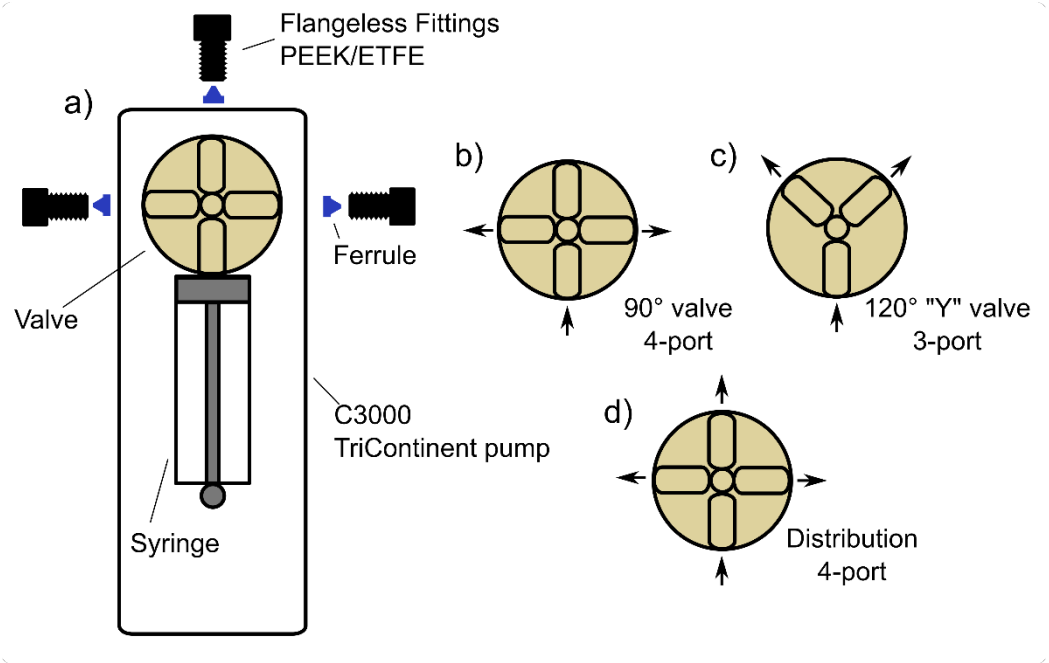


Figure S3: a) Schematic view of the front of a C3000 TriContinent pump. b) Scheme of a 90° valve, the top port is used to bypass the pump, therefore it is possible to pump from the syringe only in two directions. c) Scheme of the 120° valve, it can pump in two directions. d) Scheme of the Distribution valve, it can pump in three directions.

1.4 Building the platform

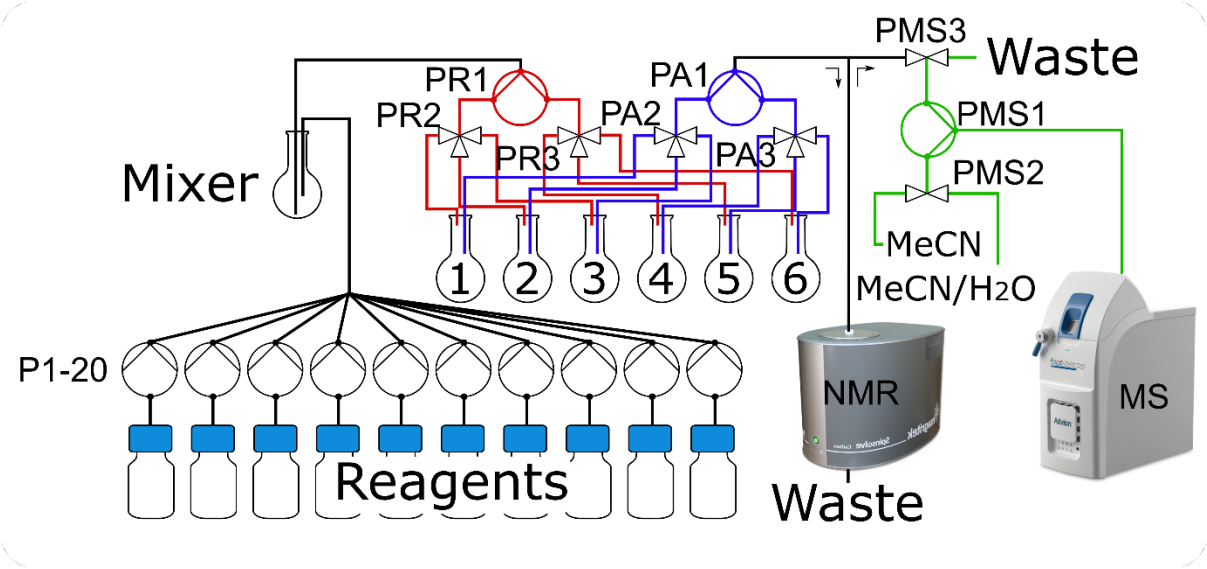


Figure S4: Scheme of the platform including pumps labels.

All fluidic connections to the pumps are made with PTFE, (1/16" OD x 0.8mm ID) tubing and PEEK/ETFE flangeless Fitting (for 1/16" OD Tubing, 1/4"-28 Flat Bottom).

Reagents pumps (P1-20): These pumps are responsible of the addition of reagents, additives and photocatalysts into the mixer flask. Each pump is assigned to a single reagent bottle. They are equipped with 90° or 120° valves. Each pump is connected on one port to a single reagent bottle and on the other to the mixer flask, this connection ends with a Luer to 1/4"-28 Flat Bottom adapter and a long needle in order to go through the septum on the mixer flask.

Reactor pumps (PR1-3): These three pumps manage the transfer of the reaction solution from the mixer flask to one of the six reactors. They are all equipped with 4-way distribution valves and two of them (PR2 and PR3) are used without the syringe to expand the number of ports of the main one (PR1). PR1 is connected from the top port to the mixer flask with a Luer to 1/4"-28 Flat Bottom adapter and a long needle. The other two ports are connected to the syringe ports of PR2 and PR3. The other three ports of PR2 and PR3 are connected to the six reactors with a Luer to 1/4"-28 Flat Bottom adapter and a long needle.

Analysis pumps (PA1-3): These pumps are assembled in the same way of PR1-2-3. They all have 4-way distribution valves. PA2 and PA3 are connected to the six reactors with the three main ports while the syringe ports are connected to the left and right ports of PR1. The top port of PR1 is connected to a 3-way block connector.

3-way block connector: This connects PA1 to the NMR flow cell and PMS3 it is a simple T-splitter.

MS pumps (PMS1-3): They have the same configuration of the reactors pumps. PMS2 and PMS3 have 90° valves and are connected to PMS1 from the syringe ports. PMS3 is connected to the 3-way block connector and to the waste tank while PMS2 is connected to a bottle of acetonitrile (used for dilution) and a bottle containing a mixture of acetonitrile/water 50:50 (used to clean the MS). PMS1 is equipped with a 0.5 ml syringe and is connected through the top port to the MS instrument.

Analysis routine: 4 ml of the reaction mixture is pumped from one of the reactor flasks through the 3-way connector block using PA1 and either PA2 or PA3. Since PMS3 is blocking one of the connector lines this movement results in the solution flowing into the NMR cell. Then PMS1 pumps 0.1 ml of solution from the connector block into its syringe. Again, since PA1 seals the other connection, this movement successfully pumps back the solution from the NMR experiment into PMS1. Finally, the mixture is diluted and pumped to the MS. Both analyses are then started at the same time.

1.5 The chemical space

In the Chemical space 1 explored molecules were selected from a pool of 6 starting materials and mixed in combinations of two and three (**Figure 3b**). 2 ml of each reagent was added from a 1M stock solution. In Chemical space 2 reagents were added in the same way from a pool of 15 compounds (**Figure 3c**). Each reaction had the additional selection of an additive: either a base (4-Dimethylaminopyridine, **23**) or a Lewis acid (scandium triflate, **24**). Furthermore the chemical space involved the presence of one of three molecules known to act as photocatalysts: 2,4,6-triphenylpyrylium tetrafluoroborate (**PC1**), tris(2,2′-bipyridyl)dichlororuthenium(II) hexahydrate (**PC2**) and rose bengal (**PC3**) (**Figure S5**). They were added in 2.5% mol to the reactors irradiated a wavelength corresponding to photocatalyst absorption (405 nm for **PC1**, 450 nm for **PC2** and 565 nm for **PC3**).

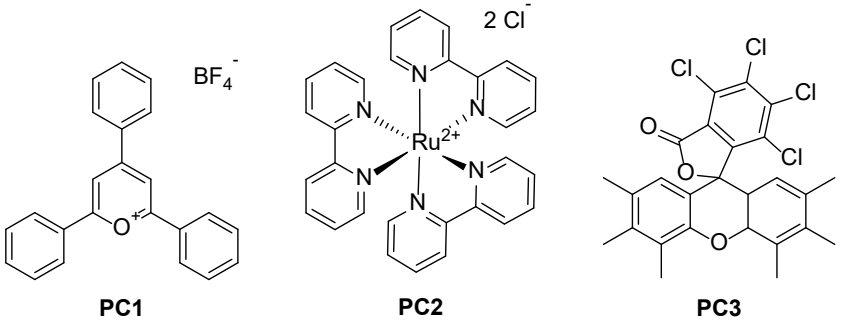


Figure S5: Photocatalysts used as additives to expand the chemical space.

1.6 The light shield

A shielding apparatus was designed and build in order to provide eye protection for the user and to isolate the three reactors from the light of each other. All parts were designed in Lightwave3D and printed on a Stratasys Connex 500 polyjet 3D printer. The apparatus was made of five parts showed as 3D models in **Figure S6**. The base (**c**) is placed on the heating plate for stirring and is furnished with a gap to sledge in the internal walls (**d**). Two holes have also been drilled in the back of the base to accommodate the back wall (**e**). Three heatsinks with the relative LEDs are placed on top of the base separated by the internal walls. The front shield (**a, b**) is then placed on top of the heatsinks. All parts were modelled around the measurements of clamps, heatsinks, and flasks, therefore the shield gently bends around the light emitters avoiding the light to shine outside or irradiate the wrong reactor. A designed cover (**f, g**) is placed on top of the front shield in order to block the light coming out from the top, as a result of a 0.5 cm spline it can sit firmly on top of the front shield.

All parts were printed in a polymer material called *verowhite;* a photosensitive polymer liquid that is solidified by UV light layer by layer. Since the final object presented a slight transparency, the internal part of the front shield was covered with aluminium foil. This solution assures that the LEDs light will reflect back into the reactors (**Figure S7**).

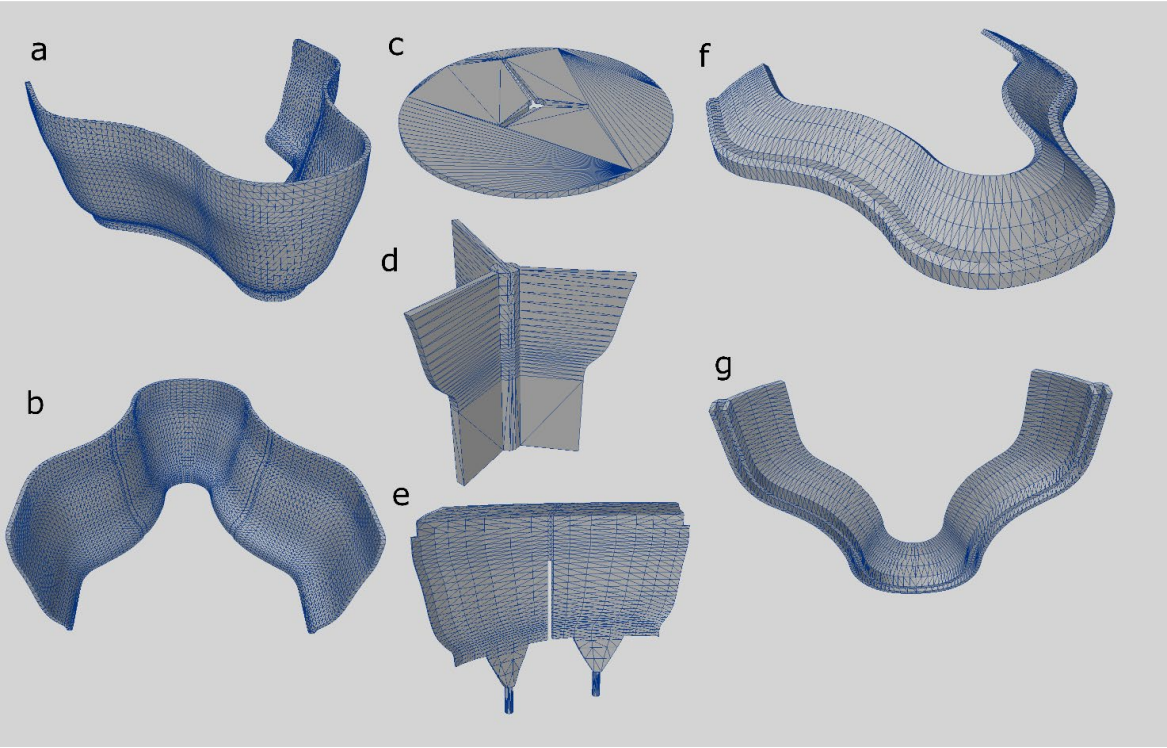


Figure S6: a,b: front shield, the internal part was covered with aluminium foil. **c**: base of the apparatus, a narrow gap in the centre allows to slot in the internal walls. **d**: internal walls, they isolate the three reactors from the unwanted LEDs light. **e**: backwall, it confines the light coming out from the back. **f, g**: cover lid, it blocks the light coming from the top.

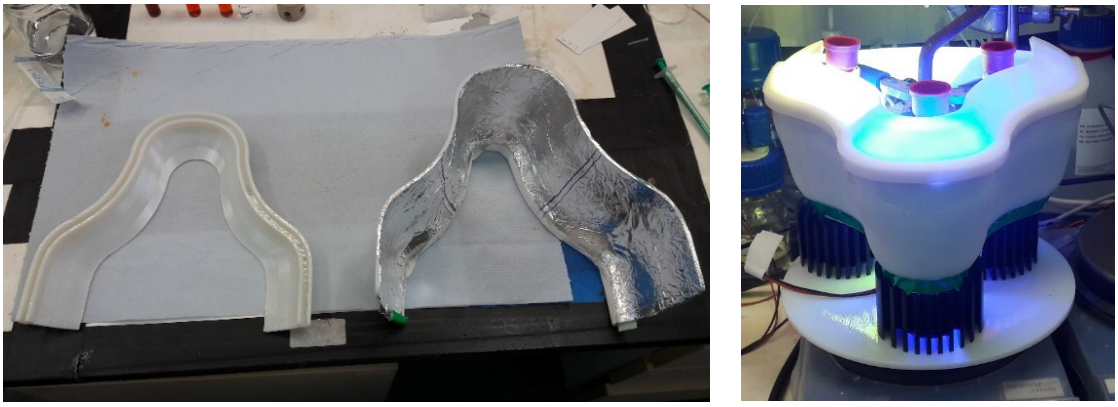


Figure S7: *Left*: front shield covered with aluminium foil and its cover. *Right*: Full setup. The LEDs are mounted on the heatsinks and placed on a stirring plate. Reactors are held on top and isolated by the 3D printed shield.

2 Neural network for reactivity assignment

2.1 Manual reactivity assignment

NMR data was checked manually and a reactivity value was assigned between 0 and 3. To do this the mixture spectrum was compared with the superimposition of the starting materials' spectra and the criteria for the assignment are based on the appearance of new peaks, their intensity, peaks shifting and reagents peaks disappearing. Although there were borderline cases between two values some general guidelines were followed: a) absolutely no difference or a slight shift = 0; b) one peak appearing or a big shift, medium intensities = 1; c) two or three peaks appearing in high intensity = 2 and d) more than three peaks appearing with a high intensity = 3. Examples of real NMR data with their manual evaluation are reported in **Figure S8**.

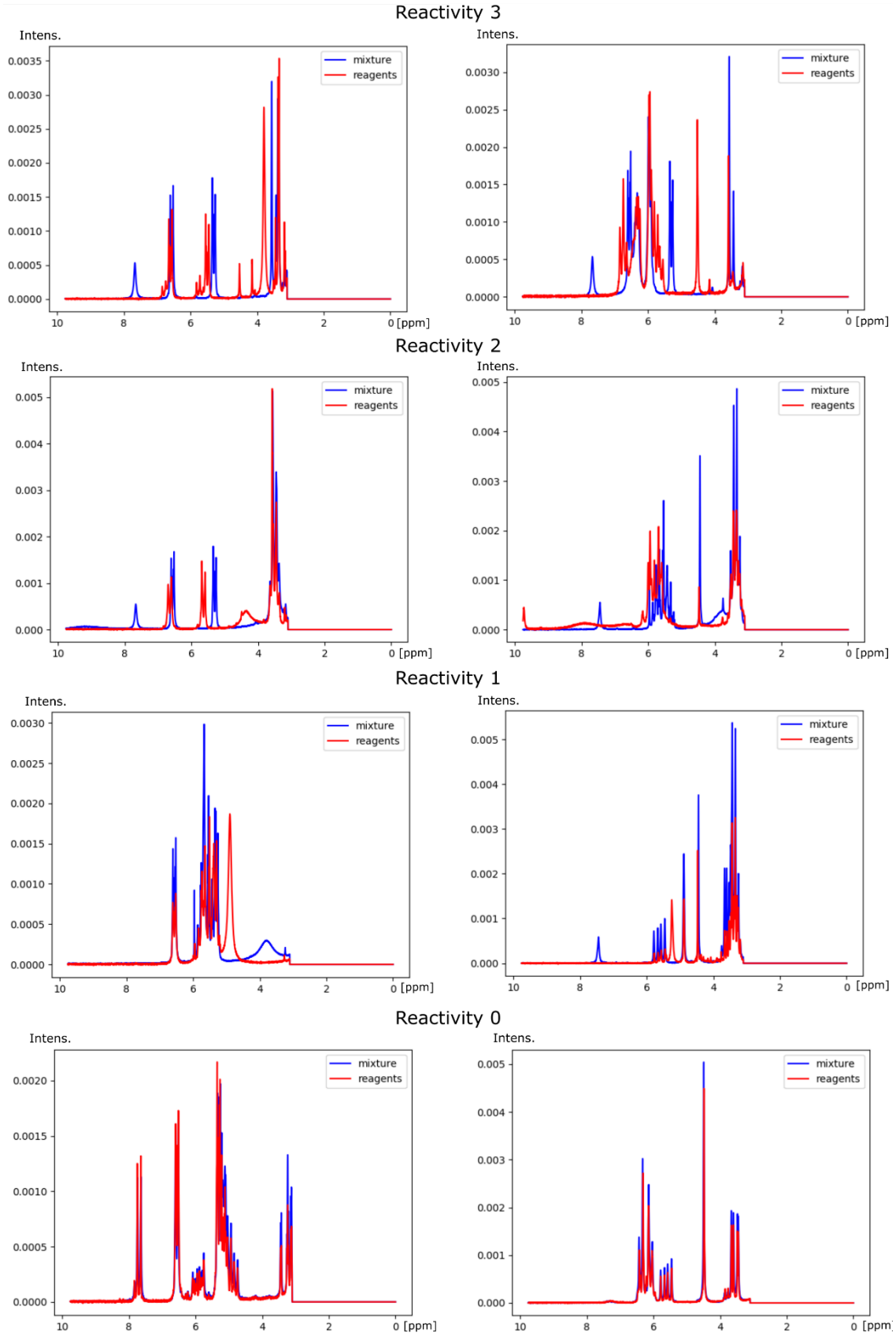


Figure S8: Examples of NMR spectra evaluated manually. By looking at the new peaks appearing and the reagents peaks disappearing a reactivity between 0 and 3 is assigned.

2.2 NMR spectra pre-processing

NMR spectra were resampled to rescale them from 4878 to 271 points. They were then normalized to 1 and the solvent peak was removed by cutting the spectrum at 3 ppm. In order to avoid overfitting a random scaling (y-axis) and shifting (x-axis) was applied on both the mixture spectrum and the reagents superimposition during training (**Figure S9**). True values classes were normalized to 1, meaning that original values from 0 to 3 corresponded to 0, 0.33, 0.66 and 1.

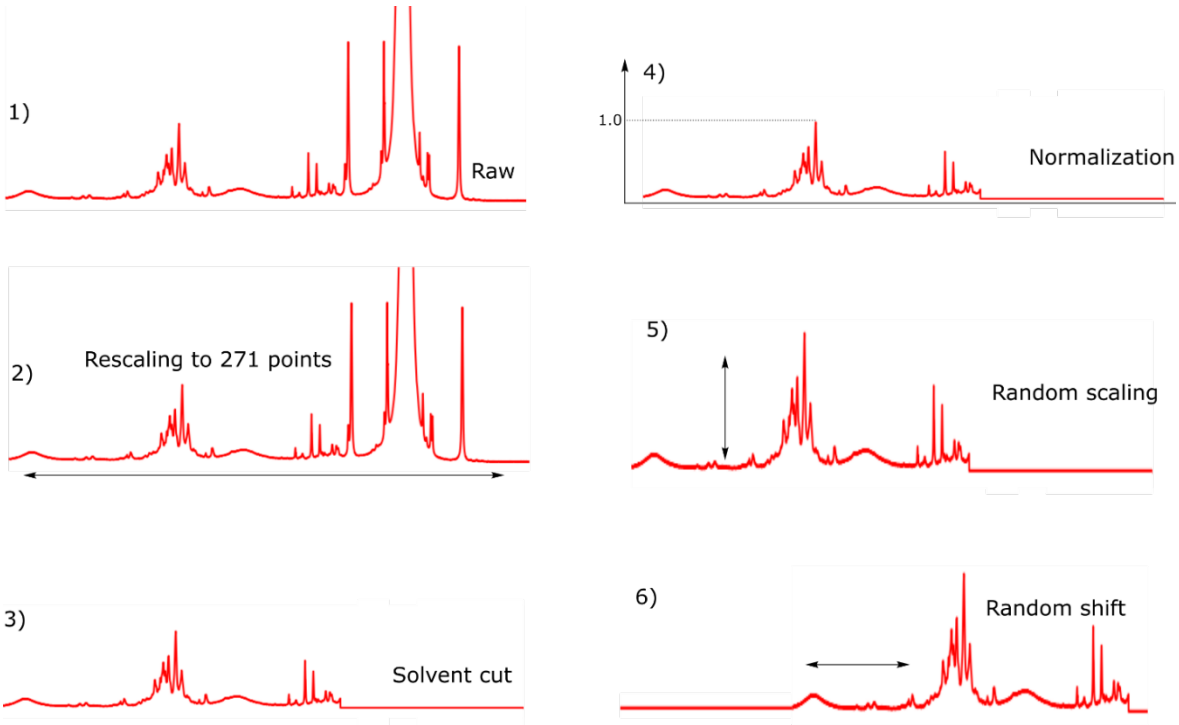


Figure S9: Pre-processing performed on the NMR spectra before using them as input. The raw spectrum (1) is rescaled to 271 points (2) and the solvent is cut out (3). It is then normalized to the highest peak (4) and a random factor on both axes is introduced during the network training to avoid overfitting (5,6)

2.3 Neural network structure

The dataset was fed into the network and each experiment was represented as a 271 x 2 matrix where the first row corresponded to the intensity values of the mixture spectrum while the second one the reconstructed spectrum obtained by superimposition the starting materials spectra. The first layer was made of 3 parallel 2D convolutional layers, they all have a pooling function set to 2 point. They produce 135 values (half the spectrum length due to pooling) for 16 channels as output. The layers differ in the filter applied: 2 x 2, 2 x 10 and 2 x 20 points. After concatenating the output, a second 2D convolutional layers with 48 values as an input and 1x1 filter is connected, it does not use pooling therefore it outputs 135 values, on 8 channels. They are fed into a 512 neurons dense layer whose output is converted into logits and predictions by the last dense layer. The parameters of the filters in the convolutional layers and the size of layer 3 were found with a random grid search, where different combinations were tried during training. The network was trained with 300 epochs and early stopping to avoid overfitting. Training data consisted in 440 NMR spectra from the 6-reagents space, split into training/validation as 0.9/0.1. Loss was calculated with cross entropy function and minimized using the AdamOptimizer algorithm[32]. The network code is based on Tensorflow[33] python library.

3 Reaction discovered with the platform and variations

3.1 Diethyl 2-bromomalonate and p-toluenesulfonylmethyl isocyanide.

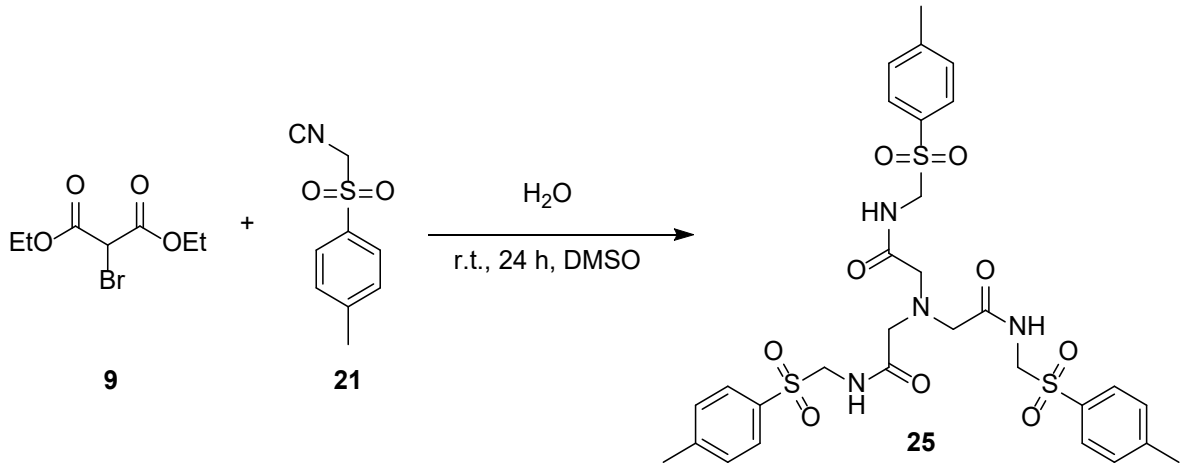


Figure S10: Scheme of the reaction yielding product **25**

Diethyl 2-bromomalonate (2 mmol, 0.41 ml), p-toluenesulfonylmethyl isocyanide (2 mmol, 0.39 g) and water (0.8 mmol, 15 µl) are mixed in 4 ml of anhydrous DMSO and stirred for 24 hours at room temperature. The reaction mixture is diluted with water (20:1) and extracted with ethyl acetate. The organic phase is separated and washed with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under reduced pressure. During the evaporation of ethyl acetate, the product precipitates as a white solid, it is isolated by filtration and washed with ethyl acetate.

IUPAC name: N,N,N-tris(2-(4-methylphenylsulfonyl)acetamide)amine

Yield: 47% (108 mg)

^{1}H NMR (600 MHz, DMSO-d_6) δ 8.96 (t, J = 6.7 Hz, 1H), 7.68 (d, J = 8.2 Hz, 2H), 7.39 (d, J = 8.0 Hz, 2H), 4.66 (d, J = 6.7 Hz, 2H), 2.97 (s, 2H), 2.38 (s, 3H).

^{13}C NMR (151 MHz, DMSO) δ 170.04, 144.69, 134.39, 129.73, 128.48, 60.00, 56.48, 21.07.

ESI-HR-MS: $[C_{30}H_{36}N_4O_9S_3Na]^+$ Calculated 715.1537 *m/z*, measured 715.1503 *m/z*

3.2 1H-benzotriazol-1-ylmethyl isocyanide and diethyl 2-bromomalonate

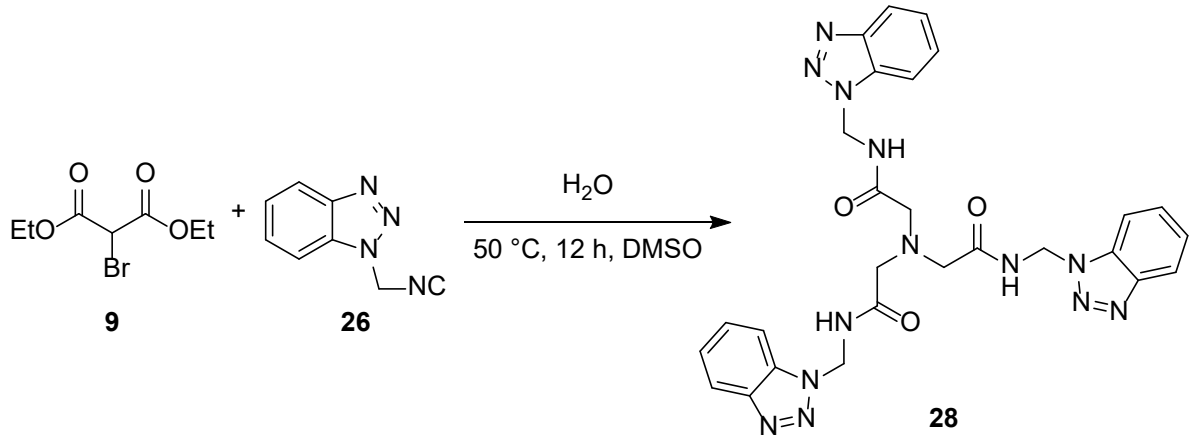


Figure S11: Scheme of the reaction yielding product **28**

Diethyl 2-bromomalonate (1.89 mmol, 0.32 ml), 1H-benzotriazol-1-ylmethyl isocyanide (1.89 mmol, 0.3 g) and water (water (0.76 mmol, 14 µl)) are mixed in 4 ml of anhydrous DMSO and stirred at 50 °C overnight. The reaction mixture is diluted with water (20:1) and extracted with ethyl acetate. The organic phase is separated and washed with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under vacuum. The crude is purified with a (silica gel) chromatographic column, gradient elution was used: EtOAc/hexane 1:1 - EtOAc 100% - EtOAc /methanol 24:1.

IUPAC name: N,N,N-tris(N-(Benzotriazol-1-ylmethyl)acetamide)amine
Yield: 31% (57 mg)
^{1}H NMR (600 MHz, $CDCl_3$) δ 8.93 (s, 3H), 7.77 (dd, *J* = 36.8, 8.4 Hz, 6H), 7.38 (t, *J* = 7.7 Hz, 4H), 7.23 (t, *J* = 7.7 Hz, 3H), 5.81 (d, *J* = 6.6 Hz, 6H), 3.39 (s, 6H).
^{13}C NMR (151 MHz, $CDCl_3$) δ 133.79, 107.78, 94.54, 90.27, 86.75, 81.52, 72.87, 21.40, 12.93.
ESI-HR-MS: $[C_{27}H_{27}N_{13}NaO_3]^+$ calculated 604.2252 m/z, measured 604.2232 *m/z*.

3.3 Silyl isocyanide and diethyl 2-bromomalonate

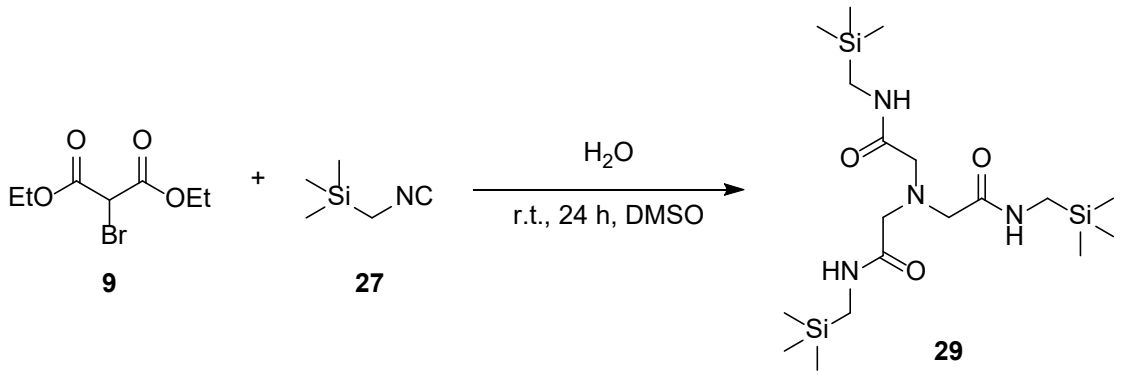


Figure S12: Scheme of the reaction yielding product **29.** The product is only observed in traces in HPLC-MS.

Diethyl 2-bromomalonate (2 mmol, 0.41 ml) and (Trimethylsilyl)methyl isocyanide (2 mmol, 0.39 g) are mixed in 4 ml of DMSO and stirred for 24 hours at room temperature. It was not possible to isolate the product. However HPLC-MS analysis of the mixture showed traces of a molecule with mass 447.2696, the expected mass for $[M+H]^+$ is 447.2623.

3.4 Variations of activators

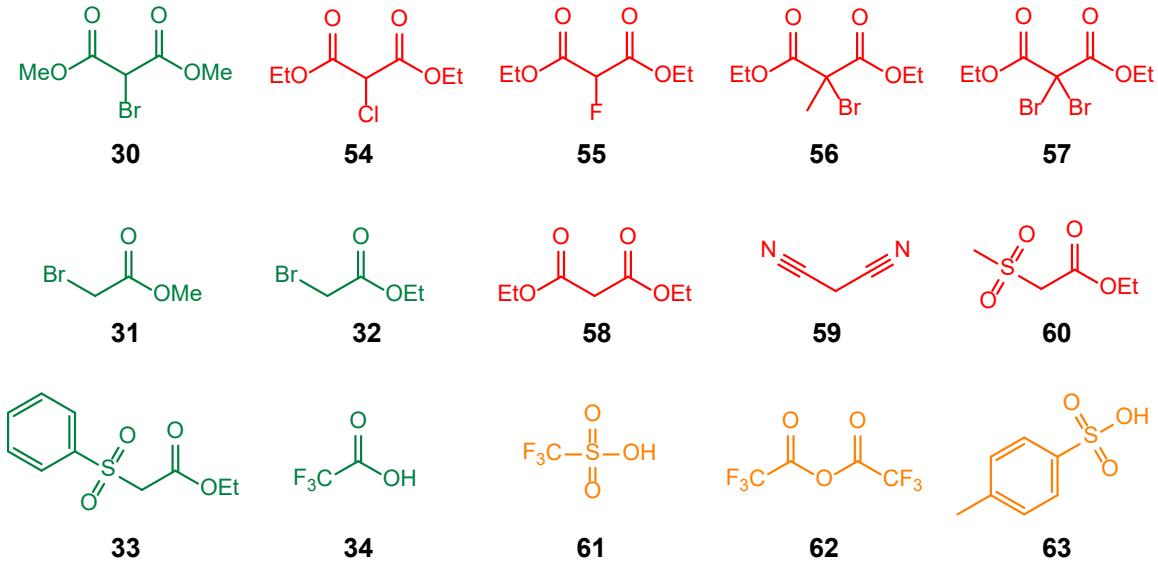


Figure S13: List of molecules tried as activator replacing diethyl bromo malonate.

The reaction has been repeated according to the conditions indicated in section 3.1, replacing the diethyl bromomalonate with various analogous molecules and organic acids. The molecules coloured in green (**30**, **31**, **32**, **33** and **34**), yielded product **25** as white precipitate. The molecules coloured in red (**54, 55, 56, 57, 58, 59, 60**) did not produce the desired product. In the reaction with molecule **61**, **62** and **63** the product was observed in the HPLC analysis but did not precipitate. The reaction with molecule **63** produced molecule **64**, while molecule **34** produced a mixture of product **25** and **65** (**Figure S14**)**.** The synthesis of molecules **69** and **70** from TosMIC as already been reported in literature[34].

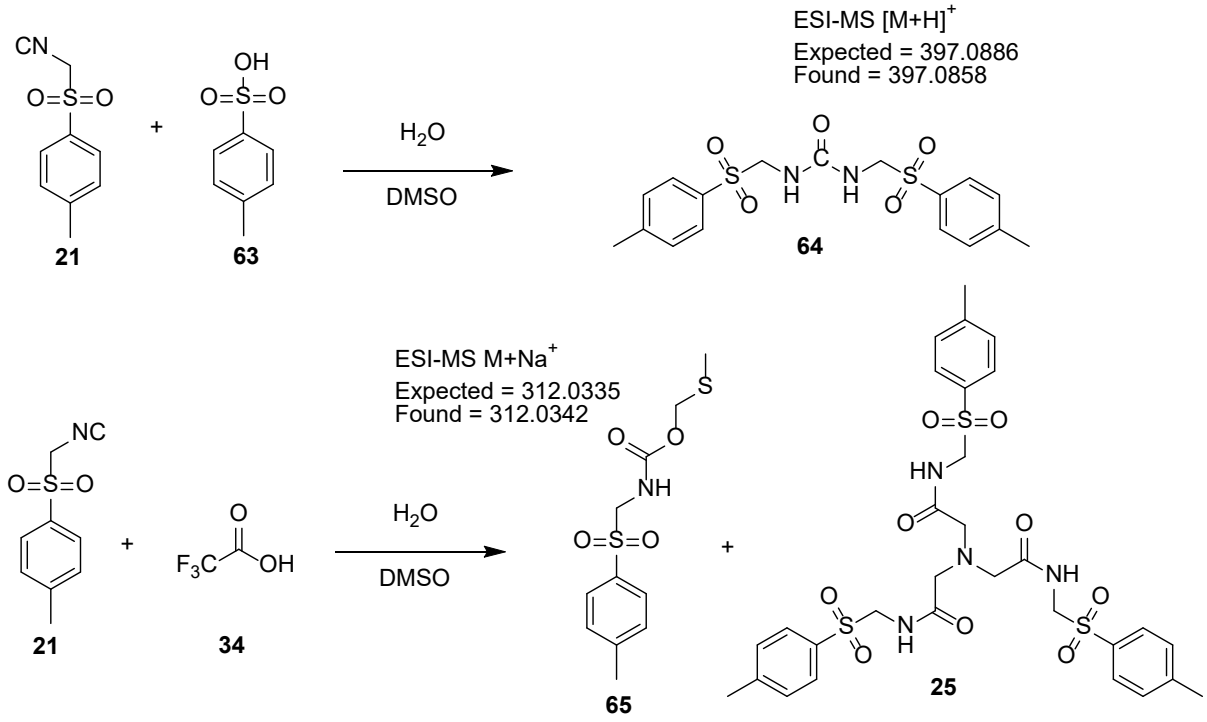


Figure S14: Side products obtained while testing PTSA (**63**) and TFA (**34**) as activators. The reactions yielding products **64** and **65** are known in literature.

Compound **64**

^{1}H NMR (600 MHz, DMSO-d_6) δ 7.66 (d, J = 7.9 Hz, 4H), 7.42 (d, J = 7.9 Hz, 4H), 7.14 (t, J = 6.7 Hz, 2H), 4.54 (d, J = 6.6 Hz, 4H), 2.40 (s, 6H).

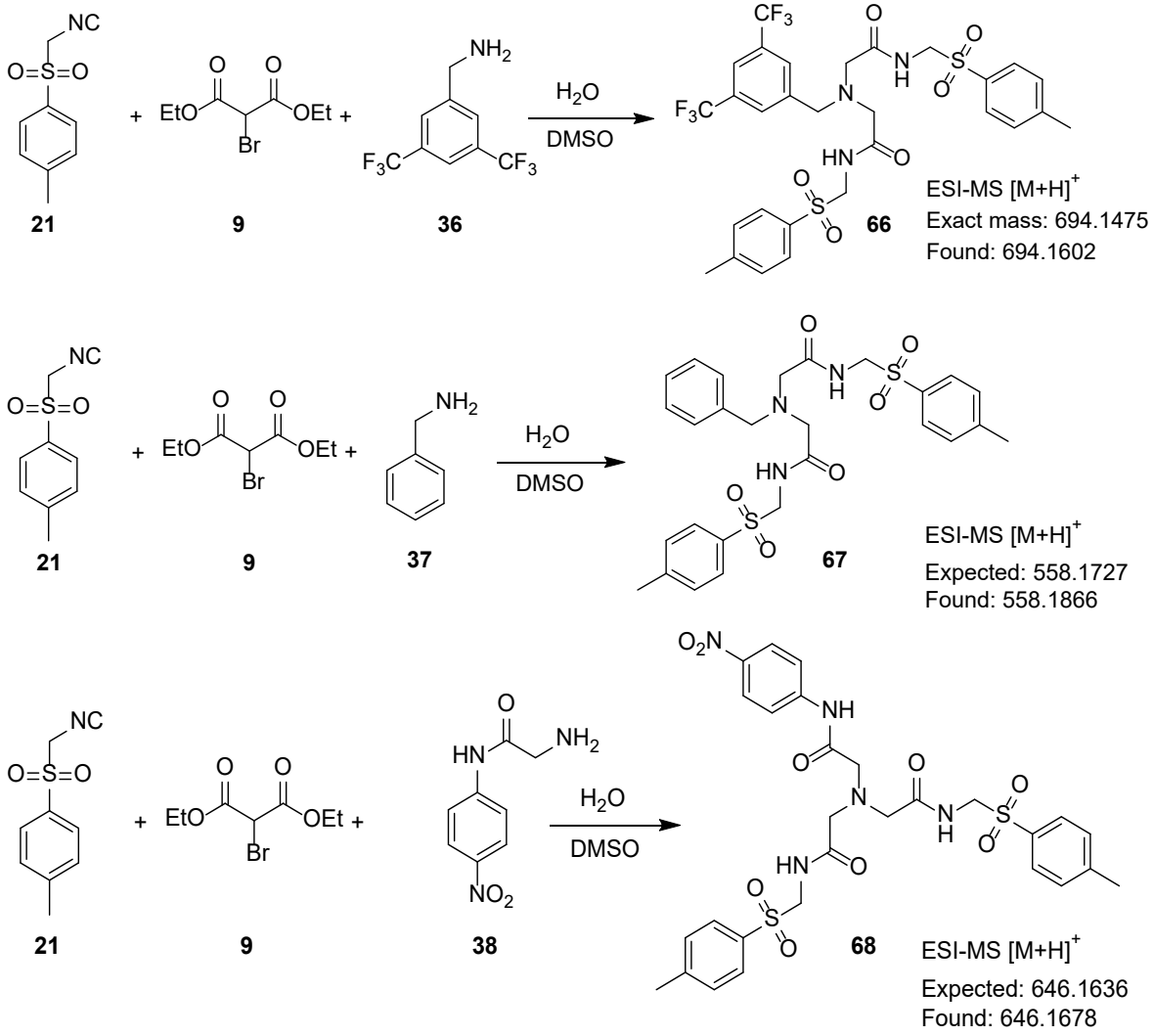
Compound **65**

^{1}H NMR (600 MHz, DMSO-d_6) δ 8.51 (t, J = 6.8 Hz, 1H), 7.71 (d, J = 7.9 Hz, 2H), 7.44 (d, J = 7.9 Hz, 2H), 5.03 (s, 2H), 4.55 (d, J = 6.7 Hz, 2H), 2.41 (s, 3H), 2.09 (s, 3H).

^{13}C NMR (151 MHz, DMSO) δ 154.92, 144.31, 134.27, 129.50, 128.16, 68.24, 62.23, 20.82, 14.17.

3.5 Asymmetric variations of product 25

The reaction has been repeated with the same conditions indicated in section 3.1 adding one equivalent of amine (**36**, **37**, **38**, **40**). The reaction mixtures were sampled after 24 h and analysed in the HPLC-MS. In all cases the expected product was observed. The masses found are reported in **Figure S15**.



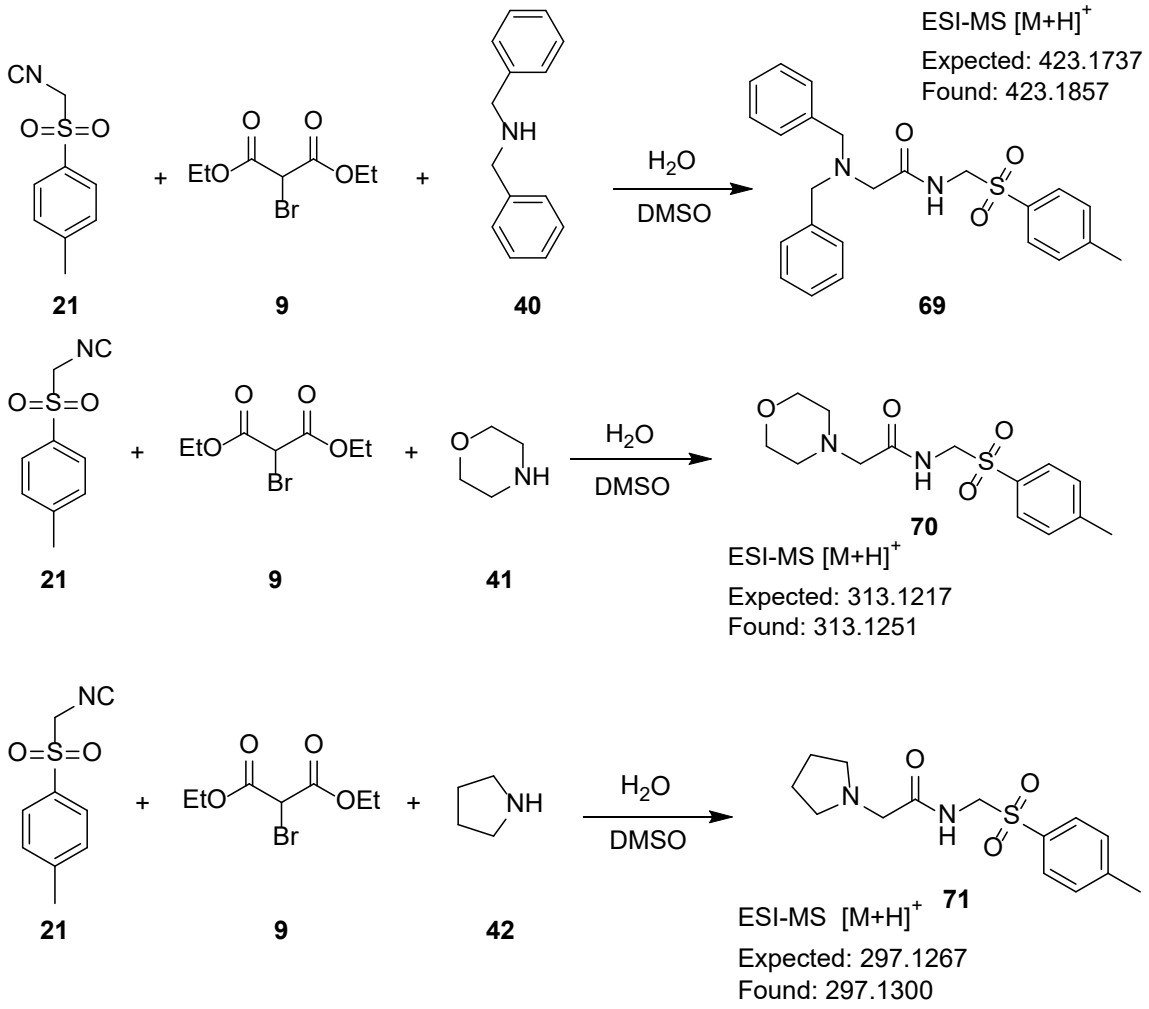


Figure S15: Scheme of the reaction variations involving the presence of a primary or secondary amine. The analogs of the mechanism intermediates have been detected by HPLC-MS.

3.6 Other Isocyanide variations

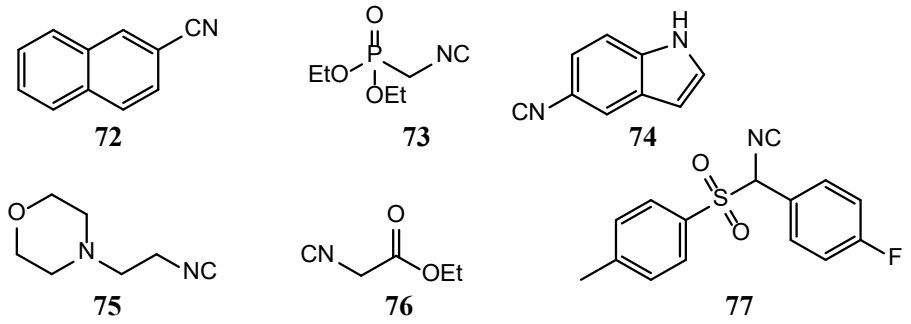


Figure S16: Other isocyanide variations tried on the reaction. None of them yielded product **25.**

The reaction has also been tried on the isocyanides reported in **Figure S16** (2-Naphthyl isocyanide **72**, Diethyl isocyanomethylphosphonate **73**, 1H-Indol-5-yl isocyanide **74**, 2-Morpholinoethyl isocyanide **75**, Ethyl isocyanoacetate **76**, α-(p-Toluenesulfonyl)-4-fluorobenzylisonitrile **77**. The reactions were prepared according to the procedure reported in section 3.1 and repeated at 50°C. HPLC-MS analysis showed no traces of the analogues of product **25**. The analysis of molecule **77** showed the presence of the one-branched (**78**) and two-branched (**79**) analogues suggesting that the intermediate is too hindered to form the third branch (**Figure S17**).

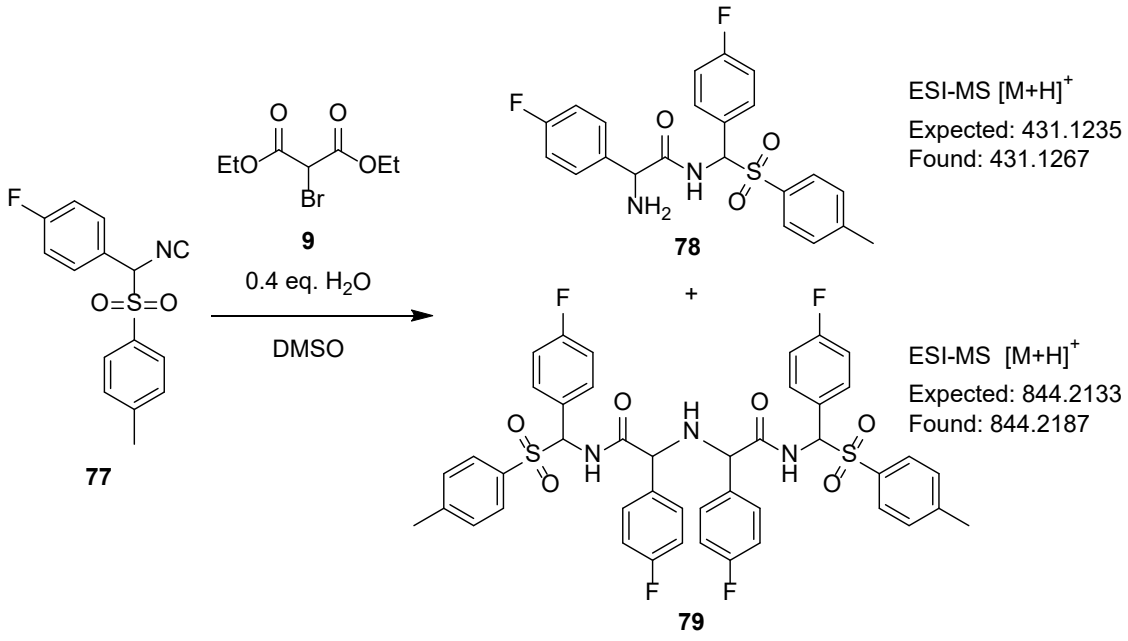


Figure S17: Isocyanide **77** did not produce the expected analogue of **25**. However, the one-branched (**78**) and two-branched (**79**) versions were detected.

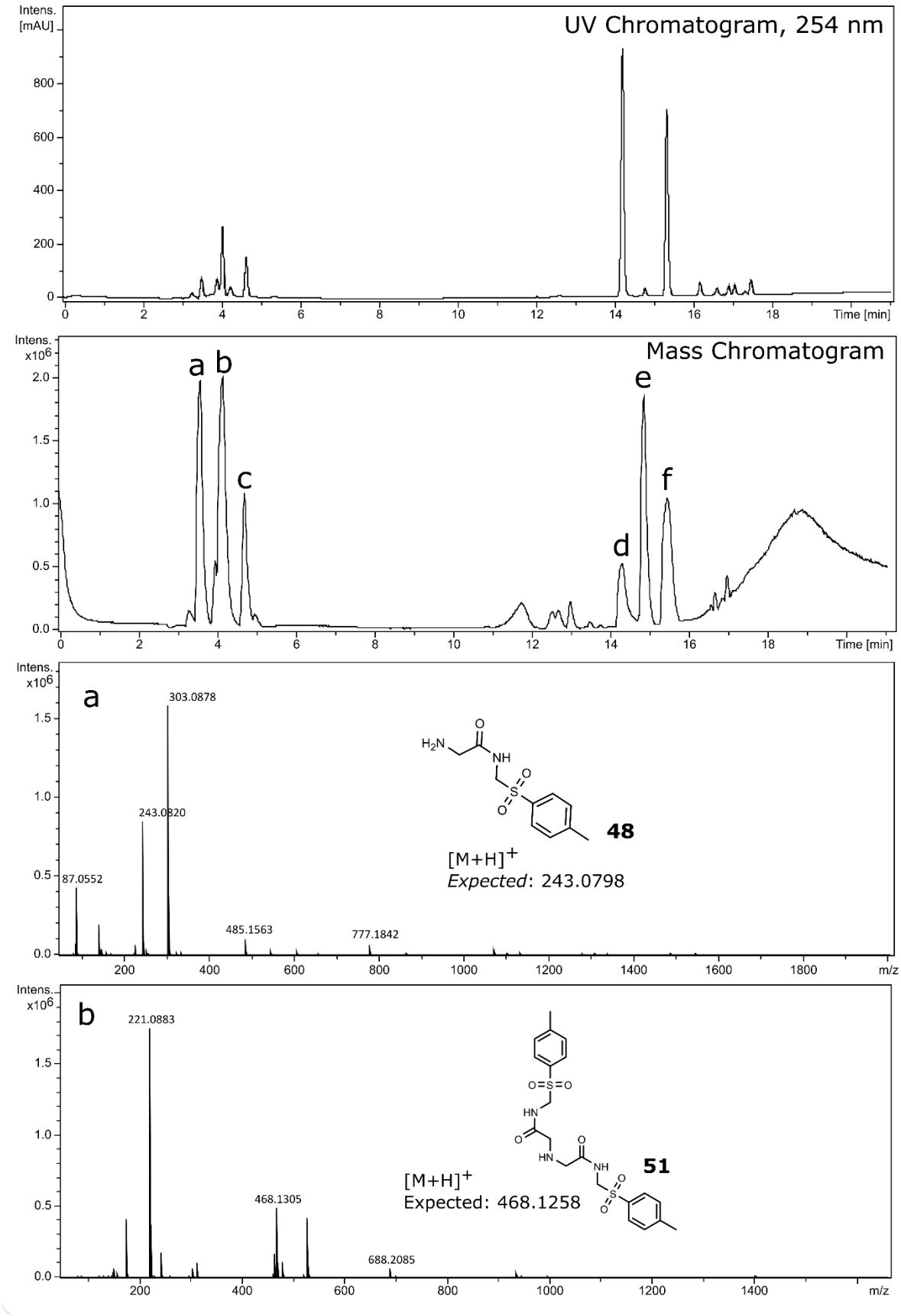
4 Reaction mechanism data

4.1 HPLC-analysis

Chromatographic separation of the reaction mixture was achieved with a reverse phase column by Agilent (Poroshell 120 HPH C18, 3.0 x 100 mm, 2.7 µm) on a Thermo Fisher UltiMate 3000 HPLC. From the reaction mixture 45 µL were sampled and diluted in 1ml of acetonitrile (0.022M final concentration). 10 µL of the sample were then injected in the instrument and eluted with a linear gradient mixture of solvents: water w/0.1% v/v formic acid and acetonitrile w/0.1% v/v formic acid at 0.5 mL per minute, over 21 minutes as indicated in **Table S1**. The column compartment was maintained at 30 °C. Results after 12h of reaction are showed in **Figure S18**. As reference the structure and the calculated mass of the compounds are reported. UV detection was performed using a diode array detector (DAD) set on 254 nm. The MS apparatus was a Bruker MaXis Impact instrument, acquisition range at 50–2000 *m/z*. Data was analysed using the Bruker DataAnalysis software suite.

Table S1: HPLC method used for the analysis of the reaction mixture.

Time [min]	Water [%]	Acetonitrile [%]
0	95	5
2	95	5
15	5	95
18	5	95
20	95	5
21	95	5


UV Chromatogram, 254 nm
Intens. [mAU]
Time [min]
Mass Chromatogram
Intens. x10^6
a
b
c
d
e
f
Time [min]
a
303.0878
243.0820
87.0552
485.1563
777.1842
m/z
48
[M+H]+
Expected: 243.0798
b
221.0883
468.1305
688.2085
m/z
51
[M+H]+
Expected: 468.1258

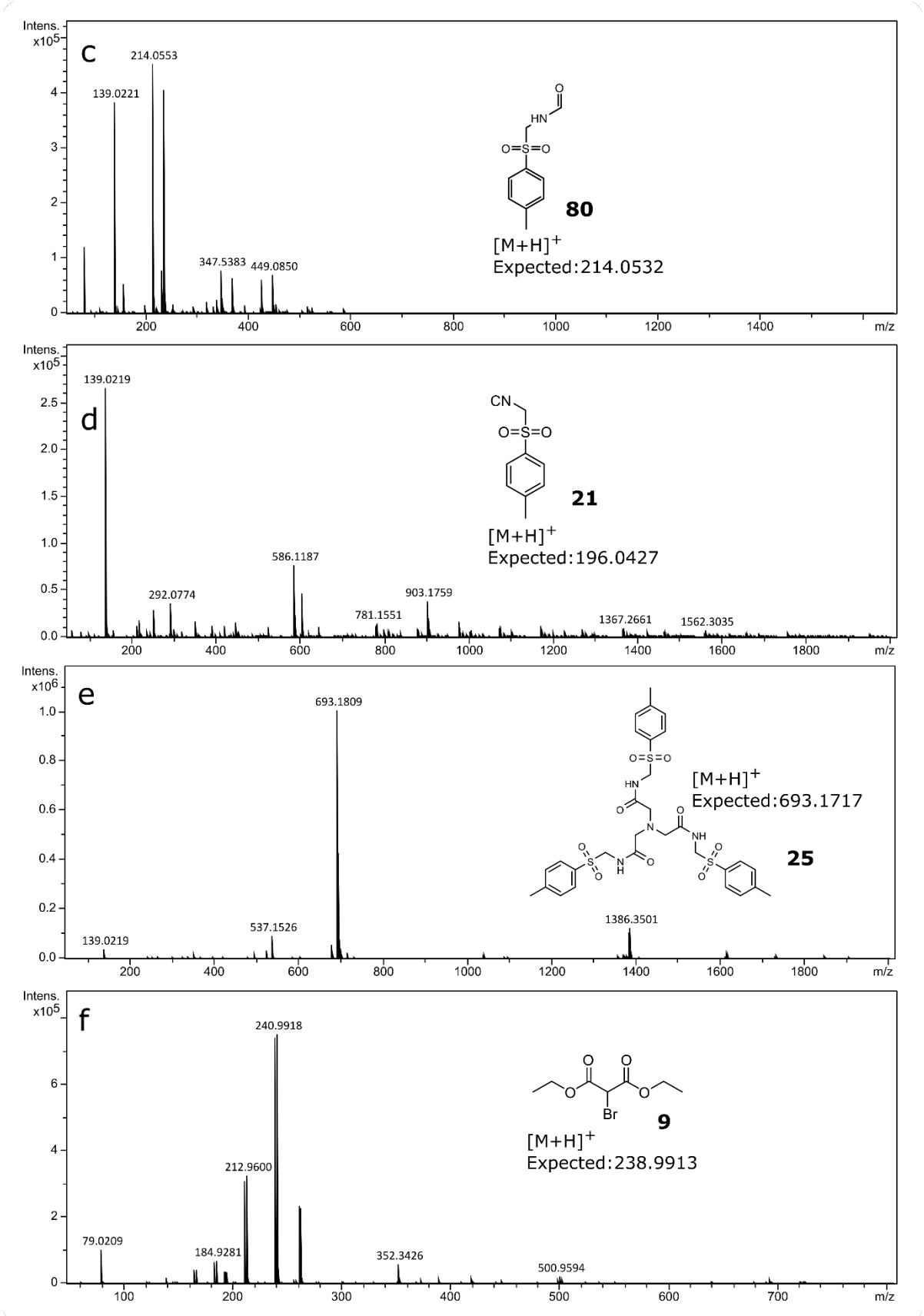
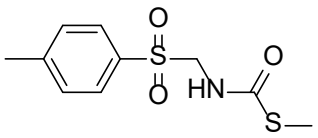


Figure S18: HPLC-MS analysis of the reaction mixture. The MS data of the major peaks are reported with the assigned molecules

4.2 Time resolved HPLC analysis

The reaction was monitored by automatically injecting a sample in the HPLC instrument (agilent 1100 equipped with DAD set on 254 nm) every two hours for 50 h. By using a simple liquid handling platform 0.3 ml were sampled from the reaction mixture and moved to a flask using a Tricontent pump. 6.5 ml of MeCN were added to the flask for dilution and the solution was pumped into a loop valve (Rheodyne, part number: MXP7920-000, equipped with a 10 µl sample loop) connected to the instrument and remotely controlled. Once the valve was switched the HPLC was triggered using a contact closure and the method was run. Results of this experiment are showed in **Figure S20**. The integrals of the main 6 peaks have been extracted and plotted vs time (**Figure 6d** in the manuscript). Results show that over 35 h TosMIC get completely consumed, the dimeric product **51** forms reaching its max at 20 h and then gets consumed while the integrals of the final product **25** and of an unknown product appear (RT: 6 minutes). The molecule corresponding to the peak at 6 minutes has been isolated with flash column chromatography and characterised with NMR and MS. It corresponds to the product **81** in **Figure S19**, an adduct of DMSO and TosMIC not involved in the proposed mechanism.



81

ESI-MS $[M+H]^+$
Expected: 260.0410
Found: 260.0409

Figure S19: Structue of the side product corresponding to the peak 3 in **Figure S20**. It is belived to not be involved in the mechanism.

^{1}H NMR (600 MHz, $CDCl_3$) δ 7.80, 7.79, 7.78, 7.78, 7.36, 7.35, 6.26, 6.25, 6.24, 4.66, 4.65, 2.45, 2.22.

^{13}C NMR (151 MHz, $CDCl_3$) δ 145.48, 133.68, 129.93, 128.85, 59.10, 24.18, 21.67, 12.31.

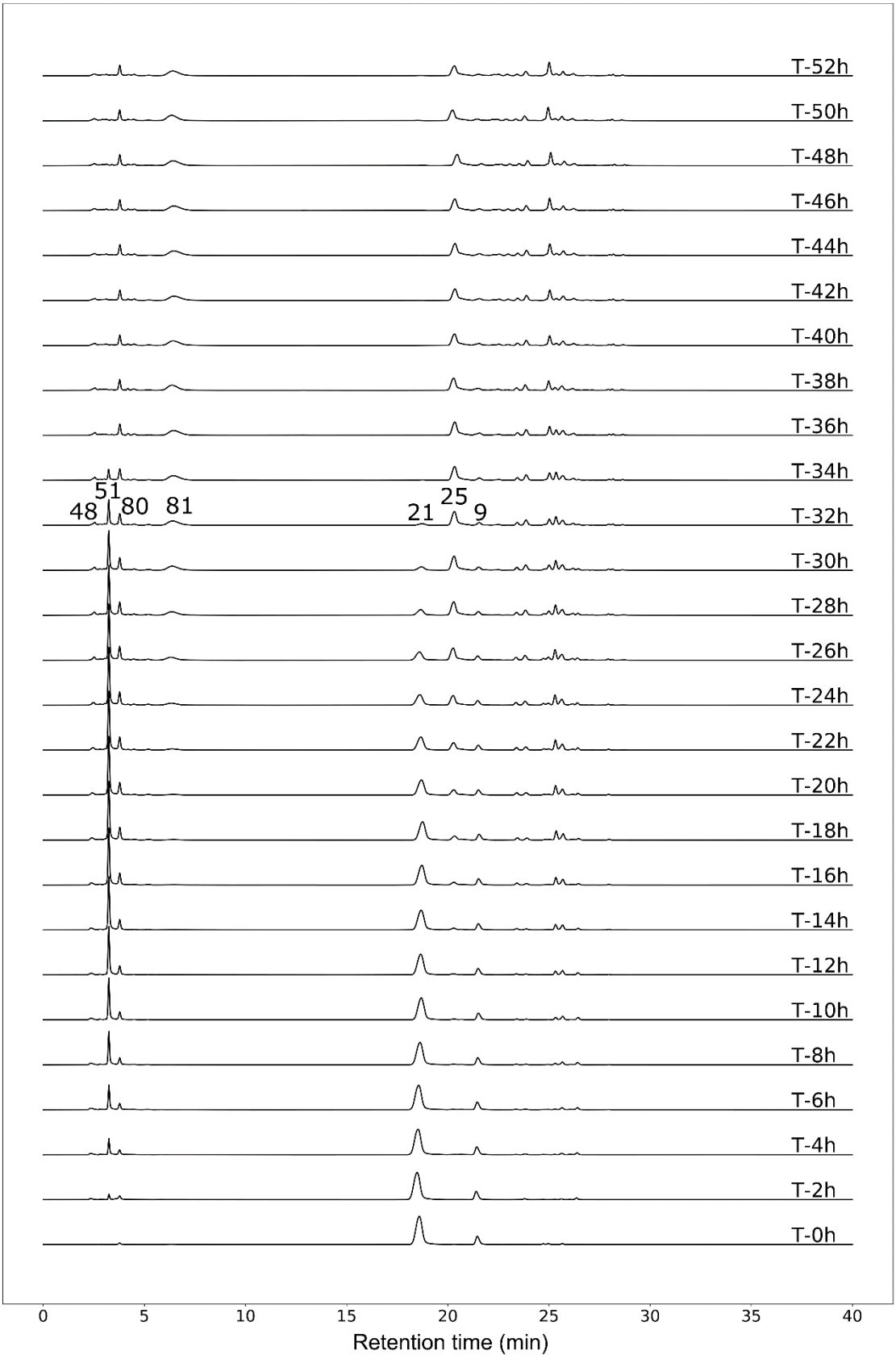


Figure S20: Online HPLC analysis of the reaction mixture. The known peaks are marked with the molecule number.

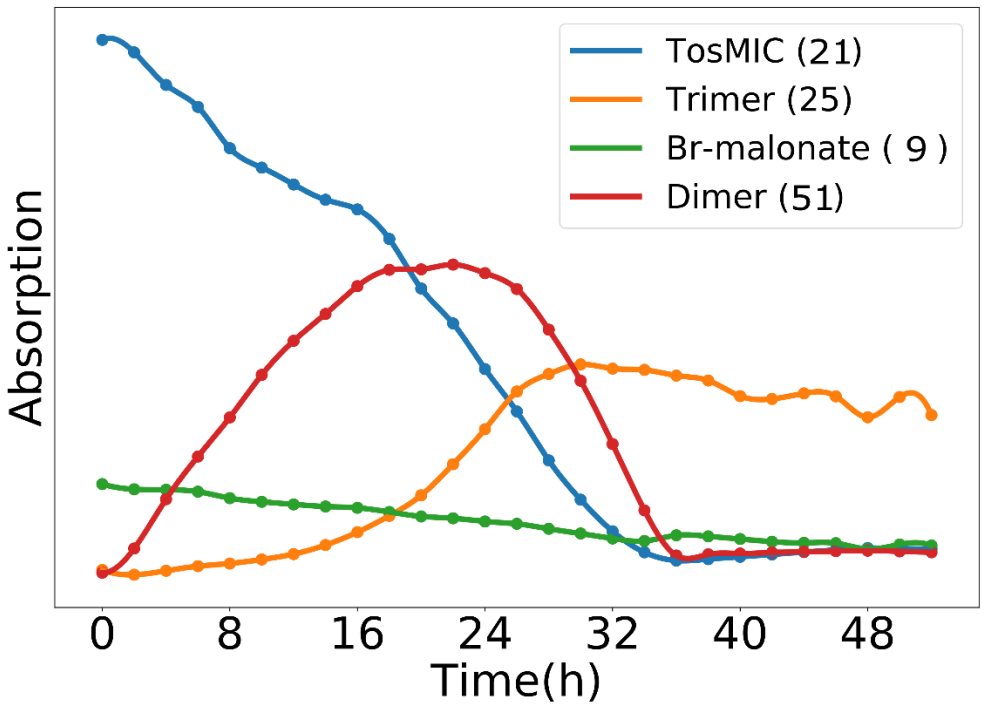


Figure S21: Visualisation of the time-resolved integrals in Figure S20 of the reagents **12** and **22**, the intermediary **50** and the product **25.**

4.3 Water influence

The calibration curve of the product **25** absorption is acquired with multiple injections of product **25** and an internal standard (hexafluorobenzene) at different ratios (**Figure S22**), integrals were integrated from the 210 nm wavelength data.

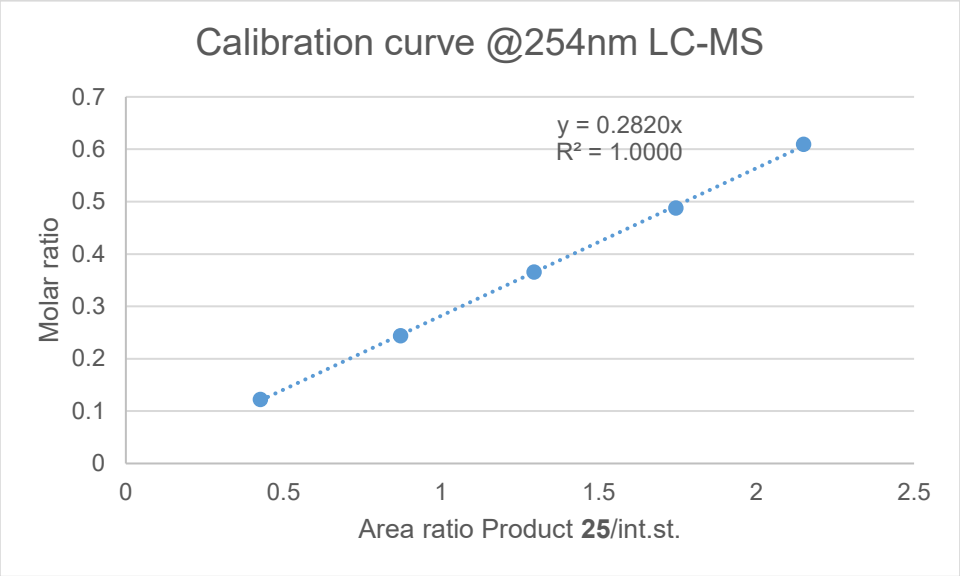


Figure S22: Calibration curve for the product **25.** Samples at different ratios with the internal standard were injected.

The slope of the calibration curve is used to calculate the yield of the product in the mixture in presence of different amount of water (see **Table S2**)

Table S2: The yield of the reaction performed in presence of different amounts of water was determined through HPLC analysis.

Water equivalents	Area product **25** 210nm	Area ISTD 254nm	Yield
0	1803	1865.826	2.2%
0.2	1946	1652.997	19.6%
0.4	4707	1603.213	48.9%
0.6	4189	1584.708	44.1%
0.8	2686	1672.036	26.8%
1	1461	1133.861	21.5%
2	1017	1754.559	9.7%
3	606	1708.202	5.9%

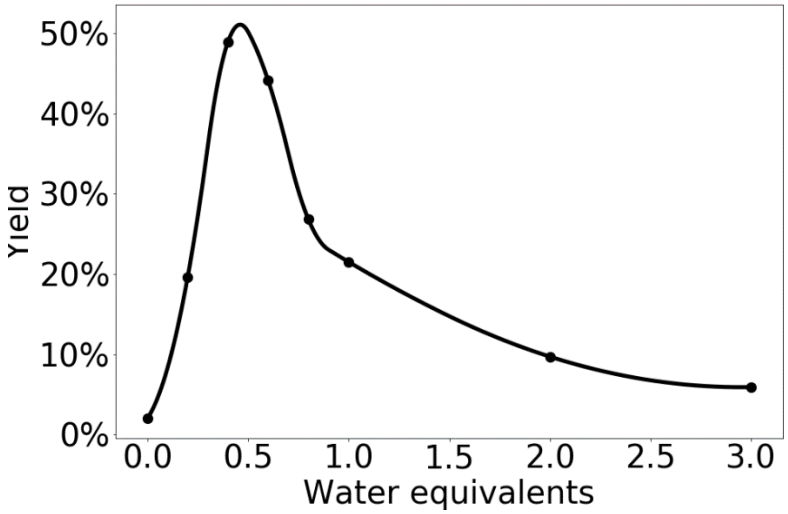


Figure S23: Visualized data points of **Table S2**.

4.4 Reaction performed in different solvents

The reaction was performed in acetonitrile and dimethylformamide. NMR analysis of the mixture showed no sign of reactivity (**Figure S24**). As further test it was also repeated in different ratio of DMSO/DMF showing the direct influence of DMSO in the yield (**Figure S25**).

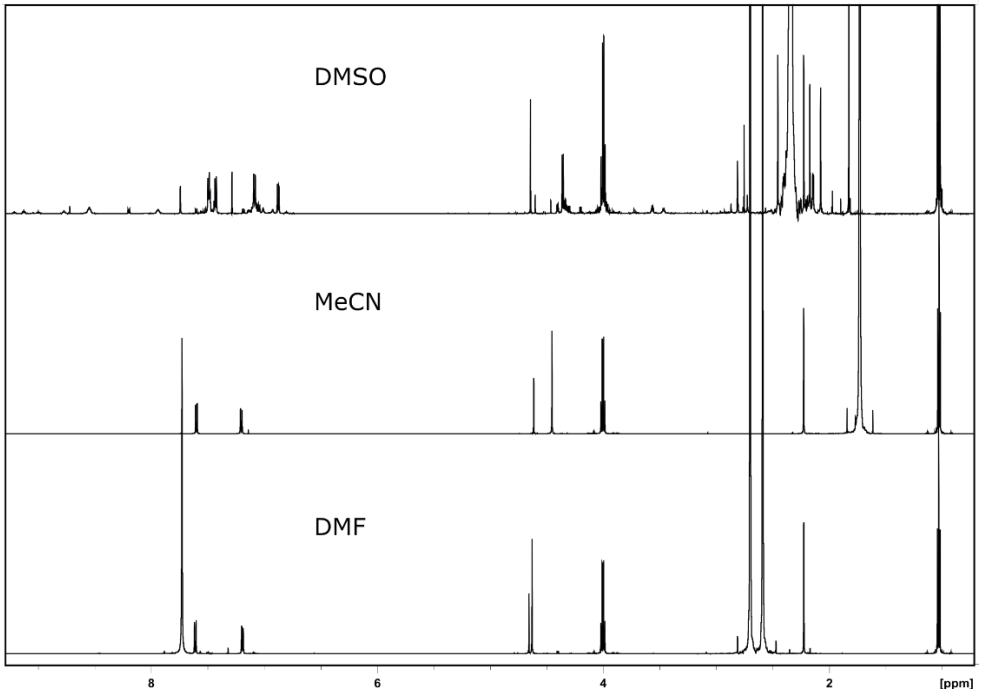


Figure S24: ^{1}H-NMR spectra of the reaction performed in different solvents. Acetonitrile and DMF show no reactivity.

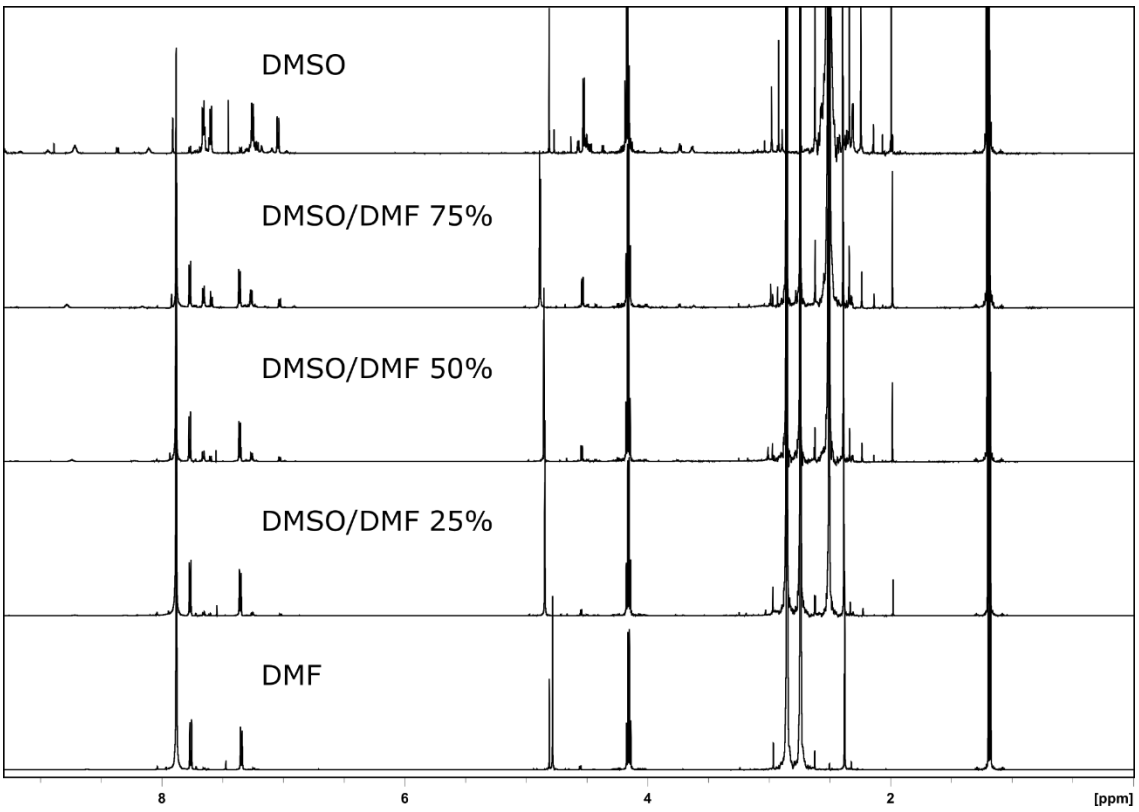


Figure S25: ^{1}H-NMR spectra of the reaction performed in a mixture of DMF and DMSO at different ratios.

4.5 Synthesis of isotopically substituted starting materials

4.5.1 Synthesis of (^{13}C)TosMIC

^{13}C-isotopically substituted TosMIC was prepared, based on literature procedures[35] [36], using either ^{13}C formamide (to substitute the isocyanide carbon) or a solution of ^{13}C-formaldehyde (to substitute the methylene carbon).

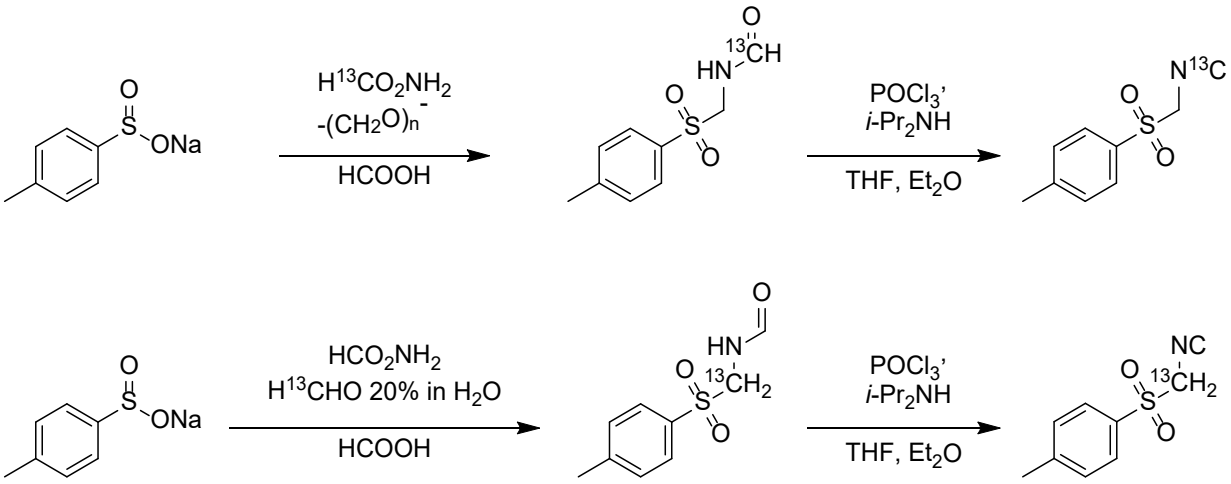


Figure S26: Synthetic approach to ^{13}C-isotopically substituted TosMIC.

4.5.2 Isocyanide Substitution

4.5.2.1 (1-13C)N-(tosylmethyl)-formamide

In a 25 ml round-bottom flask, equipped with a magnetic stir bar and a reflux condenser, were placed sodium p-toluenesulfinate (1.00 g, 5.6 mmol), paraformaldehyde (Sigma-Aldrich, 505 mg, 16.8 mmol, 3 eq.), formamide-^{13}C (Sigma-Aldrich, 99 atom% ^{13}C, 881 µl, 3.9 eq.), and formic acid (Fisher Chemicals, 1.06 ml, 5 eq.). The mixture was heated to 95 °C and stirred for 2 h at the same temperature. After the reaction was cooled to room temperature, it was diluted with cold water and transferred to a separation funnel. The mixture was extracted with ethyl acetate (3 x 40 ml) and the combined organic extracts were washed with brine, dried with $MgSO_4$, filtered and concentrated by rotary evaporation. Upon transfer to a smaller flask, the product spontaneously started to crystallize as small white crystals. The remaining solvent was then removed *in vacuo* yielding 937 mg of white crystals (85%). The pure N-(tosylmethyl)formamide was consistent with the spectral characteristic literature data for the non-isotopically substituted analogue.

^{1}H NMR (600 MHz, DMSO) δ 9.03 (q, *J* = 6.1 Hz, NH), 8.63 (dt, *J* = 11.4, 6.9 Hz, NH), 7.99 (dd, *J* = 197.9, 1.3 Hz, CHO), 7.80 (dd, *J* = 194.0, 11.1 Hz, CHO), 7.76 – 7.72 (m, 3H, Ar), 7.48 (d, *J* = 8.0 Hz, Ar), 7.44 (d, *J* = 8.0 Hz, Ar), 4.76 (t, *J* = 6.6 Hz, CH_2), 4.73 (dd, *J* = 6.8, 4.3 Hz, CH_2), 2.42 (s, CH_3), 2.40 (s, CH_3) (mixture of two rotamers).

^{13}C NMR (151 MHz, DMSO) δ 165.43, 161.00, 144.98, 144.71, 134.45, 133.65, 129.99, 129.81, 128.68, 128.51, 62.68, 58.55, 21.12, 21.11 (mixture of two rotamers).

LC-MS $[M + H]^+$ calculated for $C_8{}^{13}CH_{12}NO_3S$, 215.057; found, 215.0627.

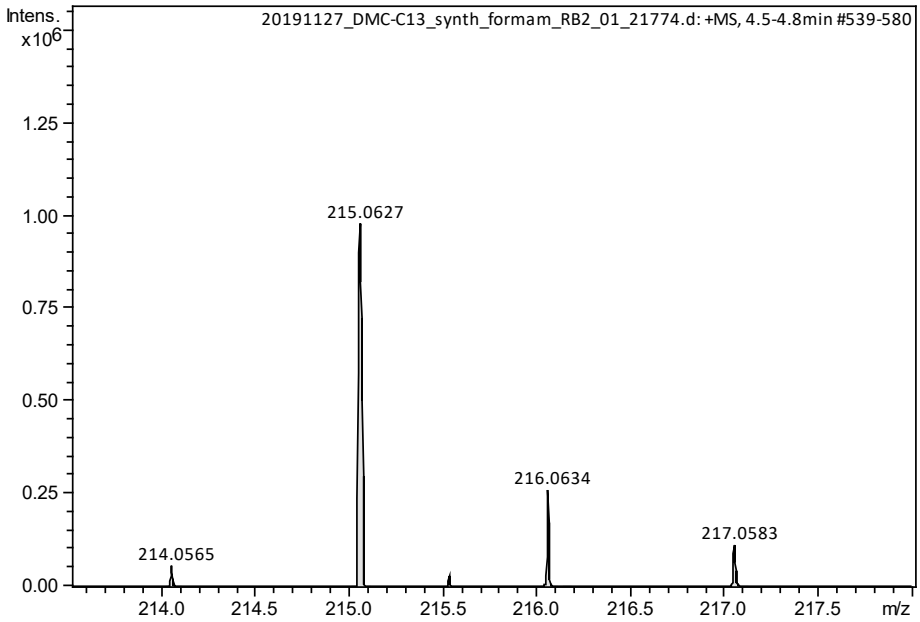


Figure S27: MS peak corresponding to (1-^{13}C)N-(tosylmethyl)-formamide.

4.5.2.2 TosCH$_2$N^{13}C

A 25 ml pear-shaped flask, equipped with magnetic stir bar was charged with ^{13}C-labelled N-(tosylmethyl)formamide (937 mg, 4.39 mmol), dry THF (3 ml), anhydrous diethyl ether (1 ml) and N,N-Diisopropylethylamine (Fluorochem, 2.29 ml, 13.2 mmol, 3 eq.). The stirred suspension was chilled to −5 °C with an ice-salt bath. Under a nitrogen atmosphere, 1.317 ml of a 4 M stock-solution of phosphorous oxychloride (1.2 eq.) in dry THF were slowly added via a syringe pump over the course of 1 hour. Towards the reaction completion the white suspension turned orange. After being stirred for another 30 min at 0 °C, the mixture was diluted with cold water (15 ml) and kept under vigorous agitation. The product then began to separate as a fine, brown crystalline solid that was collected by vacuum filtration, washed with cold water and dried in vacuum (493 mg, 57%). Spectral data in accordance with the literature for labelled-TosMIC [37] [38] [39].

^{1}H NMR (600 MHz, DMSO) δ 7.85 (d, *J* = 8.2 Hz, 2H), 7.56 (d, *J* = 8.2 Hz, 2H), 5.56 (d, *J* = 2.6 Hz, 2H), 2.45 (s, 3H).

^{13}C NMR (151 MHz, DMSO) δ 162.61, 146.14, 130.16, 128.89, 60.48, 21.18.

LC-MS $[M - H]^-$ calculated for $C_8{}^{13}CH_8NO_2S$ 195.031; found 194.9928.

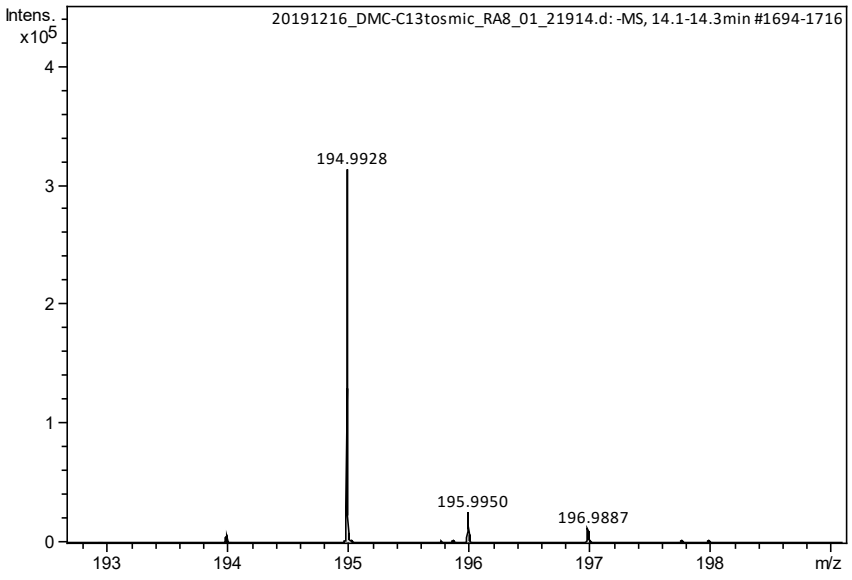


Figure S28: MS peak corresponding to TosCH$_2$N^{13}C

4.5.2.3 Product 25 - ^{13}C-1 variation

Synthesized according to standard procedure 3.1.

^{1}H NMR (600 MHz, DMSO-d_6) δ 8.96 (td, J = 6.7, 4.0 Hz, 3H), 7.68 (d, J = 8.2 Hz, 6H), 7.39 (d, J = 8.2 Hz, 6H), 4.66 (dd, J = 6.7, 3.9 Hz, 6H), 2.96 (d, J = 4.3 Hz, 6H), 2.37 (s, 9H).
^{13}C NMR (151 MHz, DMSO-d_6) δ 170.05 (^{13}C), 144.71, 134.39, 129.74, 128.49, 60.00, 56.47 (d, J = 52.6 Hz), 21.08.

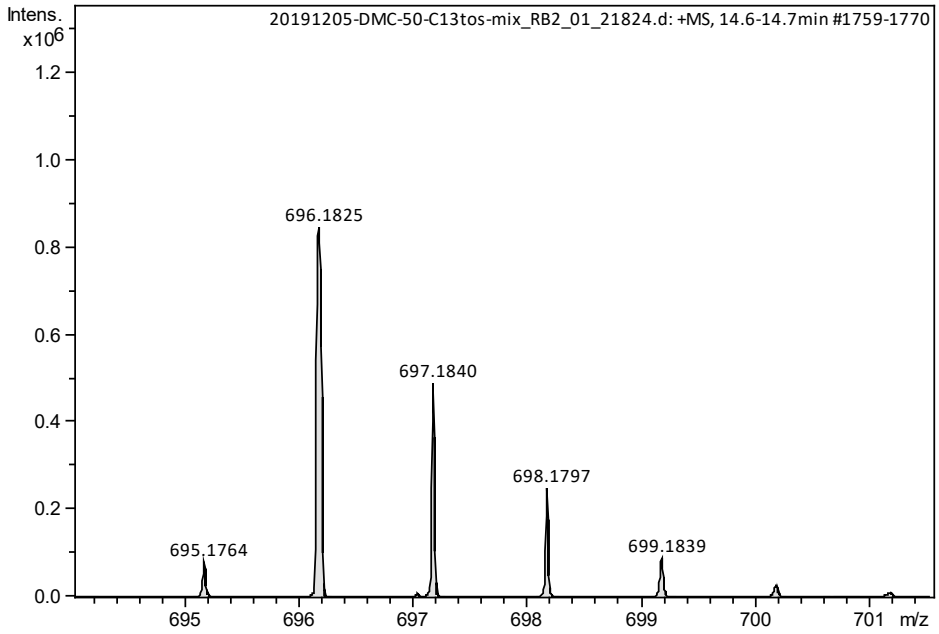


Figure S29: MS peak corresponding to Product 25 - ^{13}C-1 variation

4.5.3 Methylene Labelling

4.5.3.1 N-(tosyl(^{13}C)methyl)-formamide

Synthesized according to a modified literature procedure[35,36]. A three-necked, round-bottomed flask equipped with a magnetic stirrer is charged with 766 mg (4.3 mmol, 1 eq.) of sodium p-toluenesulfinate, 2 ml of water and 1 g (6.45 mmol, 1.5 eq.) of formaldehyde-^{13}C solution (Sigma 20 wt. % in H_2O, 99 atom % ^{13}C, 1.5eq.), 2.322 ml (2.6 g, 43 mmol, 10 eq.) of formamide and 567 ul (692 mg, 15 mmol, 3.5 eq.) of formic acid. The stirred reaction mixture is heated at 90 °C. The sodium p-toluenesulfinate dissolves during heating, and the solution is kept at 90–95 °C for 2 hours. After the reaction was cooled to room temperature, it was diluted with cold water and transferred to a separation funnel. The mixture was extracted with ethyl acetate (3 x 40 ml) and the combined organic extracts were washed with brine, dried with

$MgSO_4$, filtered and concentrated by rotary evaporation, yielding the crude N-(tosylmethyl)formamide sufficiently pure for use in the next step.

^{1}H NMR (400 MHz, DMSO-d6) δ 9.10 – 8.99 (m, NH maj), 8.68 – 8.58 (m, NH min), 7.98 (dd, J = 5.8, 1.4 Hz, CHO maj), 7.78 (d, J = 11.1 Hz, CHO min), 7.73 (d, J = 8.2 Hz, Ar, maj + min), 7.49 (d, J = 8.2 Hz, Ar, min), 7.45 (d, J = 8.2 Hz, Ar, maj), 4.76 (dd, J = 151.1, 6.9 Hz, CH_2, min), 4.71 (dd, J = 150.8, 6.8 Hz, CH_2, maj), 2.43 (s, CH_3, min), 2.41 (s, CH_3, maj) (at room temperature, mixture of two rotamers ratio 80:20).

^{13}C NMR (101 MHz, DMSO-d6) δ 160.94, 144.67, 134.47, 129.96, 129.78, 128.64, 128.47, 62.64, 58.52, 21.08 (at room temperature, mixture of two rotamers. Some peaks of the minor component were too weak to be detected).

LC-MS $[M + H]^+$ calculated for $C_8{}^{13}CH_{12}NO_3S$, 215.057; found, 215.0576.

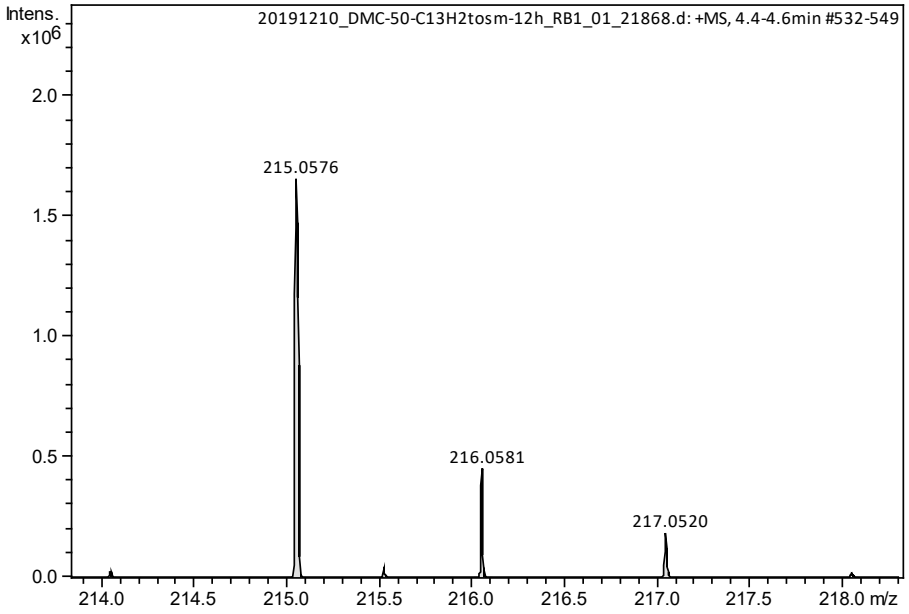


Figure S30: MS peak corresponding to N-(tosyl(^{13}C)methyl)-formamide

4.5.3.2 Tos$^{13}CH_2NC$

Same procedure as 4.5.2.2.

^{1}H NMR (600 MHz, DMSO) δ 7.85 (d, *J* = 8.1 Hz, 1H), 7.56 (d, *J* = 8.0 Hz, 1H), 5.56 (d, *J* = 157.9 Hz, 1H), 2.45 (s, 2H).
^{13}C NMR (151 MHz, DMSO-d_6) δ 162.64, 146.15, 130.16, 128.89, 60.49, 21.18.
LC-MS $[M - H]^-$ calculated for $C_8{}^{13}CH_8NO_2S$ 195.031; found 194.9984.

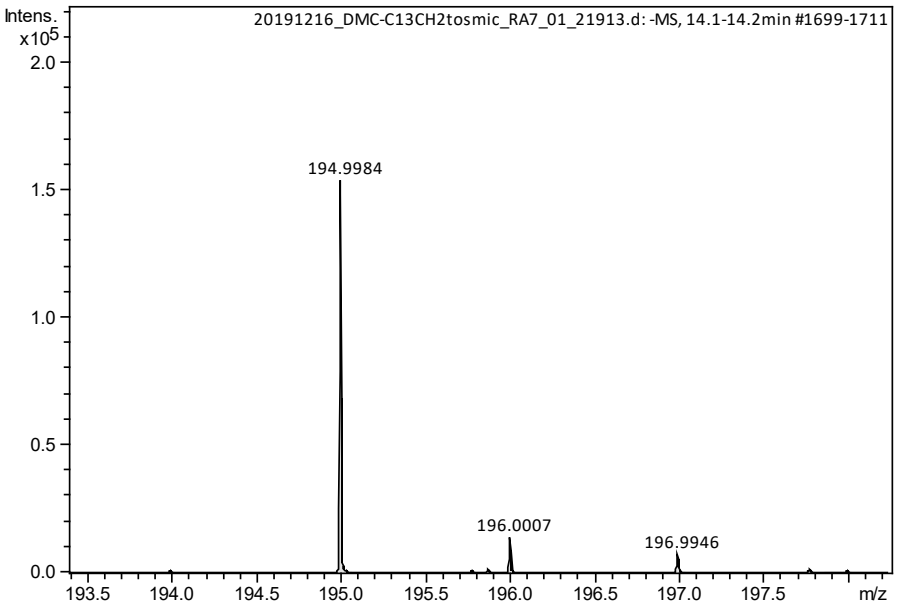


Figure S31: MS peak corresponding to Tos$^{13}CH_2NC$

4.5.3.3 Product 25 - ^{13}C-2 variation

Synthesized according to standard procedure 3.1.

^{1}H NMR (600 MHz, DMSO-d_6) δ 8.95 (t, *J* = 6.1 Hz, 3H), 7.68 (d, *J* = 7.9 Hz, 6H), 7.39 (d, *J* = 7.9 Hz, 6H), 4.65 (dd, *J* = 150.9, 6.7 Hz, 6H), 2.96 (dt, *J* = 136.3, 4.0 Hz, 6H), 2.37 (s, 9H).
^{13}C NMR (151 MHz, DMSO-d_6) δ 170.06 (d, *J* = 52.6 Hz), 144.71, 134.36, 129.74, 128.49, 60.00, 56.48, 21.08.

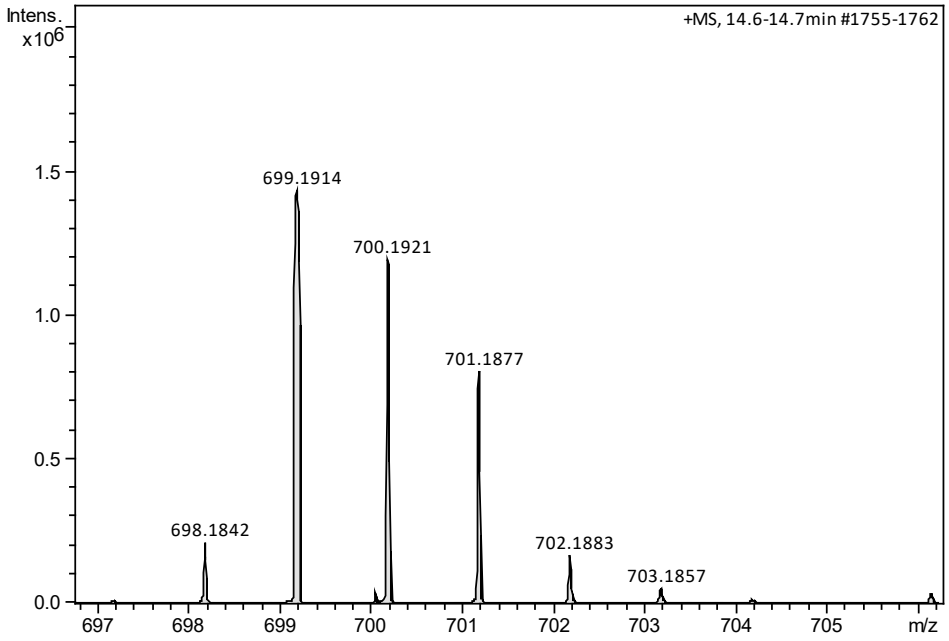


Figure S32: MS peak corresponding to Product 25 – ^{13}C-2 variation

4.5.4 DMSO Labelling

4.5.4.1 Synthesis of ^{18}O-DMSO

Method for the preparation of ^{18}O isotopically substituted DMSO:[40] solid dimethylsulfur dibromide (20 g, 90 mmol, TCI Chemicals) was added portion wise over 30 min to a vigorously stirred solution of triethylamine (25.2 ml, 180mmol) and ^{18}O-labeled water (97 atom %, Sigma-Aldrich) (1.0 ml, 50 mmol) in 60 ml of THF (fresh from solvent purification system). An ice bath was used to occasionally cool the reaction. During the addition, the orange reactant slowly dissolved giving a white precipitate of triethylamine hydrobromide. The precipitate was removed by centrifugation (5 min 4000 rpm) and washed twice with ether (same time and speed as before). The combined pale-yellow supernatant and washings were dried by rotary evaporation up to 8 mbar giving 4.6 g of orange oil, composed mostly of ^{18}O-DMSO, pure enough for use as solvent in the subsequent reaction.

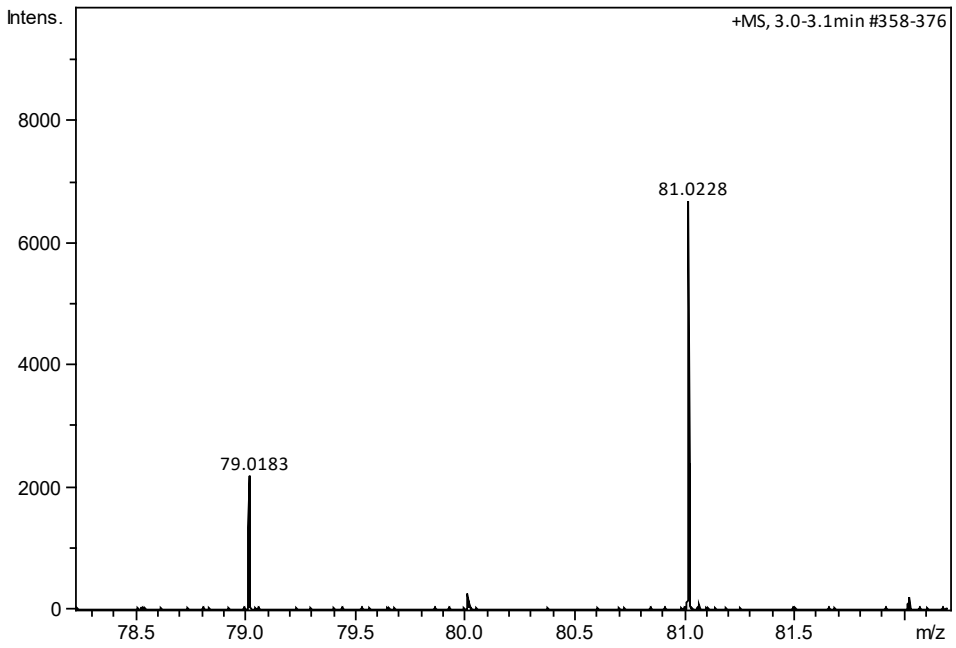


Figure S33: MS spectrum of the isotopically labelled DMSO, from the peaks intensity is it visible a 3:1 ratio between ^{18}O-DMSO and ^{16}O-DMSO.

4.5.4.2 Product 25 – 18O-DMSO Labelled

The reaction procedure followed the standard conditions presented in 3.1. The reaction mixture was analysed with HPLC-MS. The incorporation of oxygen from ^{18}O-DMSO was observed in the reaction, giving a ratio of 0.25:0.89:1:0.65 for non:mono:di:tri-labelled compounds. This is in accordance with the 3:1 ratio of the DMSO labelling and suggests that all the oxygens of the product come from DMSO.

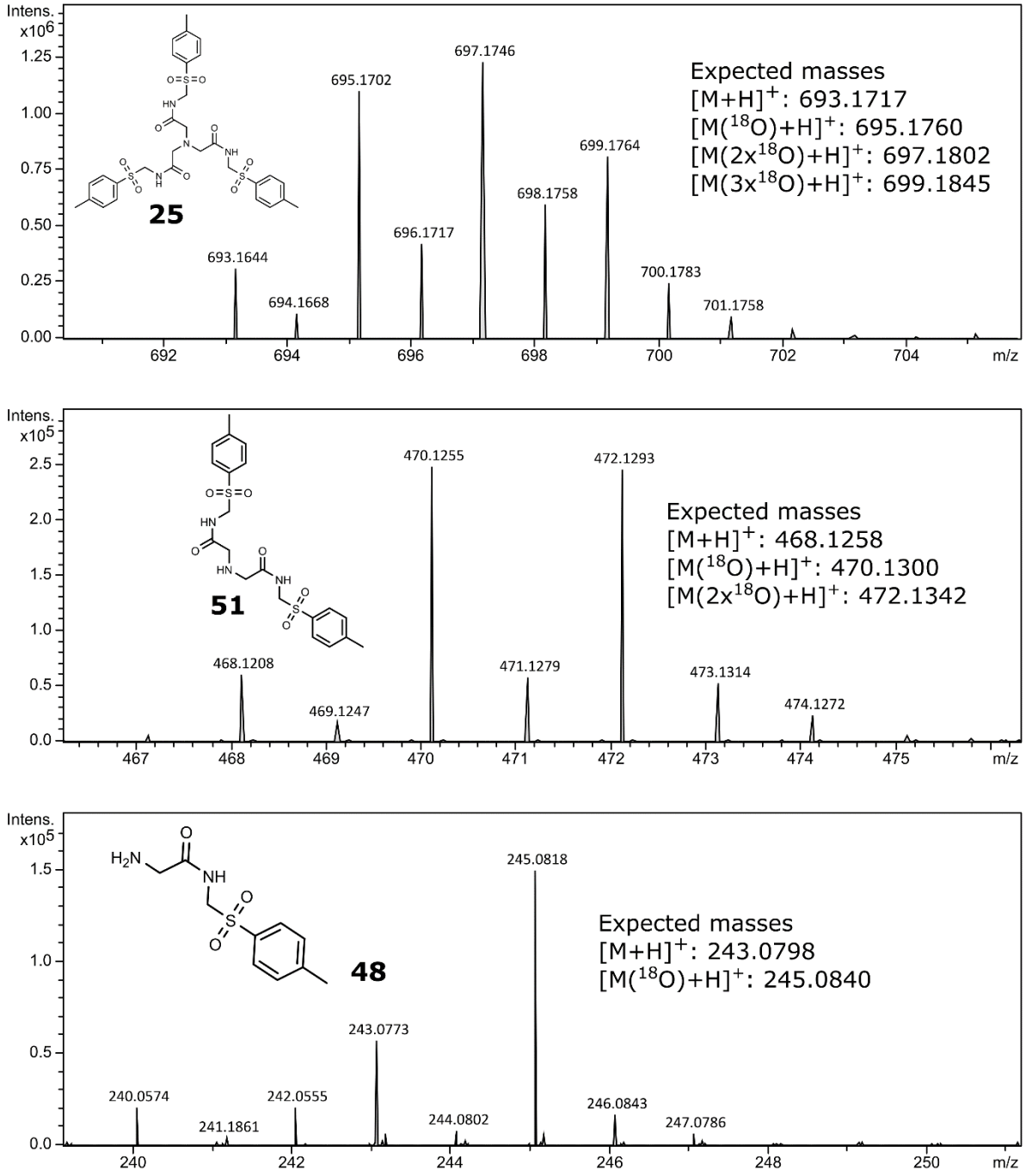


Figure S34: MS spectrum of products **25, 51** and **48** prepared in ^{18}O-DMSO. The isotopic pattern of all molecules shows incorporation in accordance with the 3:1 labelling of the solvent, suggesting that all the oxygens in the product come from DMSO.

4.5.5 Reaction in $^{18}OH_2$

The synthesis of the trimer (**25**) was undertaken according to the procedure reported in SI-3.1, using anhydrous DMSO (dried over activated molecular sieves, 4Å), under an inert atmosphere and with addition of 3 equivalents of $H_2{}^{18}O$. The mixture was analysed by HPLC-MS after 5 h

observing the formation of the product **25** and the two intermediates **48** and **51** (**Figure S35**). The most intense peaks of all compounds correspond to the non-labelled versions, indicating that the oxygens do not come from the water. However, **25** and **51** show an increase in the intensity of the $[M+H+2]^+$ isotope, suggesting the partial incorporation of a water molecule can occur. This can be explained with the presence of amines in solution reversibly forming imine groups with the carbonyls**,** essentially exchanging the oxygens with water.

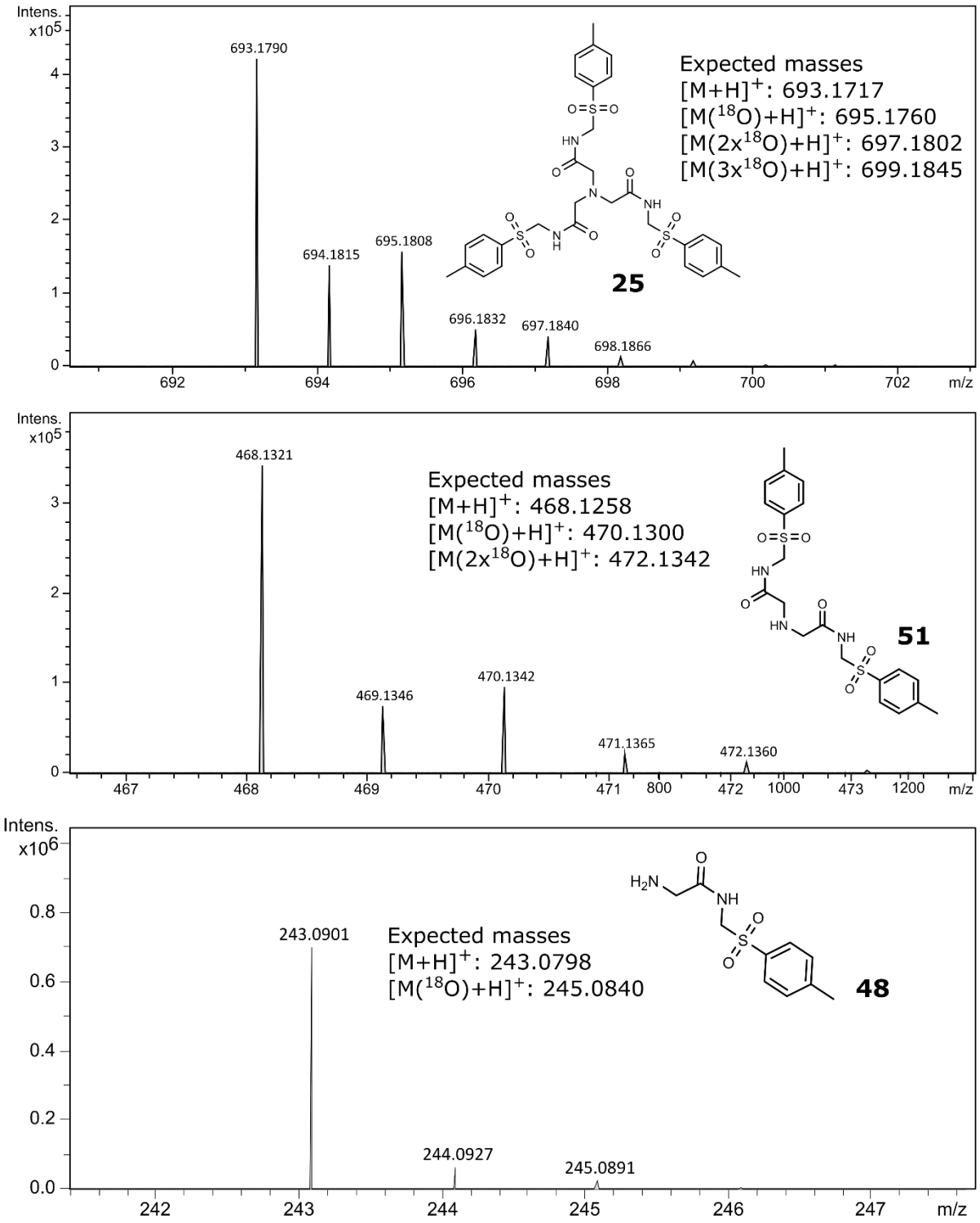


Figure S35: In order: mass isotope patterns of the product **25** and the intermediates **51** and **48**. The reaction was performed in presence of labelled water and shows no incorporation. The small increase in the +2 isotopes of **25** and **51** is explained with the imine equilibrium of the carbonyls.

4.6 IR reaction monitoring.

The reaction prepared according to the standard recipe 3.1 was monitored for 18 hours with an IR instrument (Thermo Scientific™ Nicolet™ iS™5 FTIR) equipped with a flow cell. The reaction was continuously pumped into the flow cell and back into the reactor vessel using a peristaltic pump. The instrument was remotely controlled with python and acquired a new spectrum of the mixture every 5 minutes. Results are showed in pictures **Figure S36** and **Figure S37**. The signal of the CN bond from TosMIC at 2150 cm^{-1} disappears after 10 h while the formation of isocyanate[41] (2257 cm^{-1}) and CO_2 (2349 cm^{-1}) are observed.

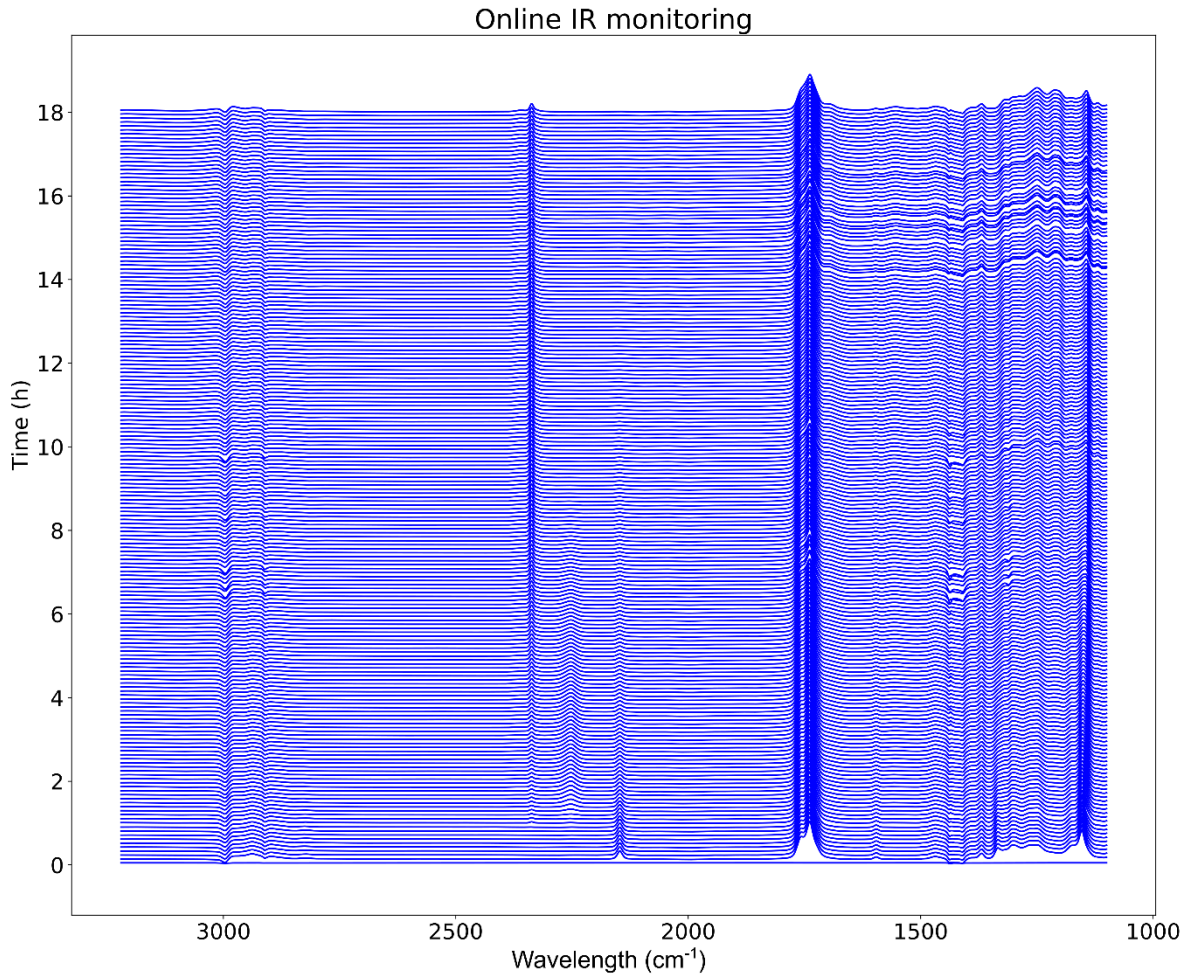


Figure S36: Online FT-IR analysis of the reaction mixture. A new spectrum is acquired every 5 minutes. For 18 hours, the wavelength are plotted vs the time of acquisition. The desappearing of TosMIC and appearing of an isocyanate bond and CO_2 are visible around 2250 cm^{-1}. Espansion of this area in **Figure S37**.

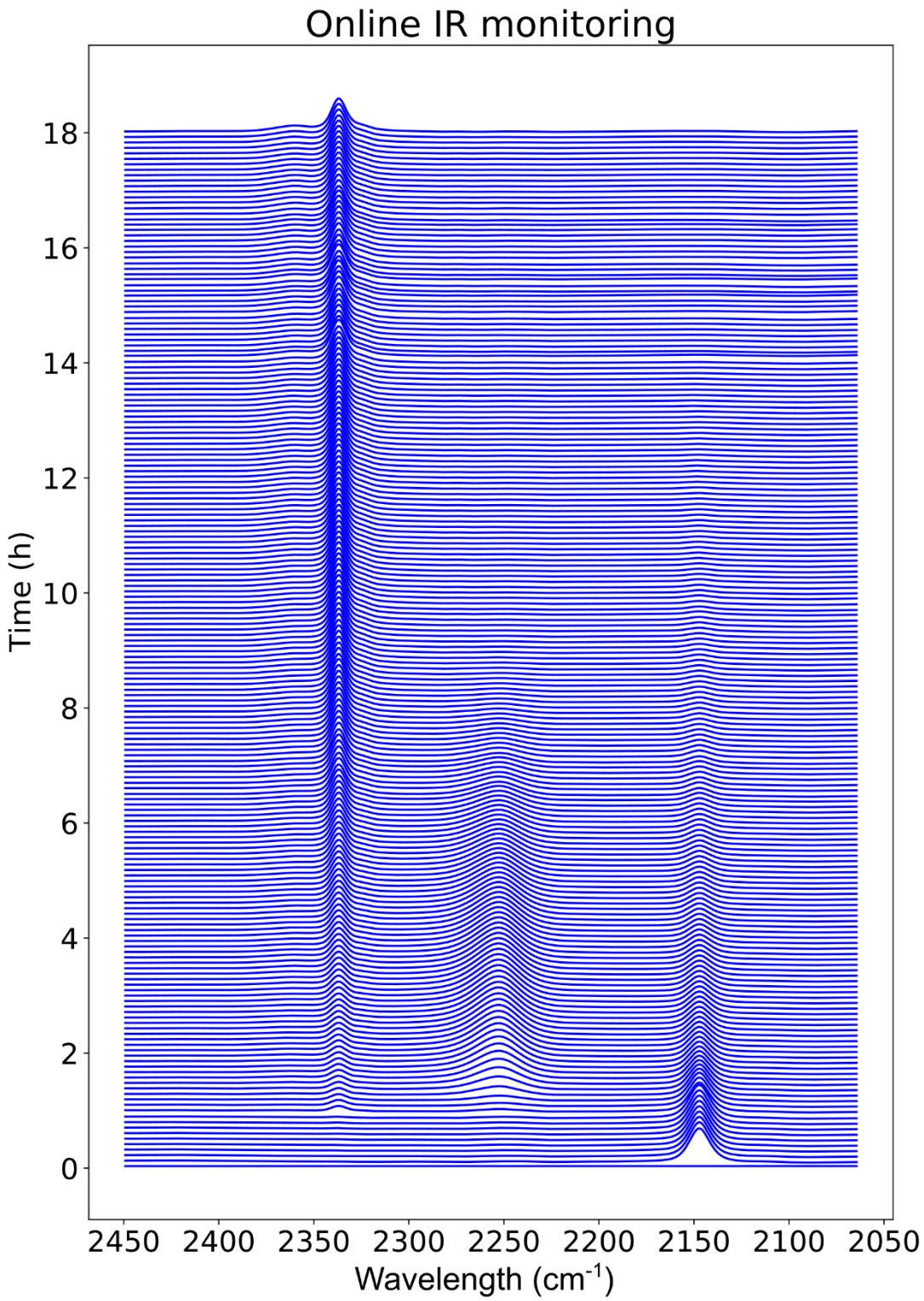


Figure S37: Detail of the online IR monitoring (**Figure S36**) showing the disappearing of the CN bond from TosMIC at 2150 cm^{-1} and the formation of isocyanate (2257 cm^{-1}) and CO_2 (2349 cm^{-1}).

4.7 Online NMR monitoring

Diethyl 2-bromomalonate (0.5 mmol, 0.1 ml) and p-toluenesulfonylmethyl isocyanide (2 mmol, 0.975 g) were mixed in 1 ml of deuterated DMSO inside an NMR tube. The reaction was left inside the NMR spectrometer for 24 h at 298K and the instrument was programmed to acquire a proton and carbon spectra every two hours. Analysis of this data showed the formation of ammonium as three peaks with 1:1:1 intensity ratio between 7.0 and 7.3 ppm[42] (**Figure S38**). The formation of dimethyl sulphide is also observed as a quintuplet at 2 ppm[43] (**Figure S39**). Data was processed using MestReNova software suite.

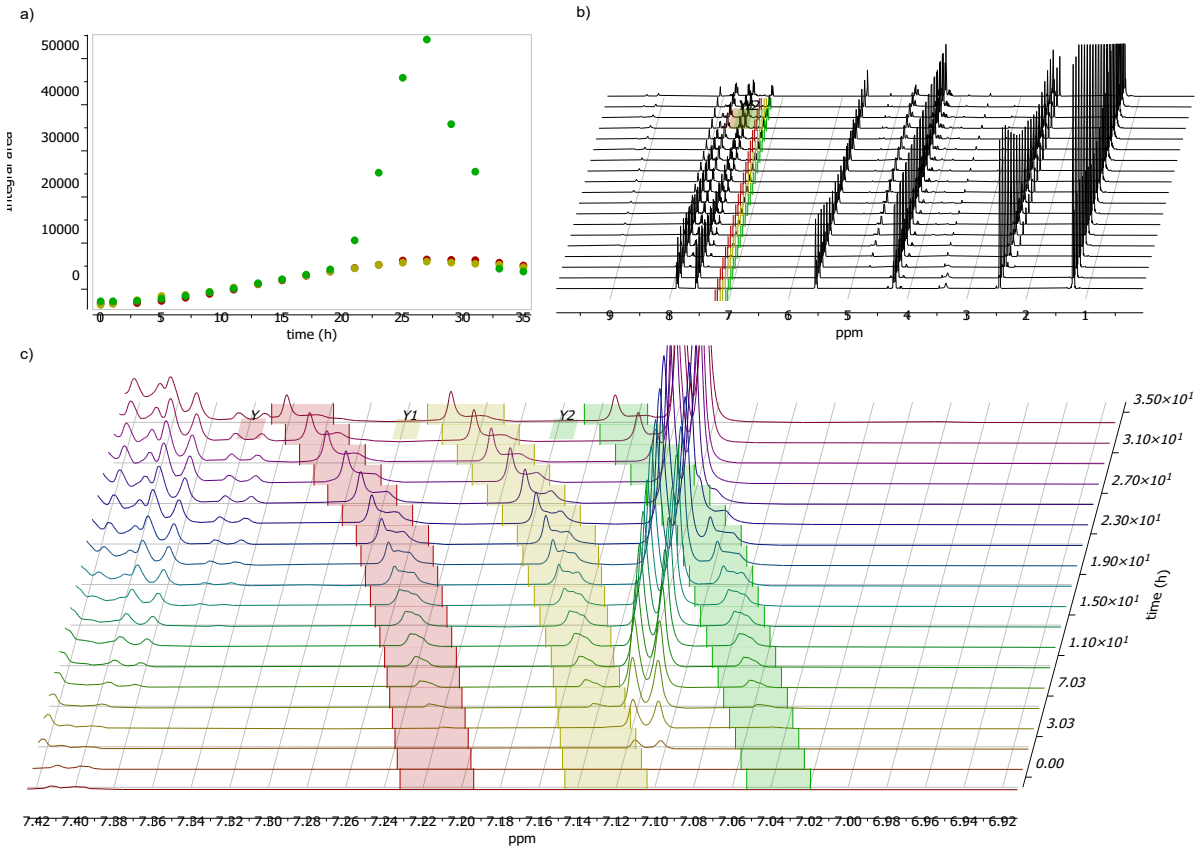


Figure S38: Online NMR monitoring of the reaction. The formation of the ammonium peaks is observed as three peaks (red, yellow and green) with 1:1:1 ratio around 7 ppm. a) Area of the three peaks over time. During the shifting peak Y2 partially overlaps with another product. b) full spectra. c) zoomed spectra showing the presence of ammonium peaks.

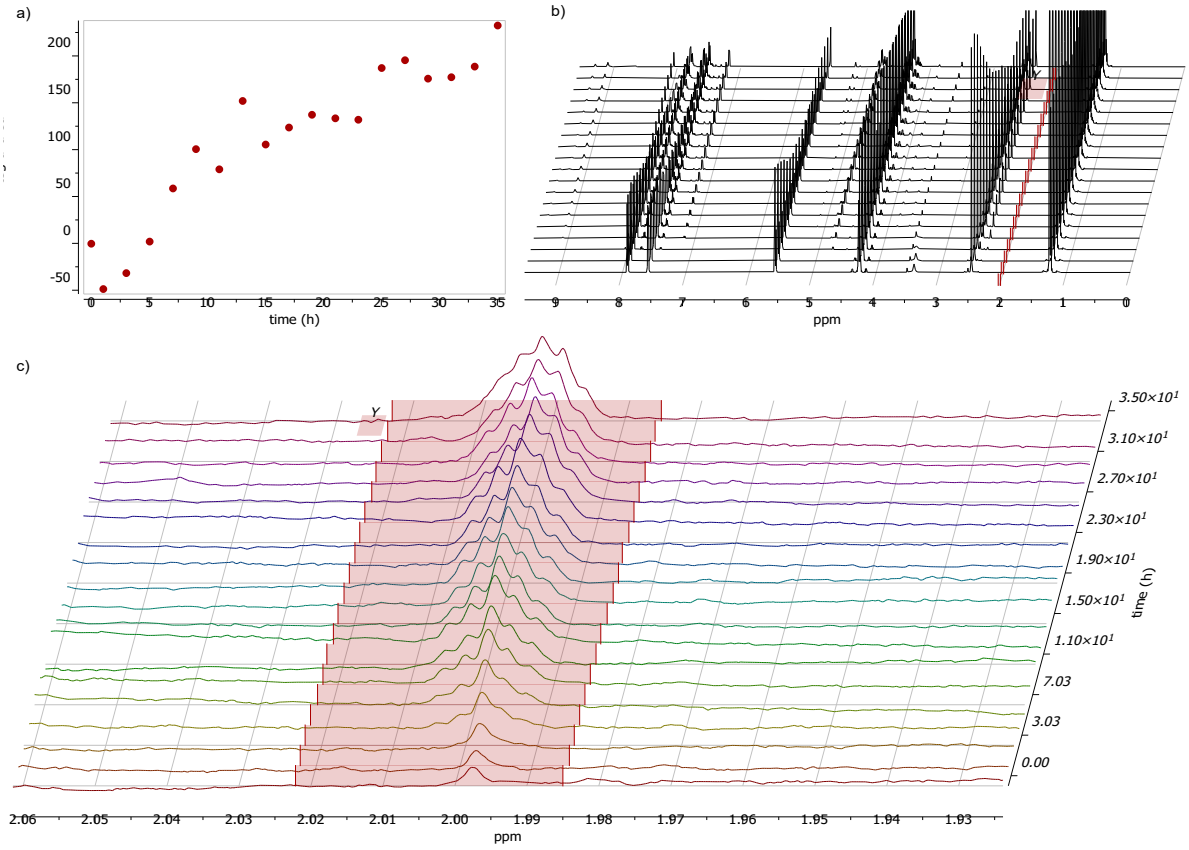


Figure S39: Online NMR monitoring of the reaction. The formation of dimethyl sulfide is observed at 2 ppm. a) Area of the peak over time. b) full spectra. c) zoomed spectra showing the presence of dimethyl sulfide peaks.

5 Cheminformatics simulation

5.1 Reaction network

A list of 21 reaction templates was encoded as SMARTS (**Figure S40**). The simulation started with a pool of three starting materials (TosMIC, water and DMSO) encoded as SMILES. The software applied the reaction templates to the starting materials and gathered the products. These were added to the pool and the process was repeated for the next step. Since the number of molecules progressively increases as result of the combinatorial explosion it was possible to calculate as further as 8 reaction steps. In order to connect the reactions graph to the product the simulation was repeated reversing the reaction SMARTS and starting from product **25**. The two resulting networks were then merged by keeping only the molecules simultaneously connected to TosMIC and **25**. The script used to perform these calculations was written in Python using the RDKit library. An estimate of the number of molecules in the reaction steps between 8 and 12 was made by fitting the natural logarithm of the number of the molecules with a second degree polynomial curve.

```
Reaction_SMARTS = {
  "imine formation 1": [C,#1:4][C:1](=O)[C,#1:5].[CX4:3][Nh2+0:2]>>[*:4][C:1](=[*+:2][*:3])[*:5].O",
  “imine formation 2":
"[C,#1:5][C:1](=O)[C,#1:6].[CX4:4][Nh+0:2][CX4:3]>>[*:5][C:1](=[*+:2]([*:4])[*:3])[*:6].O",
  "primary amine oxidation": "[Nh2+0:1][Ch:2]>>[N:1]=[*:2]",
  "secondary amine oxidation": "[Ch:2][Nh+0:1][C^3:3]>>[*:2]=[N:1][*:3]",
  “primary imine-amine attack": "[N:1]=[C^2:2].[Nh2X3:3][Ch:4]>>[Nh:1][C:2][*:3][*:4]",
  "secondary imine-amine attack":
"[N:1]=[C^2:2].[NhX3:3]([Ch:4])[Ch:5]>>[Nh:1][C:2][*:3]([*:4])[*:5]",
  "amine elimination 1": "[Nh2:1][C:2][ND1X3+0,ND2X3+0:3]>>[C:2]=[*:3].[*:1]",
  "amine elimination 2": "[Nh2:1][C:2][ND3X3+0:3]>>[C:2]=[N+:3].[*:1]",
  "imine reduction": "[N,N+1:1]=[C^2:2]>>[N+0:1][C:2]",
  "imine hydrolysis": "[C^2:1]=[NX2+0:2].[Oh2:3]>>[C:1]=[*:3].[N:2]",
  "isocyanate to urea": "O=[C:1]=[N:2].[Nh2:3]>>O=[C:1]([NH:2])[Nh:3]",
  "isocyanide alpha attack": "[C:1]=[N:2].[Ch:3][N+:4]#[C-:5]>>[C:1]([N:2])[*:3][N+:4]#[C-:5]",
  "carbanion attack": "[C:1]=[N:2].[C-:3]>>[C+0:3][C:1][N:2]",
  "isocyanide oxidation": "[N+:1]#[C-:2].[CH3]S([CH3])=[O:3]>>[N+0:1]=[C+0:2]=[O:3]",
  "isocyanate hydrolysis": "O=C=[N:1].[OH2]>>[N:1].O=C=O",
  "nitrilium hydrolysis": "[C+0:1]#[N+:2].[Oh2]>>[C:1](=O)O.[N+0:2]",
  "Ugi reaction": "[Nh:1][CX4:2][C:3]#[N+:4].O=[C:5][Oh:6]>>O=[C:5][*:1][*:2][C:3](=[*:6])[N+0:4]",
  "imine tautomerism": "[Ch^3:1]-[NX2+0:2]=[C^2:3]>>[*:1]=[*:2]-[*:3]",
  "decarboxylation": "[C:1](C(=O)[Oh])[C:2](=[O:3])>>[C:1][C:2](=[O:3]).O=C=O",
  "Ts elimination": "O=[S:3](=O)[CX4:1][ChX4:2]>>[C:1]=[C:2]",
  "Ts
reduction":“O=[S:3](=O)[CX4:1]([C^2,C^1,N^1:2])[C^2,C^1,N^1:4]>>[C:1]([C^2,C^1,N^1:2])[C^2,C^
1,N^1:4]"
}
```

Figure S40: List of reaction SMARTS used to create the chemical space network.

5.2 Similarity index with known TosMIC reactions

1656 reactions involving TosMIC as reagent were gathered from the Reaxys database. The reaction fingerprints were calculated by subtracting reagent(s) fingerprints of from product fingerprint(s). Since the fingerprint values were generated from the molecular structure the difference for a similar reaction will be close to a vector of zeroes, while two distinct reactions will produce a vector with high values, positive and negative. In order to quantify and compare the reaction fingerprints we calculated the ℓ-2-norm, the square root of the sum of the squares. Values closed to zero indicate a high structural similarity between reagent(s) and product(s). The examples of the two reactions of TosMIC reactions showing a higher ℓ-2-norm than the discovered one are reported in **Figure S41**. They are variations of polymerization reactions and Ugi four-component reactions.

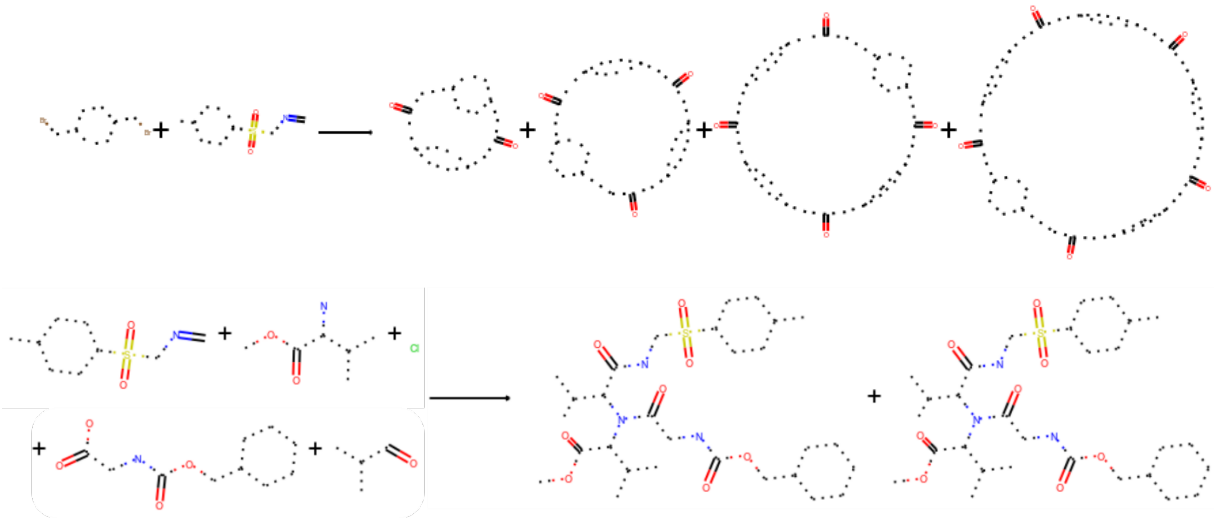


Figure S41: Templates of the two main types of reactions presenting a higher difference generation.

Next, the dataset has also been used to compare the discovered reaction with known TosMIC reactions using the Tanimoto similarity index[44]. A matrix of cross correlation between the reaction fingerprints is showed in **Figure S42**. The Tanimoto index between reaction leading to product **25** and the other literature reactions is in the last column/row and expanded in the bottom part of the figure for visualization. The average similarity was 0.091. A histogram comparing this value with the similarity averages of the other reactions is showed in **Figure S43**. As reference the reactions with highest similarity to the discovered reaction are showed in **Figure S44**. The script used to perform these calculations was written in Python using the RDKit library.

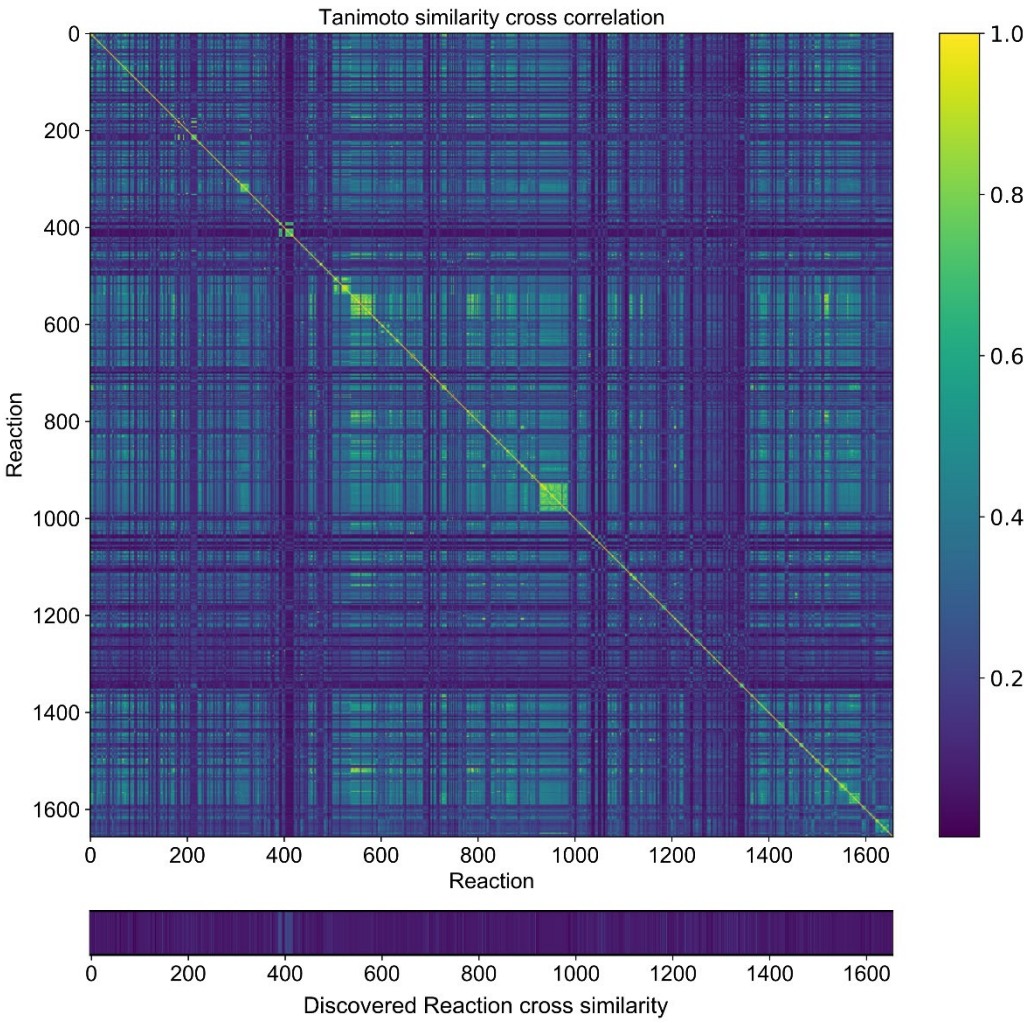


Figure S42: Tanimoto similarity index between reaction fingerprints of the literature reactions of TosMIC. In the bottom the similarity index between literature reactions and discovered reaction giving product **25.**

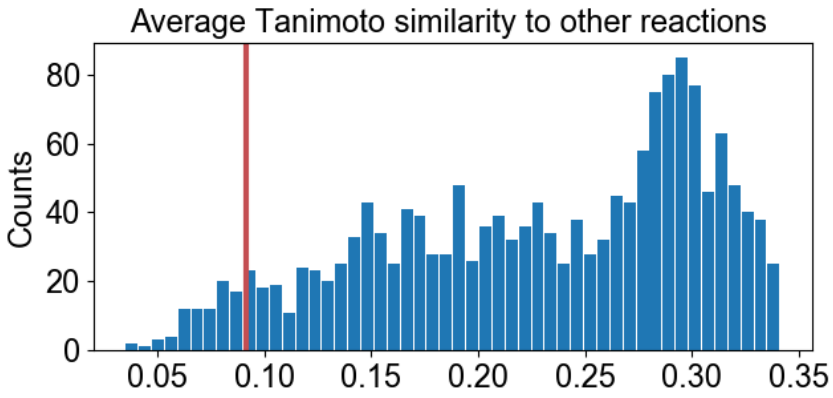


Figure S43: Histogram comparing the average similarity for each reaction to other reactions in **Figure S42**.

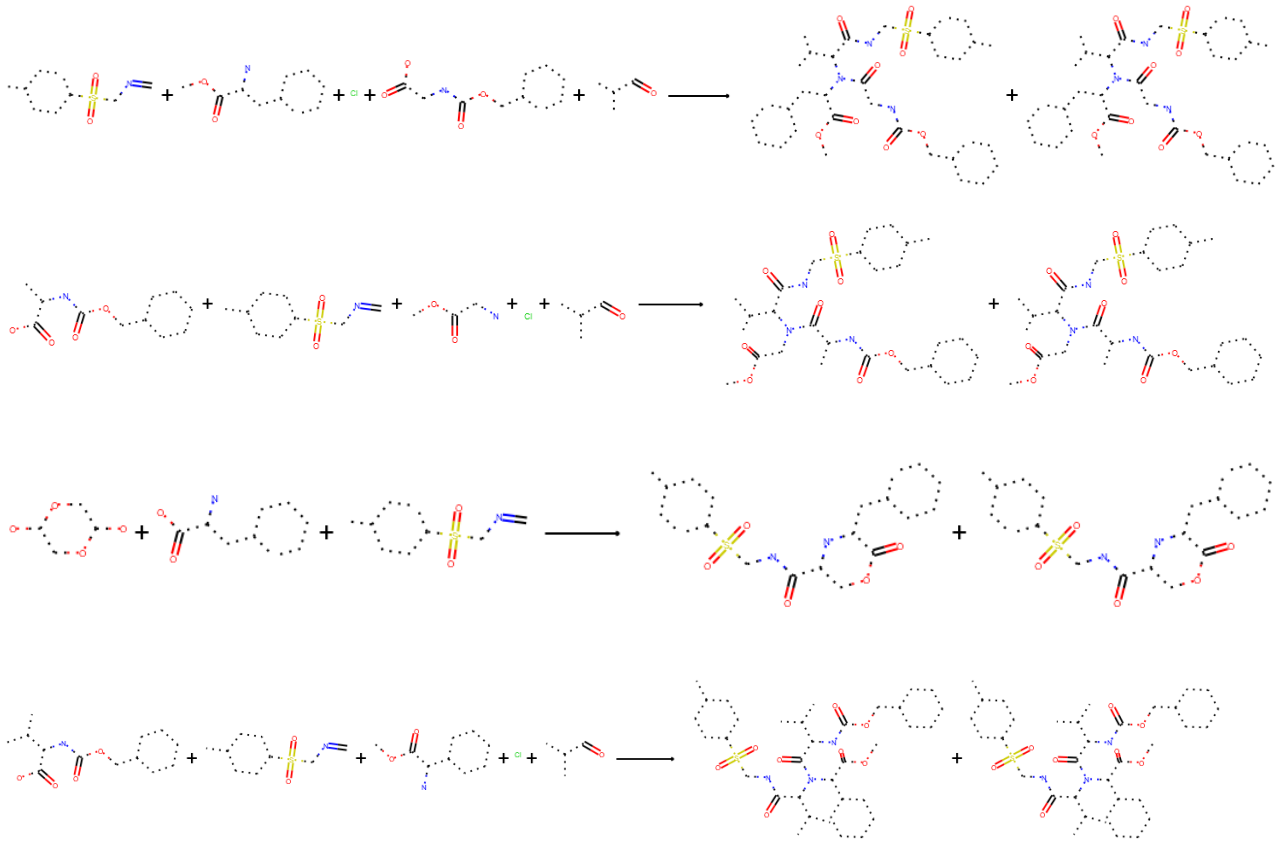


Figure S44: Reactions having the most similar reaction fingerprints to discovered reaction**,** according to Tanimoto similarity index.

6 Other reactions discovered and re-discovered

A total of 7 reactions have been analysed. One of the yielded product **25** and was further investigated. From two of them despite the evidence of reactivity in NMR analysis and TLC it was not possible to isolate any products in purity high enough for characterisation. The first one was the reaction of 1,3-Diethyl-2-thiobarbituric acid and TosMIC while the second was the reaction of 1,3-Diethyl-2-thiobarbituric acid and phenylhydrazine. Both of them involved the presence of tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate and 450 nm irradiation. The remaining four reactions are reported below.

6.1 Phenylhydrazine and bromoacetonitrile under 450 nm irradiation

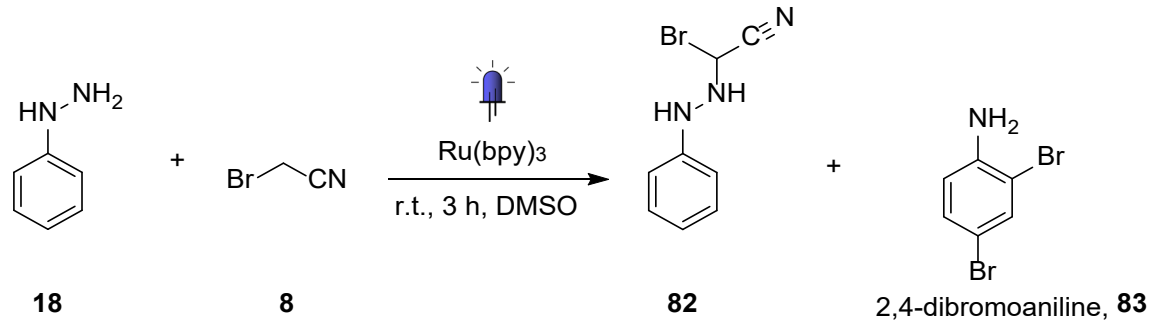


Figure S45: Scheme of the reaction between phenylhydrazine and bromoacetonitrile, under 450 nm irradiation and the $Ru(bpy)_3$ photocatalyst. The first product is unreported in literature.

Phenylhydrazine (2 mmol, 0.23 ml), Bromoacetonitrile (2 mmol, 0.41 ml) and Tris(2,2'-bipyridyl) tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate (2.5% mol, 32 mg) are mixed in 4.5 ml of DMSO. The reaction is stirred at room temperature and irradiated with 450 nm LED for 3 hours. The reaction mixture is diluted with water and extracted with ethyl acetate. The organic phase is separated and washed three times with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under reduced pressure. The crude is purified with a chromatographic column (silica gel, Hexane/ EtOAc 80:1), isolating both Compound **82** and 2,4-dibromoaniline **83**. **82** is obtained with a photocatalytic addition where a new C-N bond is formed but the bromide, usually a leaving group, is kept in its place. This reaction is unreported in literature. As further tests the reaction was repeated in presence of light but without photocatalysts and in presence of 2,4,6-triphenylpyrylium tetrafluoroborate. Both tests resulted negative as no product formation was detected by TLC.

Compound **82**

Yield: 7% (31 mg)

IUPAC name: Bromo-(N'-phenyl-hydrazino)-acetonitrile

^{1}H NMR (600 MHz, $CDCl_3$) δ 8.90 (s, 1H), 7.37 (d, J = 16.0 Hz, 2H), 7.18 (d, J = 7.7 Hz, 2H), 7.09 (t, J = 7.4 Hz, 1H), 6.24 (s, 1H)

^{13}C NMR (151 MHz, $CDCl_3$) δ 141.22, 129.14, 123.22, 113.77, 110.68, 99.62.

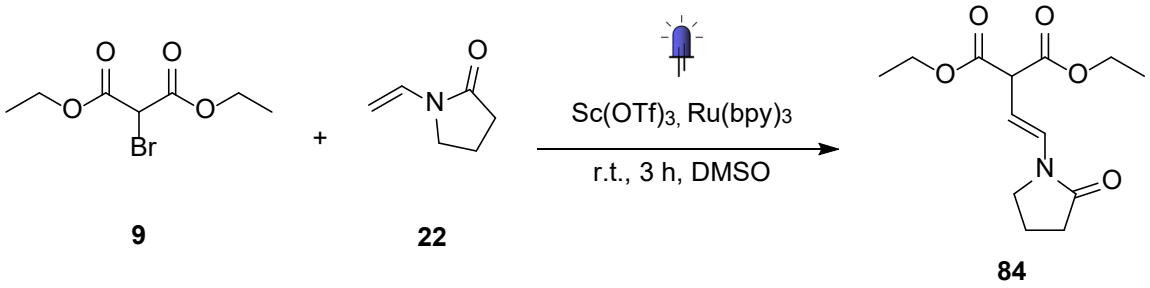
Crystallographic data: See section 7

2,4-dibromoaniline, **83**

Yield: 11% (55 mg)

^{1}H NMR (600 MHz, $CDCl_3$) δ 7.56 (d, *J* = 2.2 Hz, 1H), 7.22 (dd, *J* = 2.2, 8.5 Hz, 1H), 6.66 (d, *J* = 8.6 Hz, 1H), 4.11 (s, 1H), validated with commercial material.

6.2 Diethyl 2-bromomalonate and 1-vinyl-2-pyrrolidinone under 450 nm irradiation



9 **22** **84**

Figure S46: Scheme of the reaction between diethyl 2-bromomalonate and 1-vinyl-2-pyrrolidinone under 450 nm irradiation and the $Ru(bpy)_3$ photocatalyst.

Diethyl 2-bromomalonate (2 mmol, 0.34 ml), 1-Vinyl-2-pyrrolidinone (2 mmol, 0.21 ml) and tris(2,2'-bipyridyl)dichlororuthenium(II) hexahydrate (2.5% mol, 32 mg), scandium triflate (2.5% mol, 25 mg) are mixed in 4.5 ml of DMSO. The reaction is stirred at room temperature and irradiated with 450 nm LED for 3 hours. The reaction mixture is diluted with water and extracted with ethyl acetate. The organic phase is separated and washed three times with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under vacuum. The crude mixture is purified with chromatographic column (silica gel, hexane/ EtOAc 2:1). The reaction is a C-H functionalization made through photoredox catalysis. A similar reaction has been already described in 2012[45] where the authors used the same reagents, $[Ir(ppy)_2(dtbbpy)]PF_6$ as photocatalyst, 2 equivalents of Na_2HPO_4 and acetonitrile as solvent.

Yield: 53% (285 mg)

IUPAC name: (E)-Diethyl 2-[2-(2-oxopyrrolidin-1-yl)vinyl]malonate

^{1}H NMR (600 MHz, $CDCl_3$) δ 7.00 (d, J = 14.4 Hz, 1H), 5.09 (dd, J = 14.5, 9.5 Hz, 1H), 4.14 (qd, J = 7.1, 3.6 Hz, 4H), 3.96 (d, J = 9.5 Hz, 1H), 3.51 (t, J = 7.2 Hz, 2H), 2.42 (t, J = 8.1 Hz, 2H), 2.06 (p, J = 7.7 Hz, 2H), 1.21 (t, J = 7.1 Hz, 6H).

13C NMR (151 MHz, CDCl3) δ 173.17, 168.20, 127.74, 102.82, 61.55, 53.24, 44.82, 30.83, 17.23, 13.80.
ESI-HR-MS: $[C_{13}H_{19}NNaO_5]^+$ Calculated 292.1155 *m/z*, measured 292.1147 *m/z*

6.3 *N,N*-dimethyl-4-nitrosoaniline, bromoacetonitrile and diethyl 2-bromomalonate

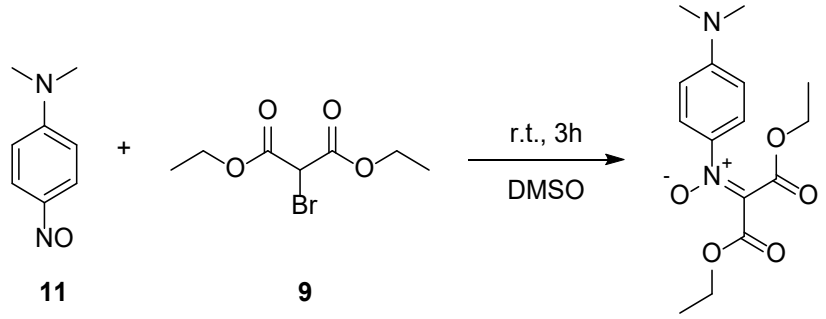


Figure S47: Reaction reported between *N,N*-dimethyl-4-nitrosoaniline and diethyl-2-bromomalonate.

Bromoacetonitrile (2 mmol, 0.41 ml), N,N-Dimethyl-4-nitrosoaniline (2 mmol, 0.3 g) and Diethyl 2-bromomalonate (2 mmol, 0.34 ml) are mixed in 6 ml of DMSO. The mixture is stirred at room temperature for 3 hours. The reaction is then diluted with water and extracted with ethyl acetate. The organic phase is separated and washed three times with brine. Mg_2SO_4 is then added to the reaction mixture and after filtration the solvent is removed under vacuum. The crude is purified with a chromatographic column (silica gel, Hexane / EtOAc 10:1). The reaction is an addition of nitroso group on the diethyl-2-bromomalonate to form the respective nitrone. It is already known in literature[46] where it usually involves sodium hydroxide and THF as solvent.

Yield: 18% (111 mg)
IUPAC name: N-(4-(dimethylamino)phenyl)-C,C-diethoxy-carbonylnitrone
^{1}H NMR (600 MHz, $CDCl_3$) δ 7.37 (d, *J* = 9.1 Hz, 1H), 6.63 (d, *J* = 9.1 Hz, 2H), 4.43 (q, *J* = 7.1 Hz, 2H), 4.19 (q, *J* = 7.1 Hz, 2H), 3.05 (s, 6H), 1.40 (t, *J* = 7.1 Hz, 3H), 1.21 (t, *J* = 7.1 Hz, 3H)
^{13}C NMR (151 MHz, $CDCl_3$) δ 160.72, 159.43, 151.93, 135.63, 130.37, 124.47 110.18, 61.79, 61.53, 39.79, 13.46
ESI-HR-MS: $[C_{15}H_{20}N_2NaO_5]^+$ Calculated 331.1264 *m/z*, measured 331.1248 *m/z*.

6.4 Diethyl 2-bromomalonate, 4-phenylurazole and 1-vinyl-2-pyrrolidinone

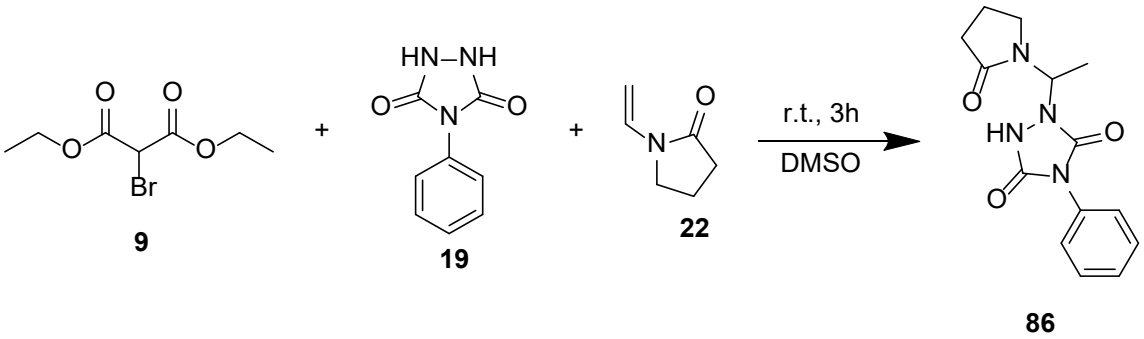


Figure S48: Scheme of the reaction between 4-phenylueazole, 1-vinyl-2-pyrrolidinone, and diethyl-2-bromomalonate.

Diethyl 2-bromomalonate (2 mmol, 0.34 ml), 4-phenylurazole (2 mmol, 0.354 g) and 1-Vinyl-2-pyrrolidinone (2 mmol, 0.23 ml) are mixed in 6 ml of DMSO. The reaction is stirred at room temperature for 3 hours. The mixture is diluted with of water, extracted with ethyl acetate and washed three times with brine. The organic phase is dried with Mg_2SO_4 and the solvent is removed under vacuum. During the evaporation the product precipitates as white crystals, they are filtered and washed with cold ethyl acetate.

The reaction is a nucleophile addition of phenylurazol nitrogen on the pyrrolidinone double bond. A similar reaction has been reported by Senogles[47] *et al* in 1980 and involved the hydrolysis of 1-vinyl-2-pyrrolidinone in aqueous solutions.

Yield: 16% (92 mg)

IUPAC name: 1-[1-(2-Oxo-pyrrolidin-1-yl)-ethyl]-4-phenyl-[1,2,4]triazolidine-3,5-dione

^{1}H NMR (600 MHz, DMSO) δ 10.69 (s, 1H), 7.42 (m, 5H), 5.90 (s, 1H), 3.46 (m, J = 7.0 Hz, 2H), 2.26 (m, J = 3.5 Hz, 2H), 1.96 (m, J = 8.7 Hz, 2H), 1.50 (s, 3H)

^{13}C NMR (151 MHz, DMSO) δ 174.37, 153.00, 152.32, 131.56, 128.91, 128.00, 126.21, 60.13, 43.20, 30.44, 17.77

ESI-HR-MS: $[C_{14}H_{15}N_4O_3]^-$ Calculated 287.1150 *m/z*, measured 287.1430 *m/z*

Crystallographic data: See section 7

7 Crystal structure details

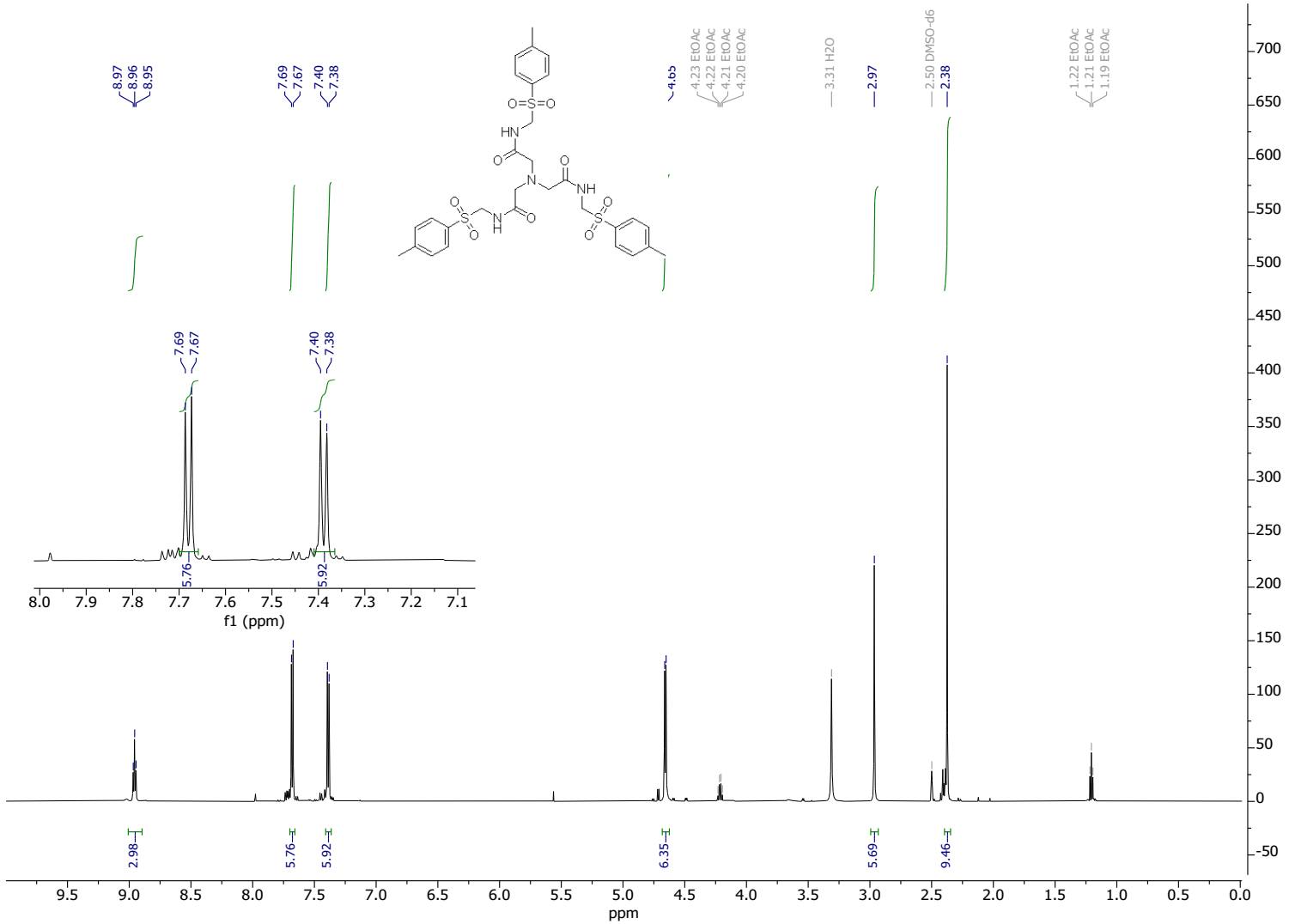
Single Crystal X-ray Diffraction: Suitable single crystals were selected and mounted by using the MiTeGen MicroMountsTM kit with Fomblin oil. X-ray diffraction intensity data were measured at 150(2) K on Bruker Apex II Quasar diffractometer using Mo Kα [λ = 0.71073 Å] radiation. Structure solution and refinement were carried out with SHELXT[48] and SHELXL-2018[49] via WinGX[50]. Corrections for incident and diffracted beam absorption effects were applied using empirical methods[51].

Name	**Compound 25**	
Empirical formula	C31 H39 N4 O9.50 S3.50	
Formula weight	731.87	
Temperature	150(2) K	
Wavelength	0.71073 Å	
Crystal system	Trigonal	
Space group	P -3 c 1	
Unit cell dimensions	a = 22.260(8) Å	α= 90°
	b = 22.260(8) Å	β= 90°
	c = 9.328(4) Å	γ = 120°
Volume	4003(3) Å3	
Z	4	
Density (calculated)	1.214 Mg/m3	
Absorption coefficient	0.263 mm-1	
F(000)	1540	
Crystal size	0.232 x 0.030 x 0.016 mm3	
Theta range for data collection	2.113 to 23.318°	
Index ranges	-24<=h<=24, -24<=k<=24, -10<=l<=10	
Reflections collected	30636	
Independent reflections	1937 [R(int) = 0.2800]	
Completeness to theta = 23.318°	99.70%	
Max. and min. transmission	0.745 and 0.634	
Refinement method	Full-matrix least-squares on F2	
Data / restraints / parameters	1937 / 0 / 140	
Goodness-of-fit on F2	1.043	
Final R indices [I>2sigma(I)]	R1 = 0.0631, wR2 = 0.1369	
R indices (all data)	R1 = 0.1366, wR2 = 0.1760	
Extinction coefficient	n/a	
Largest diff. peak and hole	0.31 and -0.45 e.Å-3	

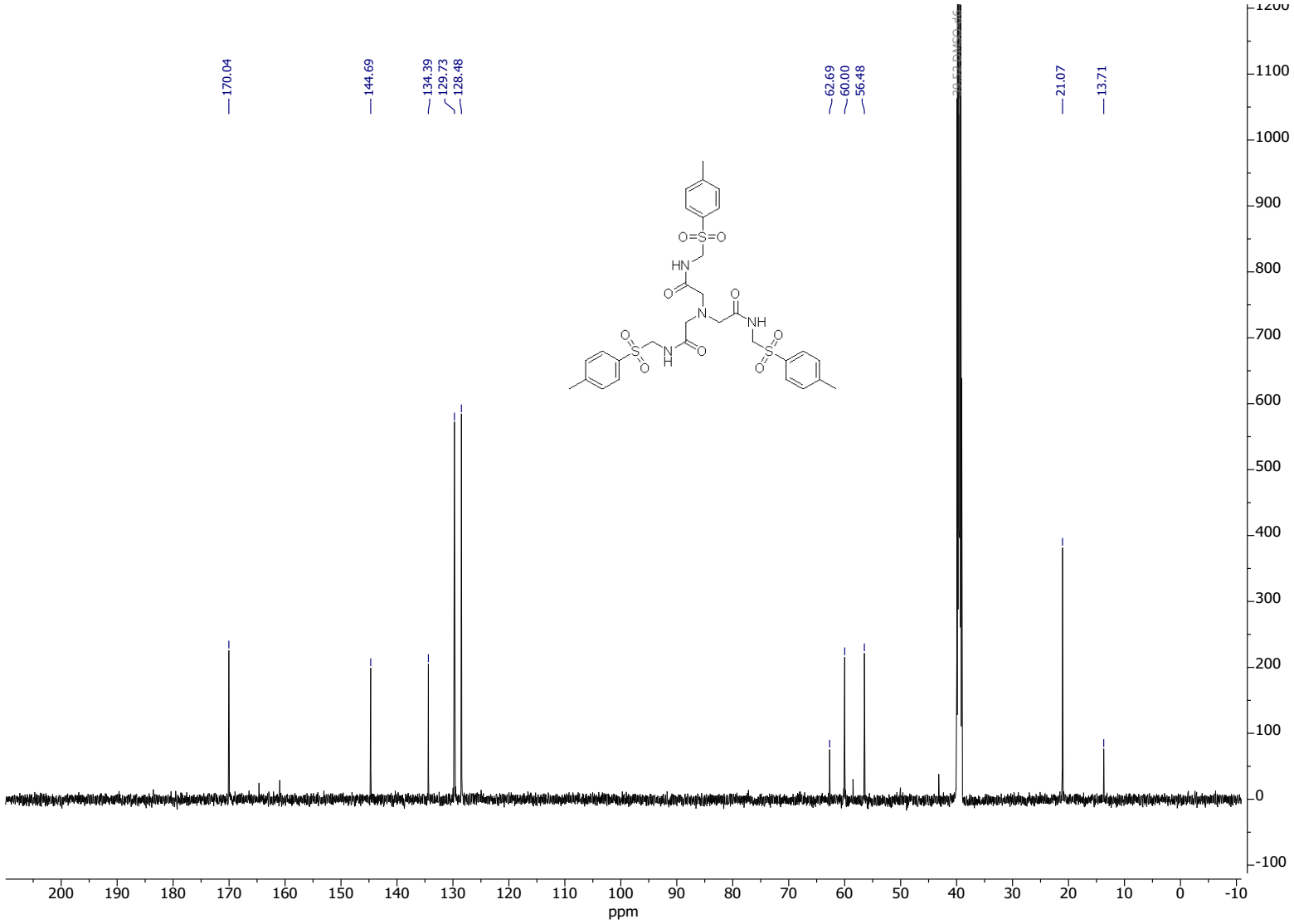
Name	**Compound 82**	
Empirical formula	$C_8H_5BrN_3$	
Formula weight	223.06	
Temperature	150(2) K	
Wavelength	0.71073 Å	
Crystal system	Orthorhombic	
Space group	Pna21	
Unit cell dimensions	a = 12.777(4) Å	α= 90°
	b = 4.5611(14) Å	β= 90°
	c = 14.595(5) Å	γ = 90°
Volume	850.6(5) $Å^3$	
Z	4	
Density (calculated)	1.742 Mg/m^3	
Absorption coefficient	4.775 mm^{-1}	
F(000)	436	
Crystal size	0.157 x 0.052 x 0.040 mm^3	
Theta range for data collection	2.791 to 25.993°	
Index ranges	-15<=h<=15, -5<=k<=5, -18<=l<=18	
Reflections collected	13942	
Independent reflections	1677 [R(int) = 0.0588]	
Completeness to theta = 25.242°	100.00%	
Absorption correction	Empirical	
Max. and min. transmission	0.728 and 0.605	
Refinement method	Full-matrix least-squares on F^2	
Data / restraints / parameters	1677 / 1 / 109	
Goodness-of-fit on F^2	1.107	
Final R indices [I>2sigma(I)]	R_1 = 0.0520, wR_2 = 0.1558	
R indices (all data)	R_1 = 0.0582, wR_2 = 0.1615	
Absolute structure parameter	0.082(12)	
Extinction coefficient	n/a	
Largest diff. peak and hole	0.75 and -1.17 $e.Å^{-3}$	

Name	**Compound 86**	
Empirical formula	$C_{14}H_{16}N_4O_3$	
Formula weight	288.31	
Temperature	150(2) K	
Wavelength	0.71073 Å	
Crystal system	Orthorhombic	
Space group	Pbca	
Unit cell dimensions	a = 11.972(4) Å	α= 90°
	b = 12.045(5) Å	β= 90°
	c = 18.406(7) Å	γ = 90°
Volume	2654.1(17) $Å^3$	
Z	8	
Density (calculated)	1.443 Mg/m^3	
Absorption coefficient	0.105 mm^{-1}	
F(000)	1216	
Crystal size	0.140 x 0.042 x 0.012 mm^3	
Theta range for data collection	2.213 to 26.000°	
Index ranges	-14<=h<=14, -14<=k<=14, -22<=l<=22	
Reflections collected	30451	
Independent reflections	2600 [R(int) = 0.1131]	
Completeness to theta = 25.242°	100.00%	
Absorption correction	Empirical	
Max. and min. transmission	1.000 and 0.852	
Refinement method	Full-matrix least-squares on F^2	
Data / restraints / parameters	2600 / 0 / 191	
Goodness-of-fit on F^2	1.032	
Final R indices [I>2sigma(I)]	R_1 = 0.0514, wR_2 = 0.1132	
R indices (all data)	R_1 = 0.0938, wR_2 = 0.1368	
Extinction coefficient	n/a	
Largest diff. peak and hole	0.33 and -0.55 e.$Å^{-3}$	

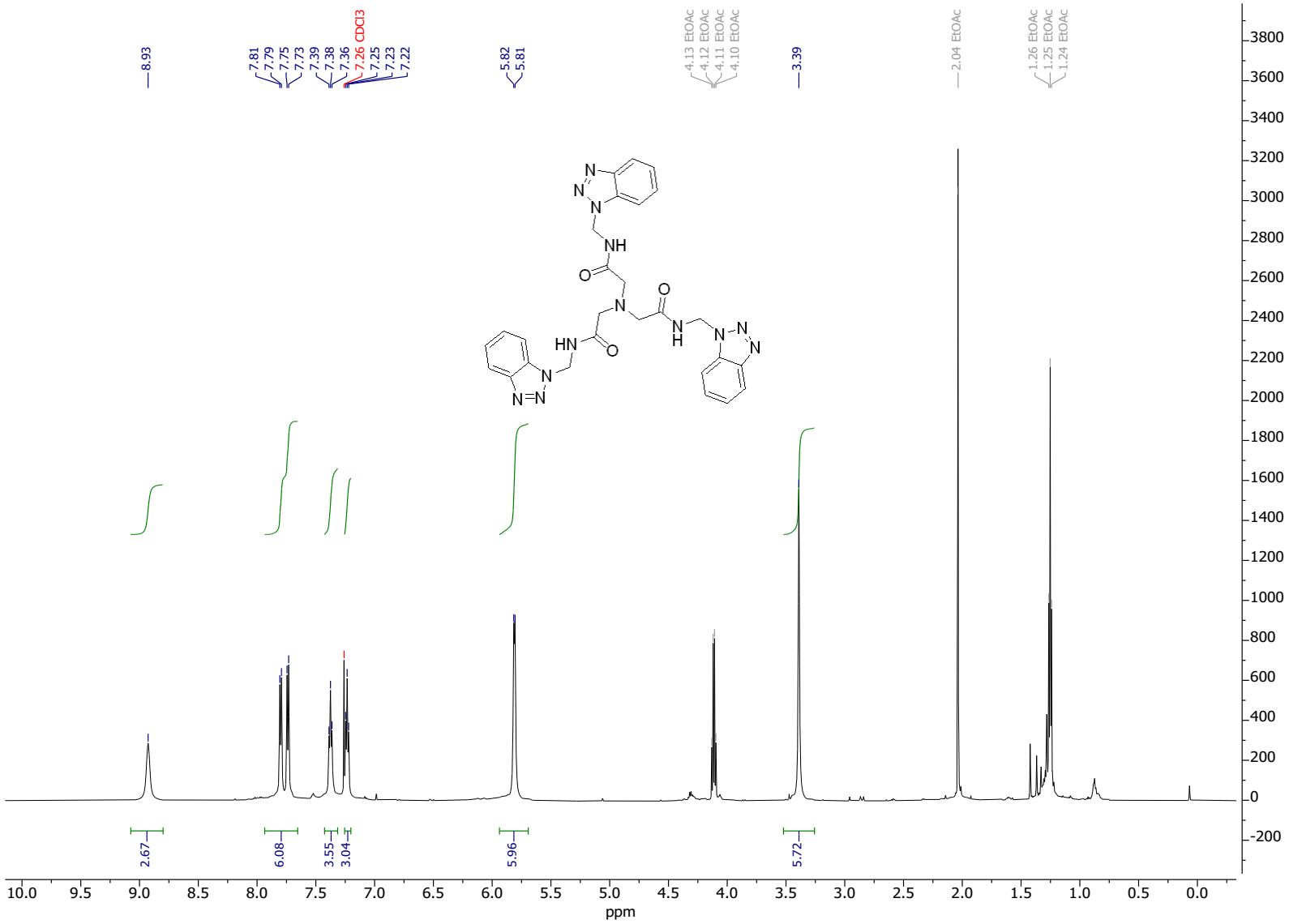
8 ^{1}H and ^{13}C NMR spectra



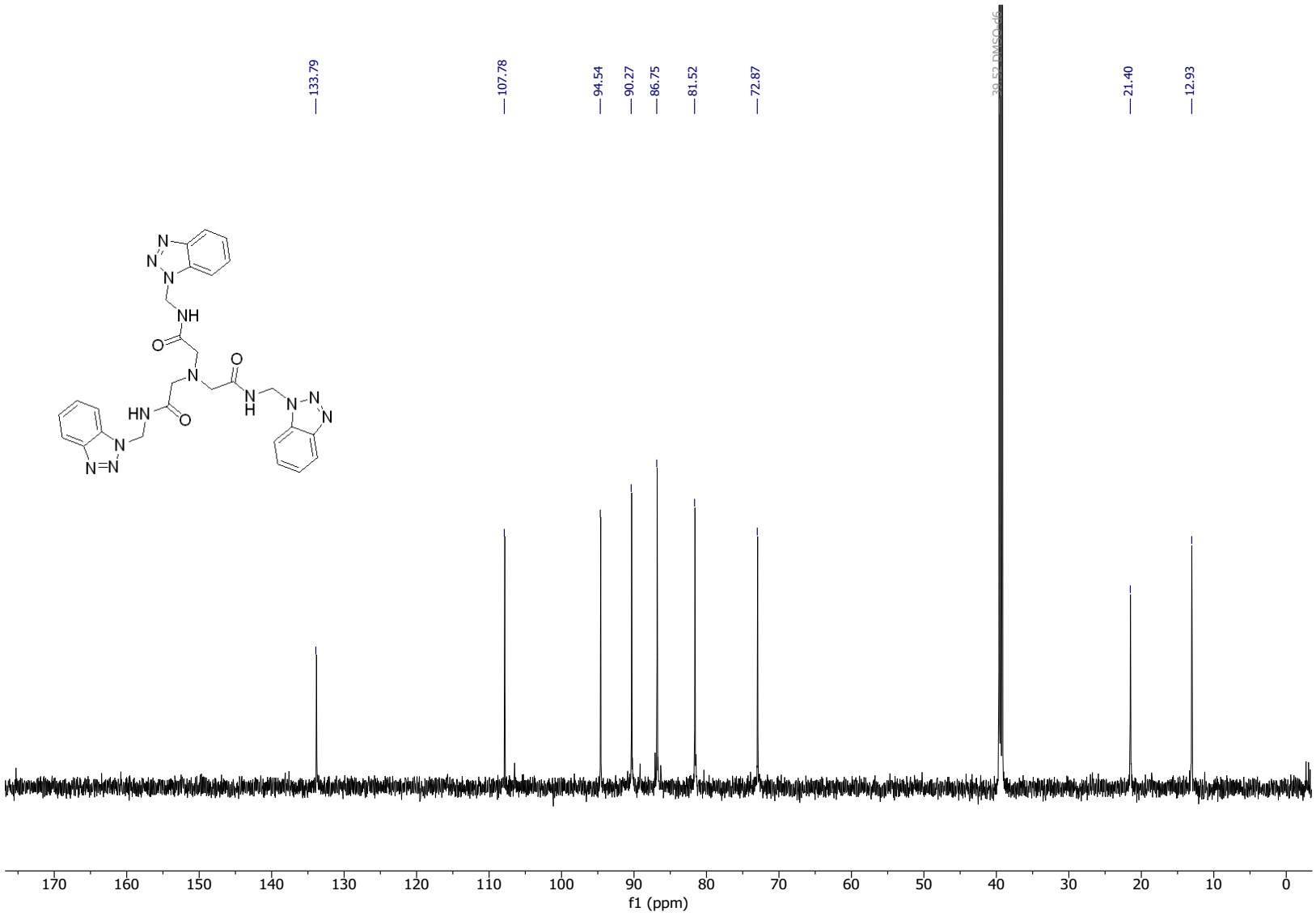
Spectrum S1: ^{1}H-NMR of Product **25**



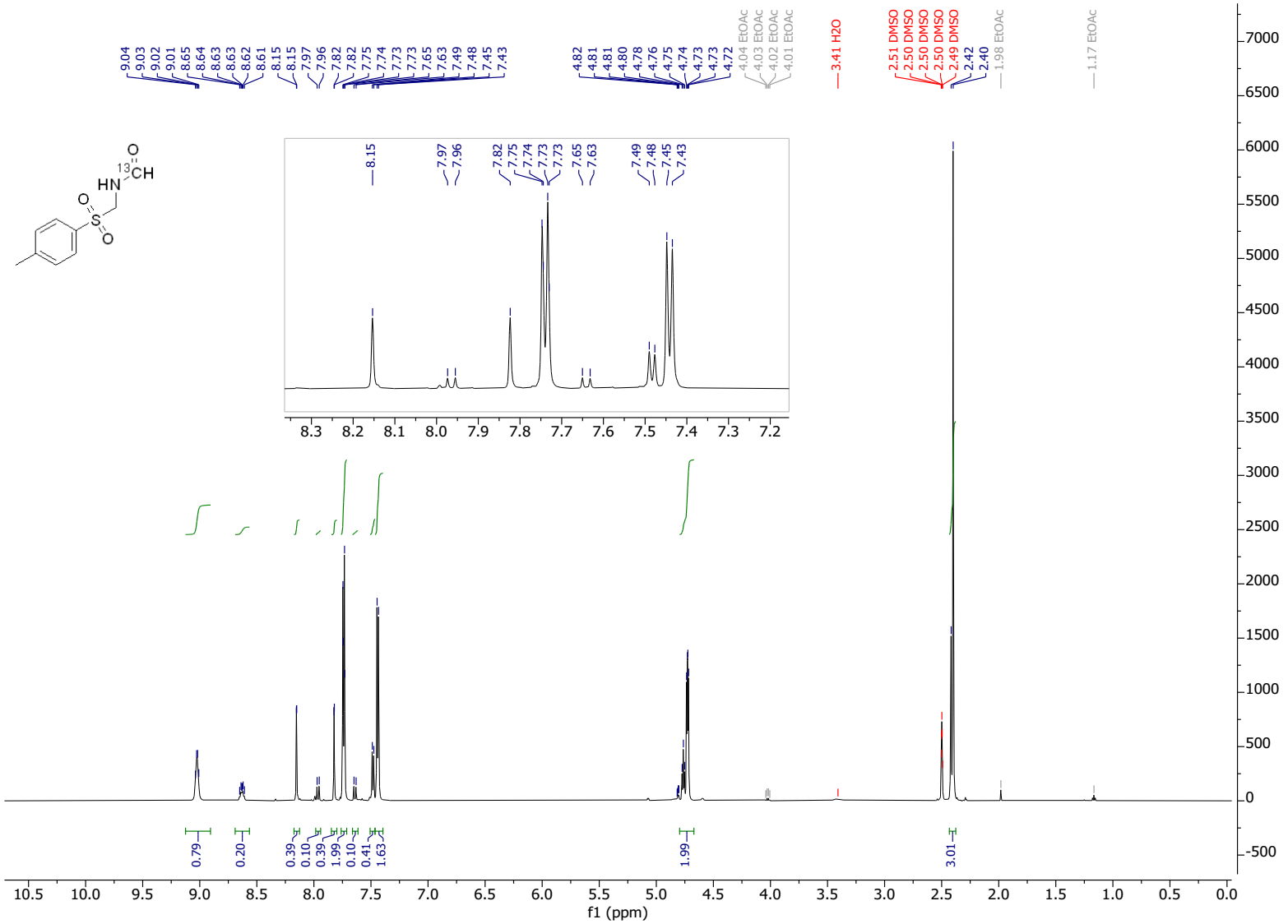
Spectrum S2: ^{13}C-NMR of Product **25**



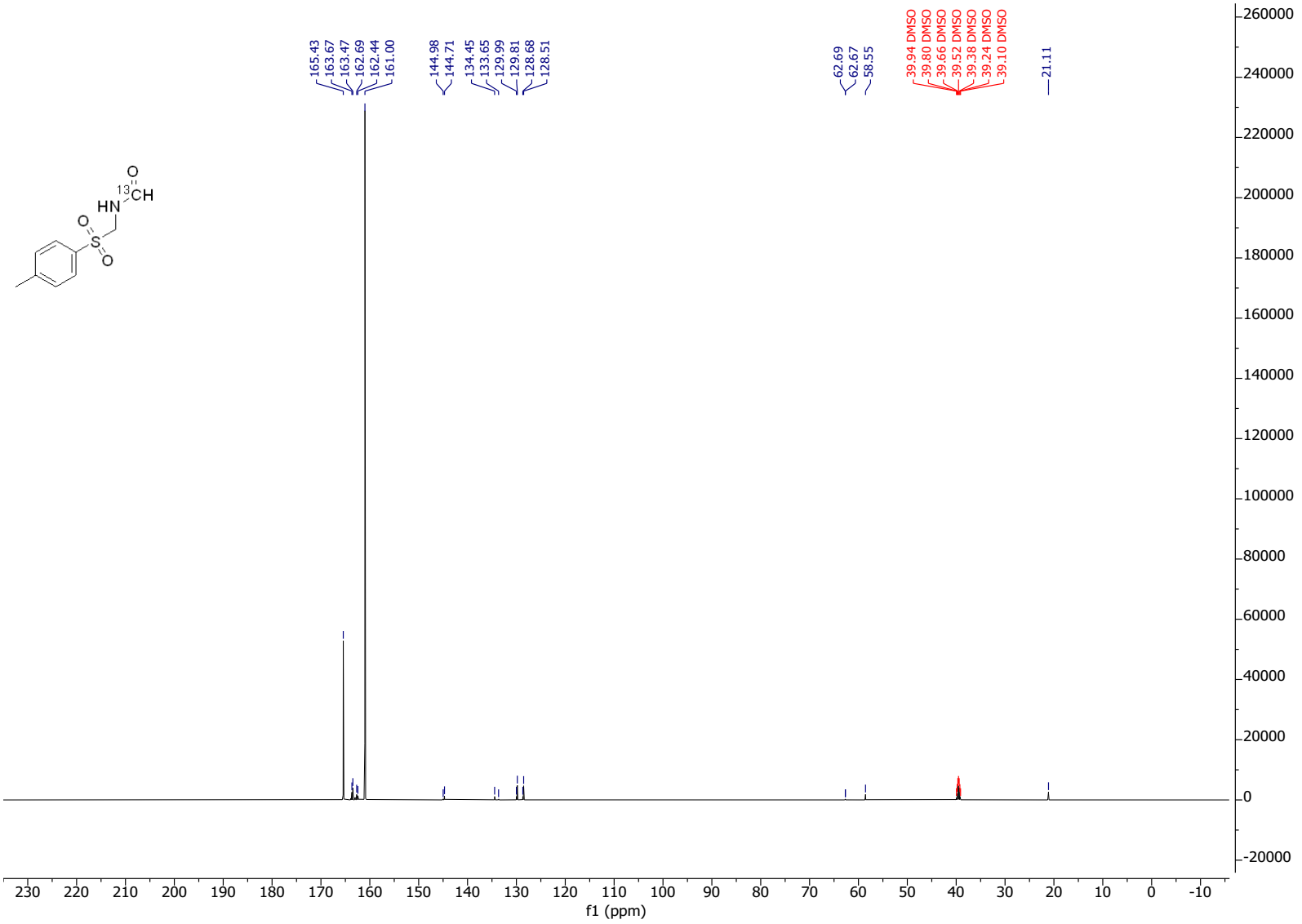
Spectrum S3: ^{1}H-NMR of Product **28**



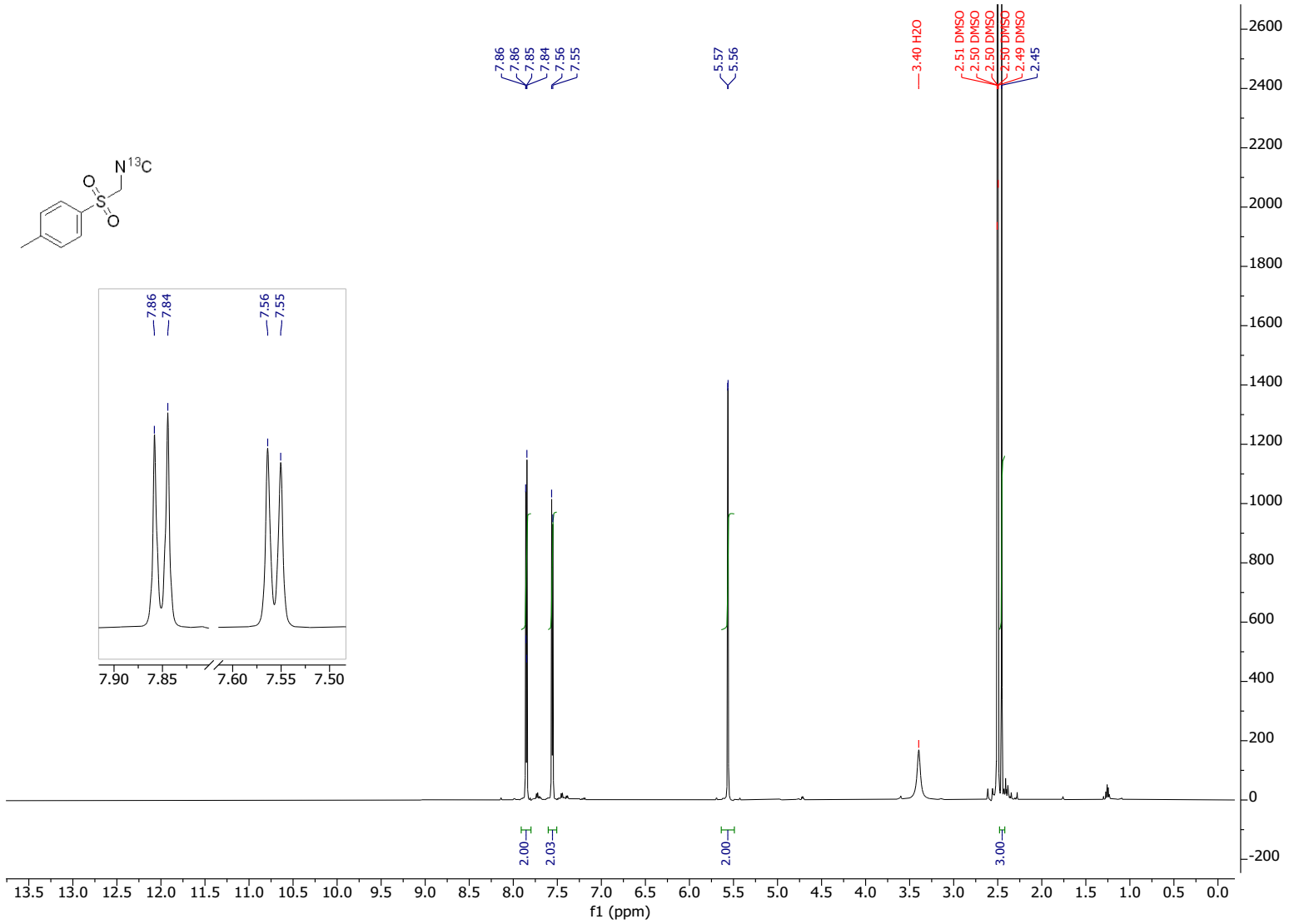
Spectrum S4: ^{13}C-NMR of Product **28**



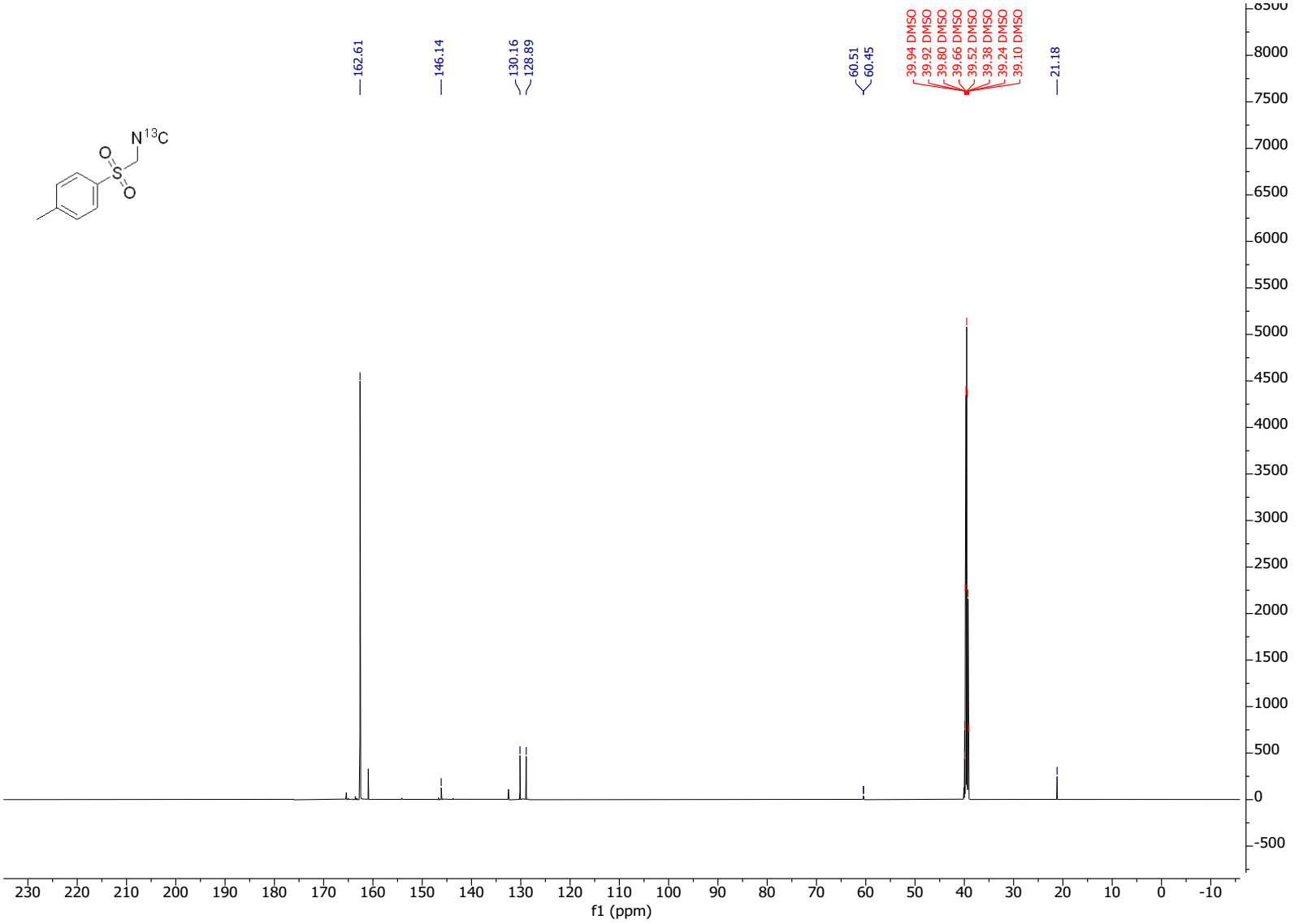
Spectrum S5: ^{1}H-NMR of (1-^{13}C)N-(tosylmethyl)-formamide



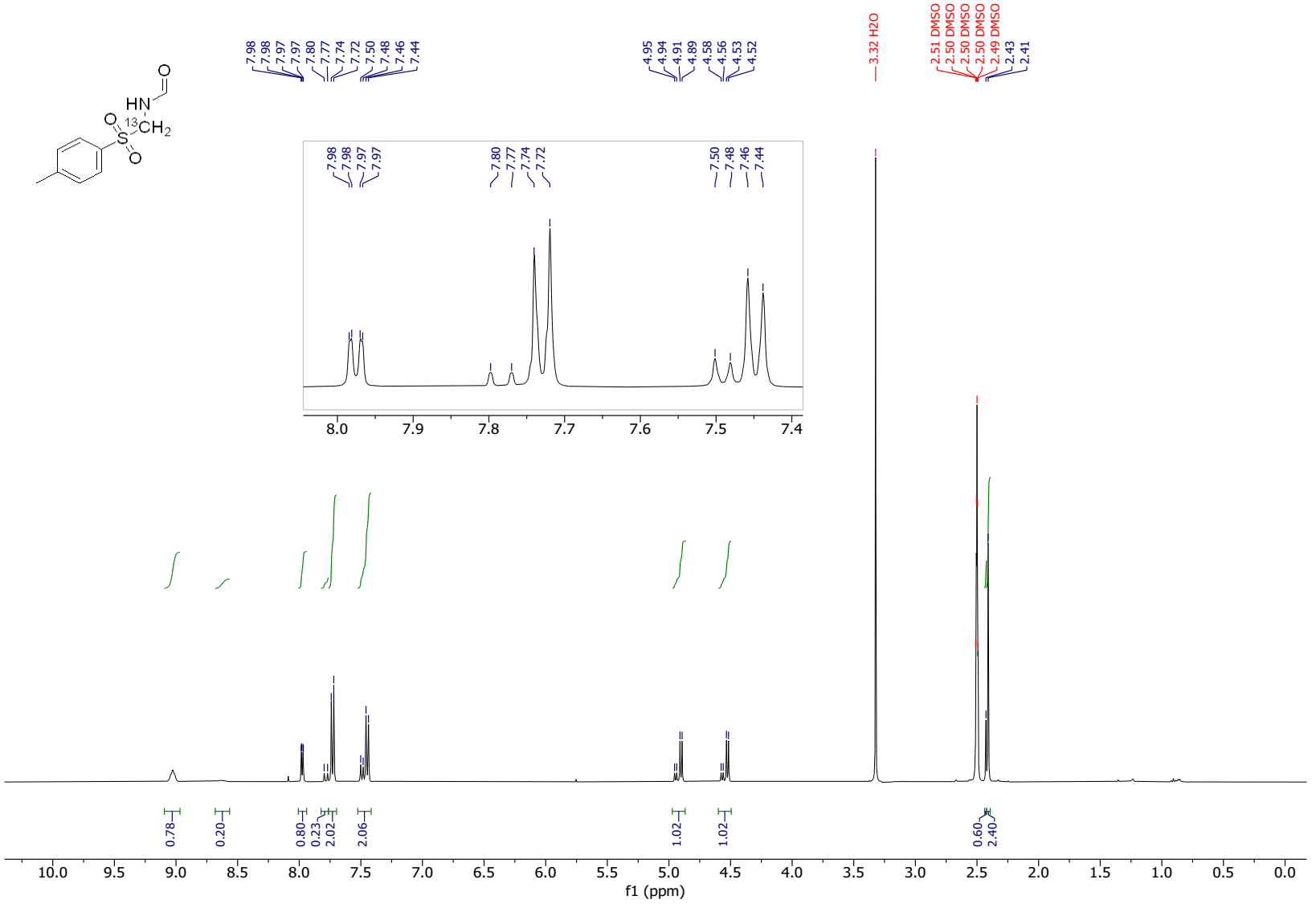
Spectrum S6: ^{13}C-NMR of (1-^{13}C)N-(tosylmethyl)-formamide



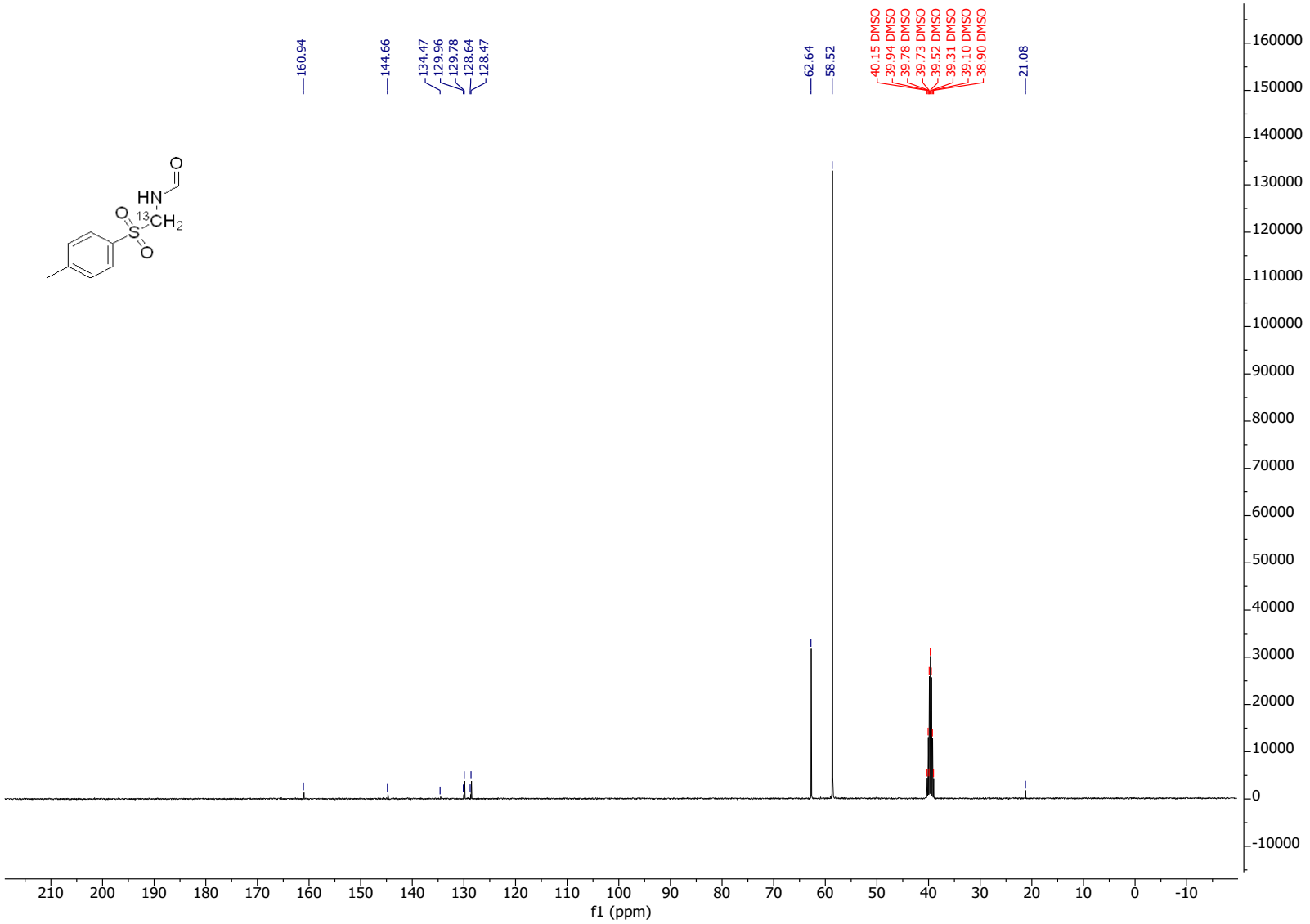
Spectrum S7: ^{1}H-NMR of $TosCH_2N^{13}C$



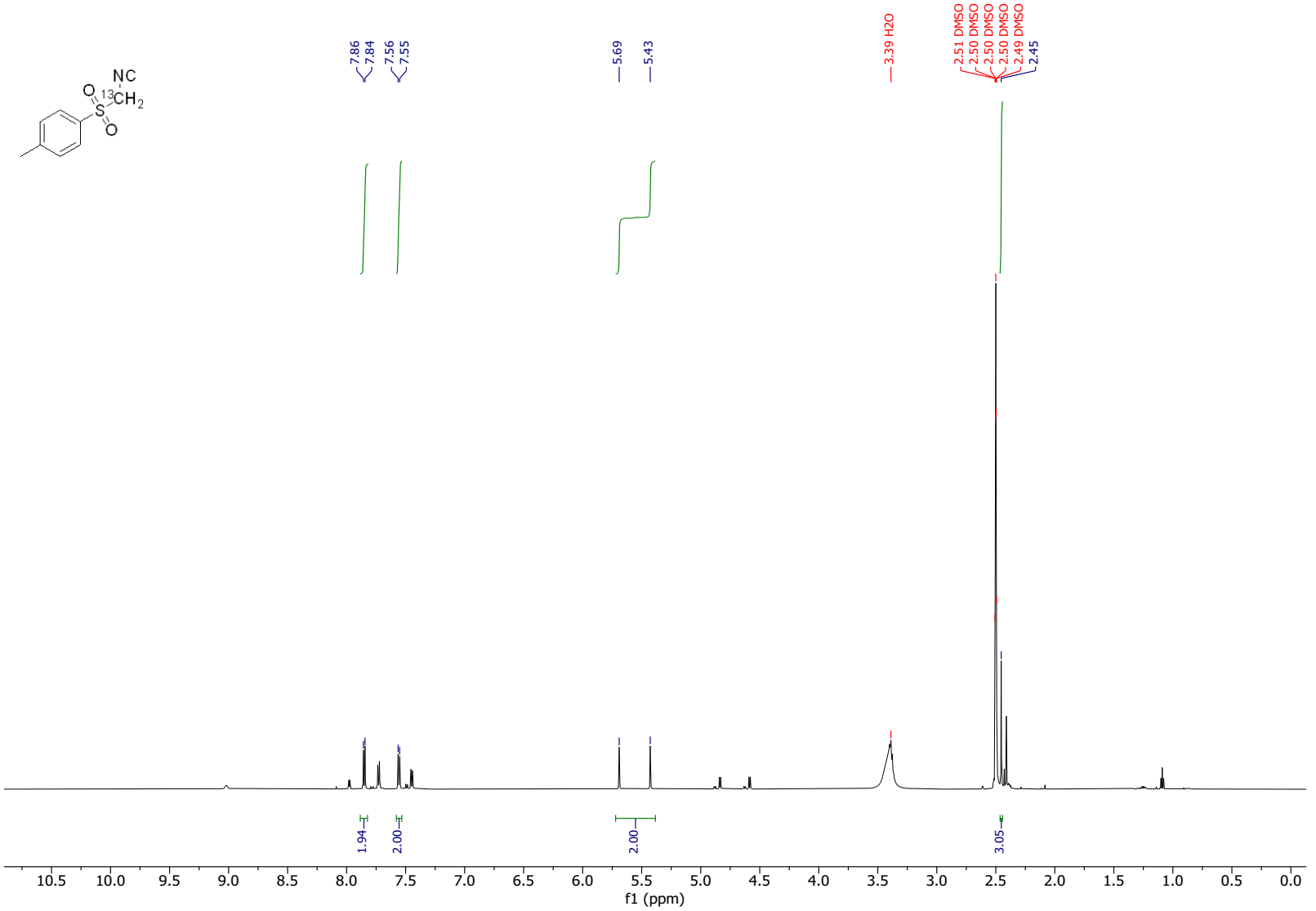
Spectrum S8: ^{13}C-NMR of $TosCH_2N^{13}C$



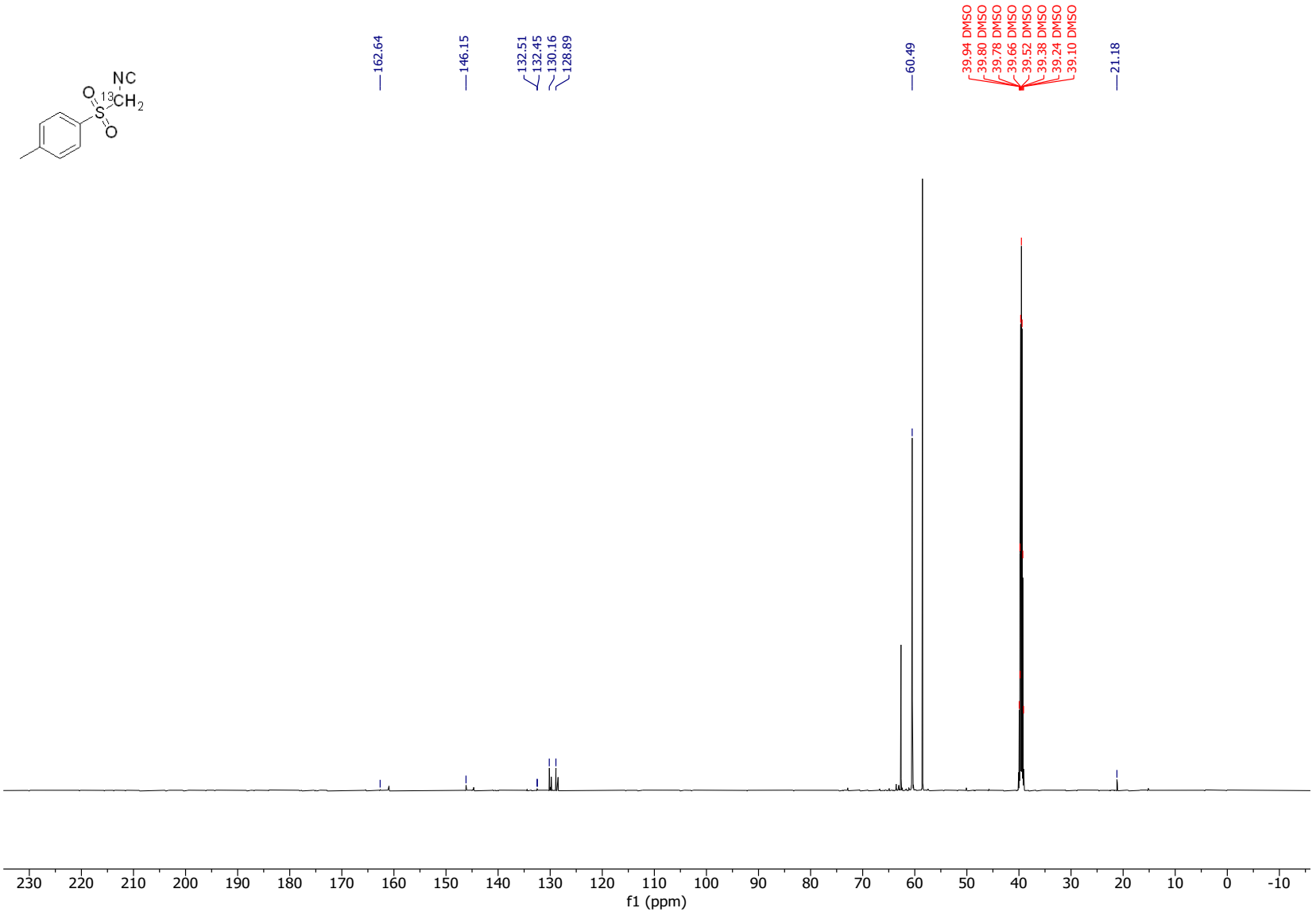
Spectrum S9: ^{1}H-NMR of N-(tosyl(^{13}C)methyl)-formamide



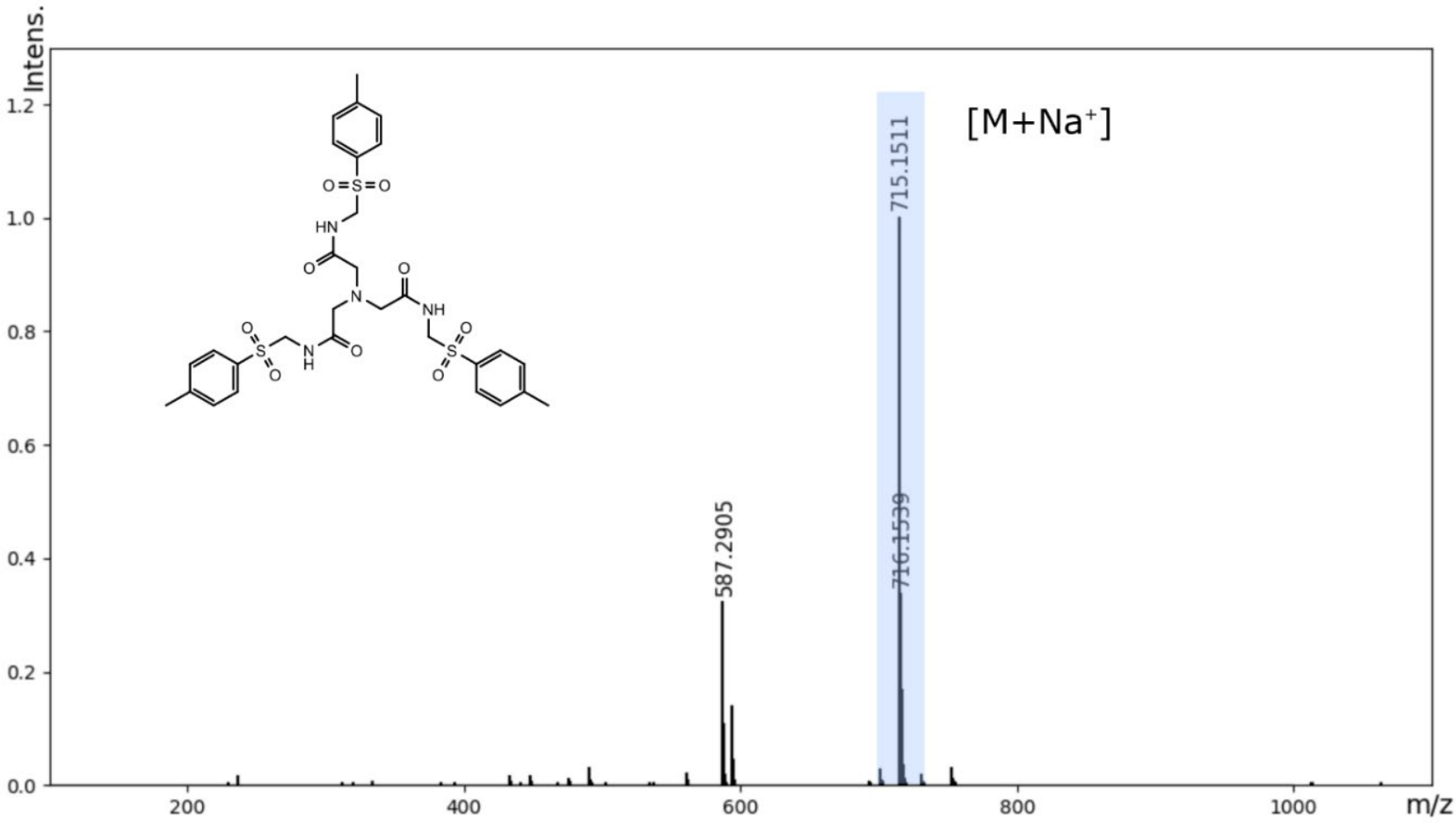
Spectrum S10: ^{13}C-NMR of N-(tosyl(^{13}C)methyl)-formamide



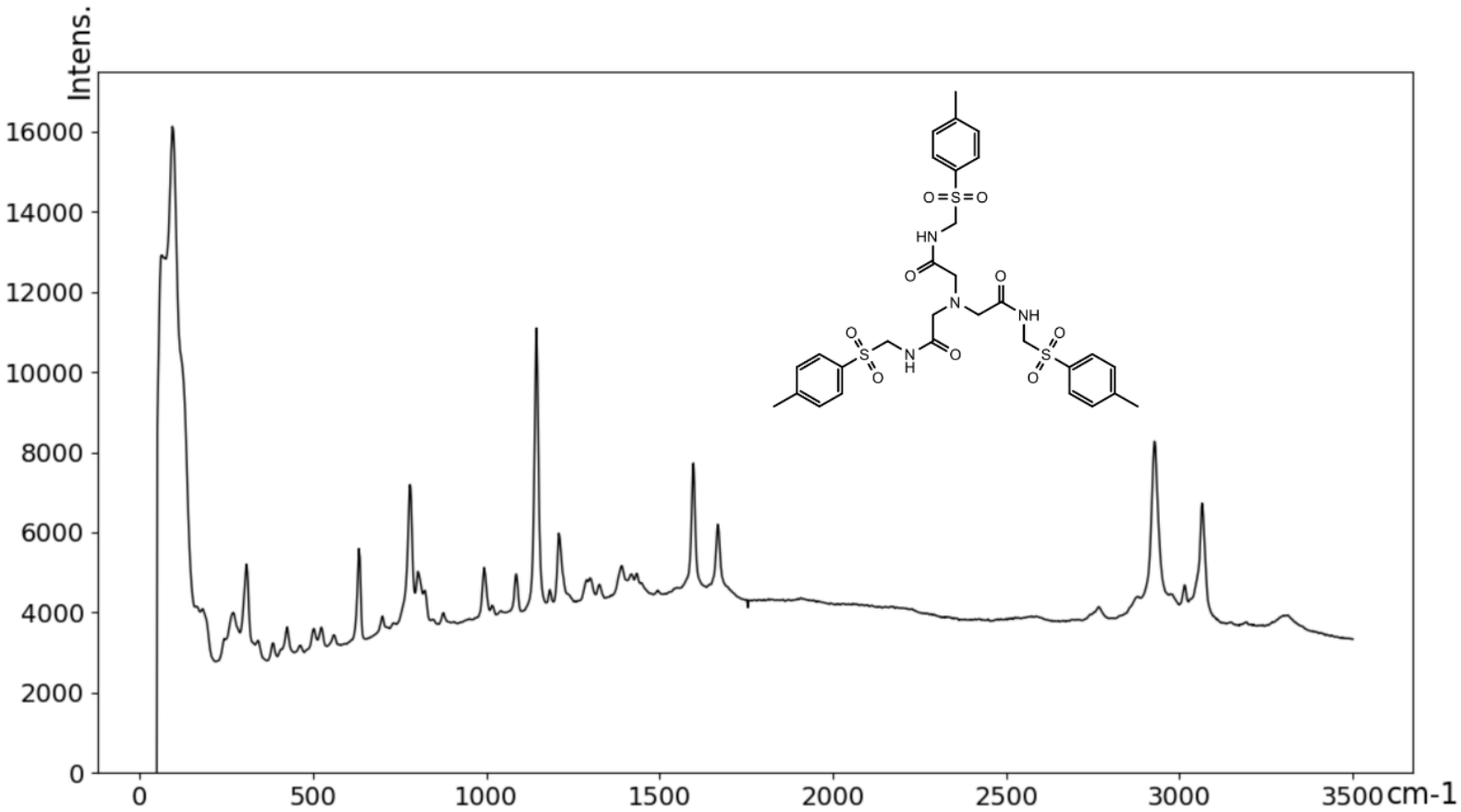
Spectrum S11: ^{1}H-NMR of Tos^{13}CH$_2$NC



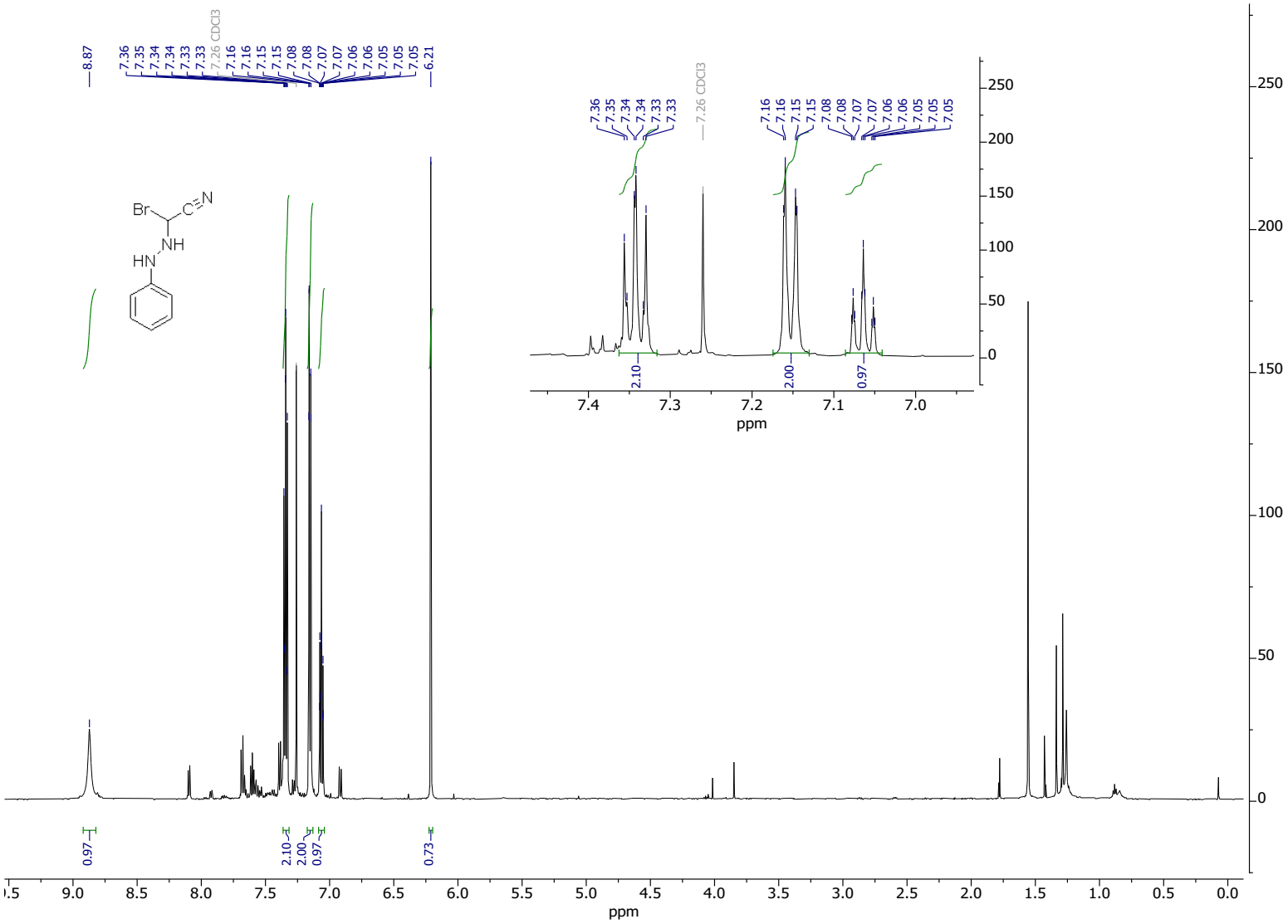
Spectrum S12: ^{13}C-NMR of Tos$^{13}CH_2NC$



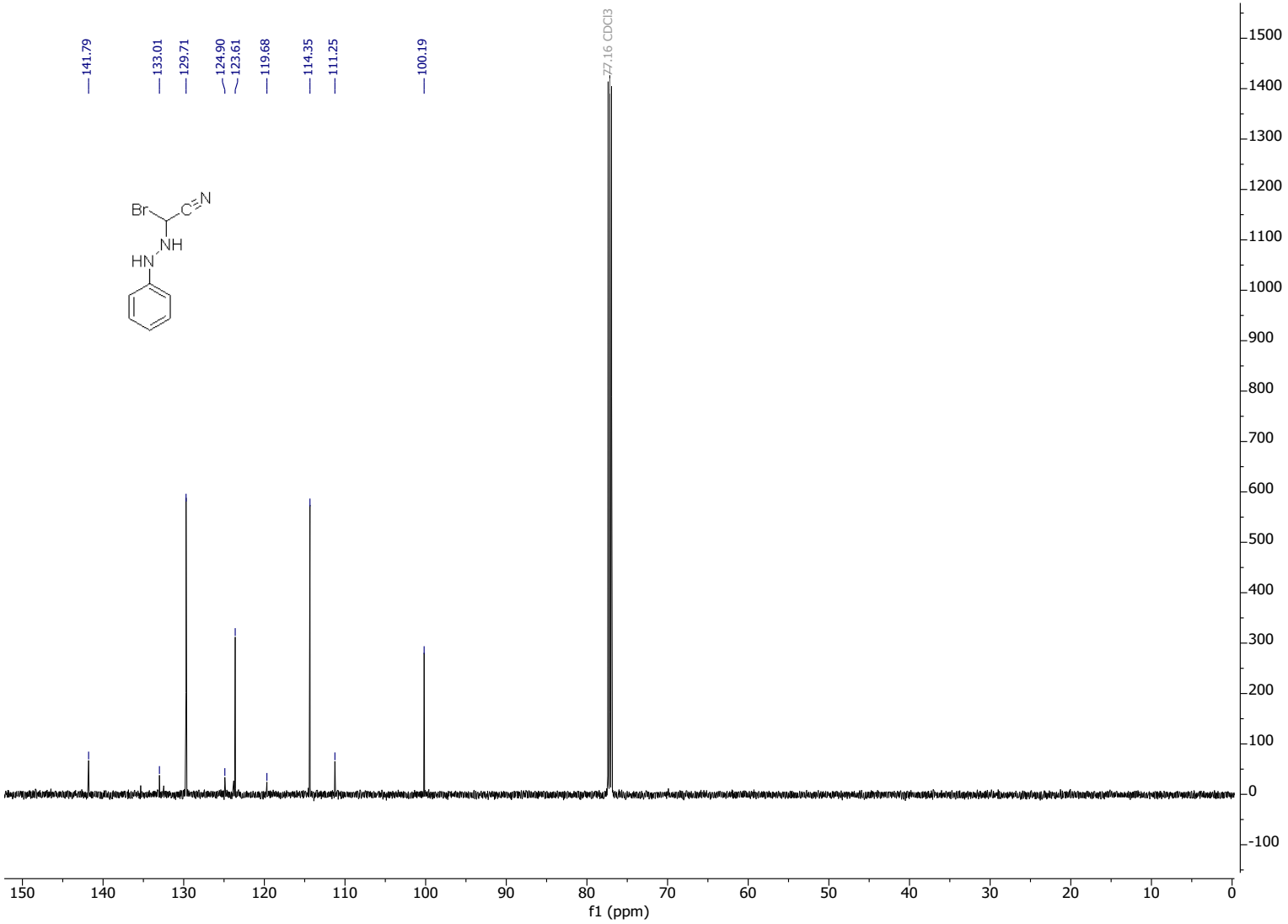
Spectrum S13: MS spectrum of compound **25**



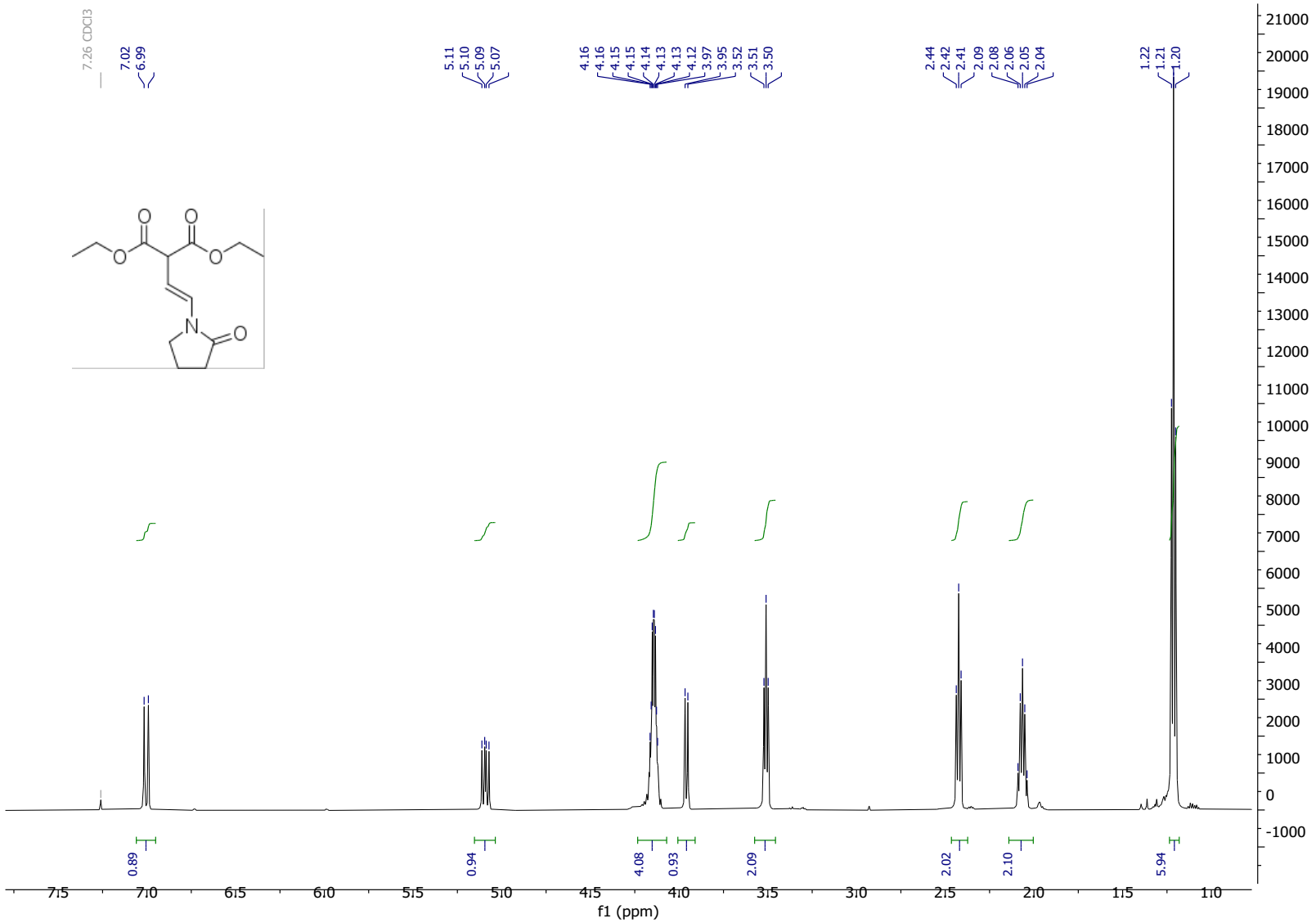
Spectrum S14: Raman spectrum of compound **25**



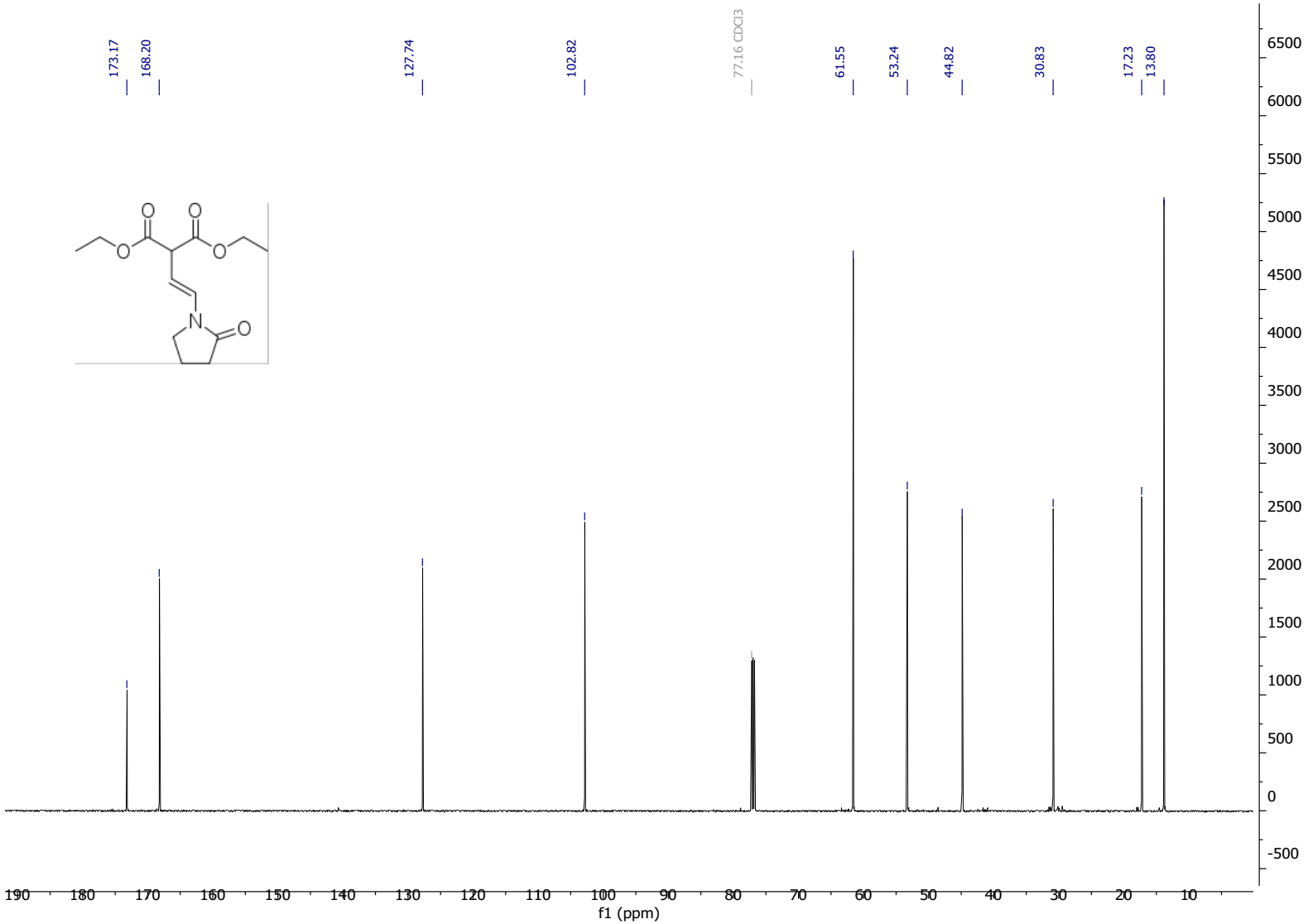
Spectrum S15: ^{1}H-NMR of compound **82**



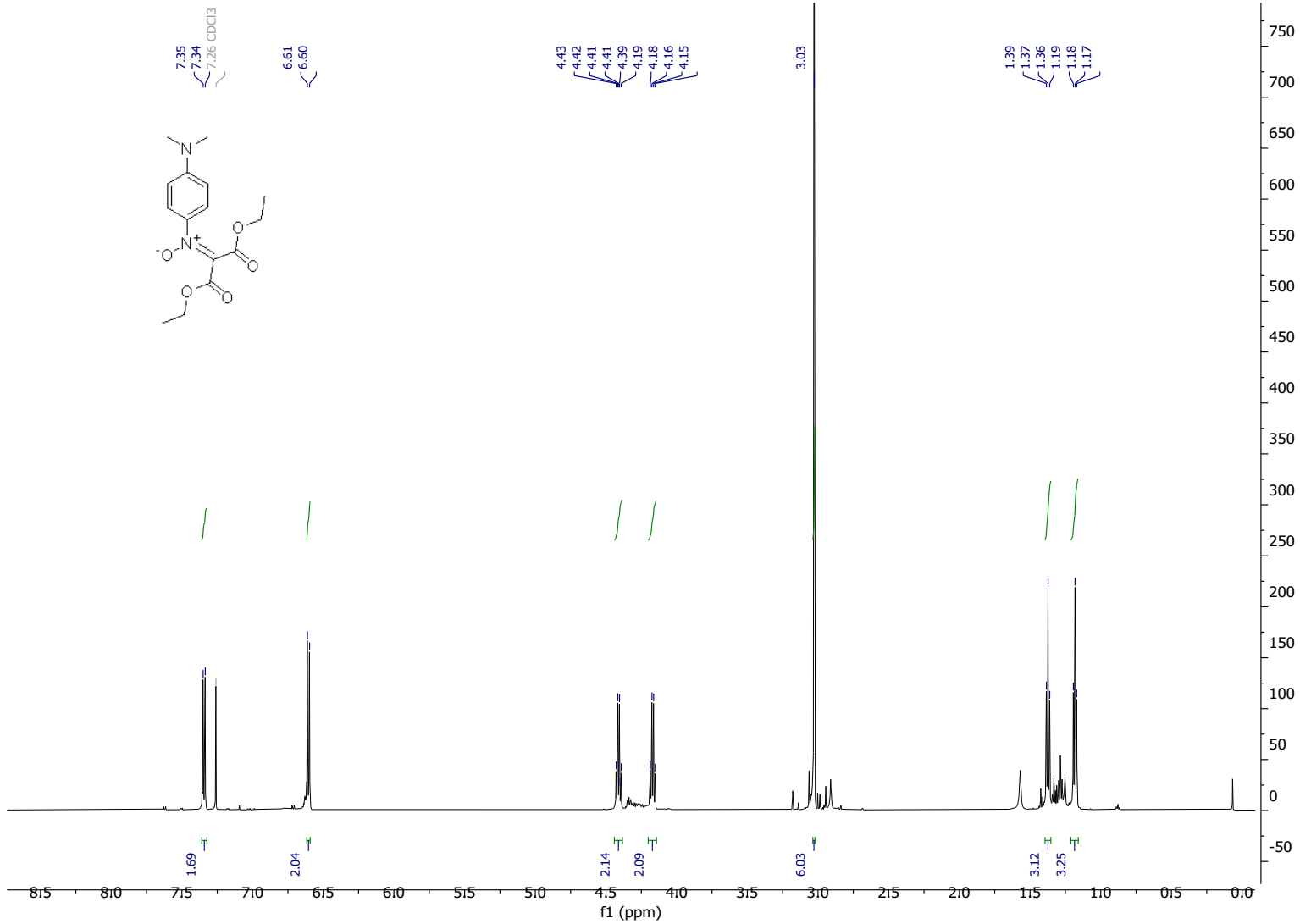
Spectrum S16: ^{13}C-NMR of compound **82**



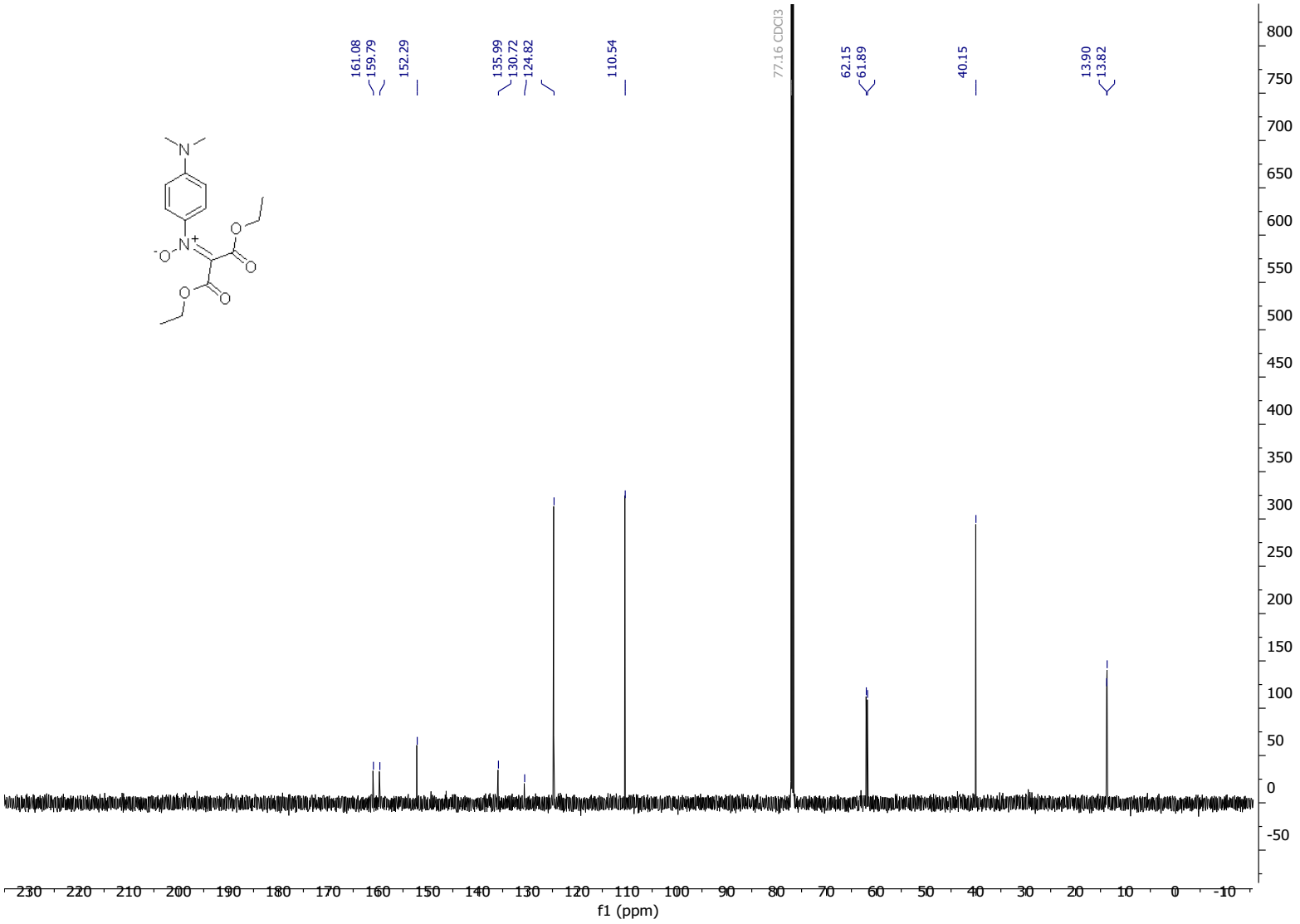
Spectrum S17: ^{1}H-NMR of compound **84**



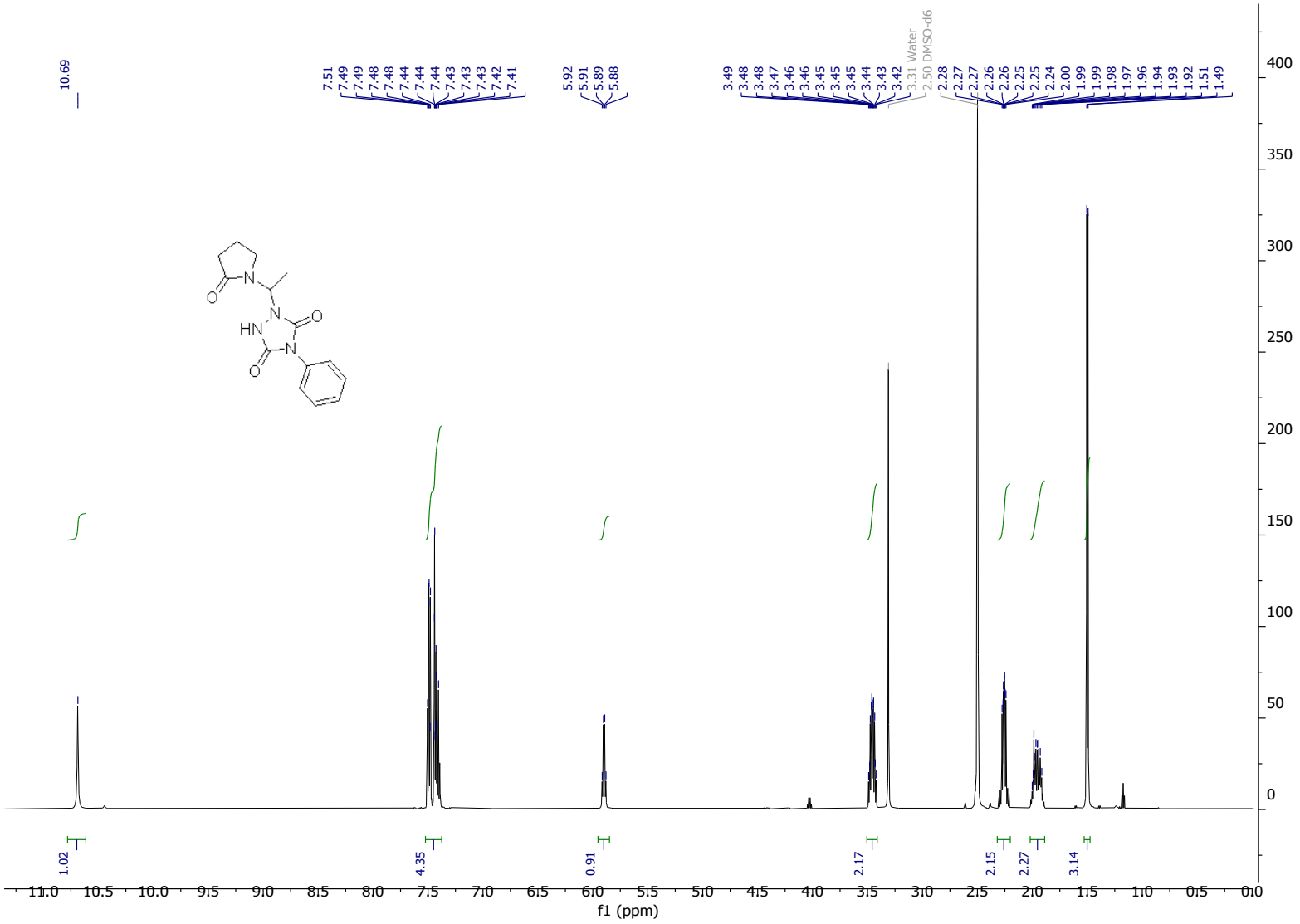
Spectrum S18: ^{13}C-NMR of compound **85**



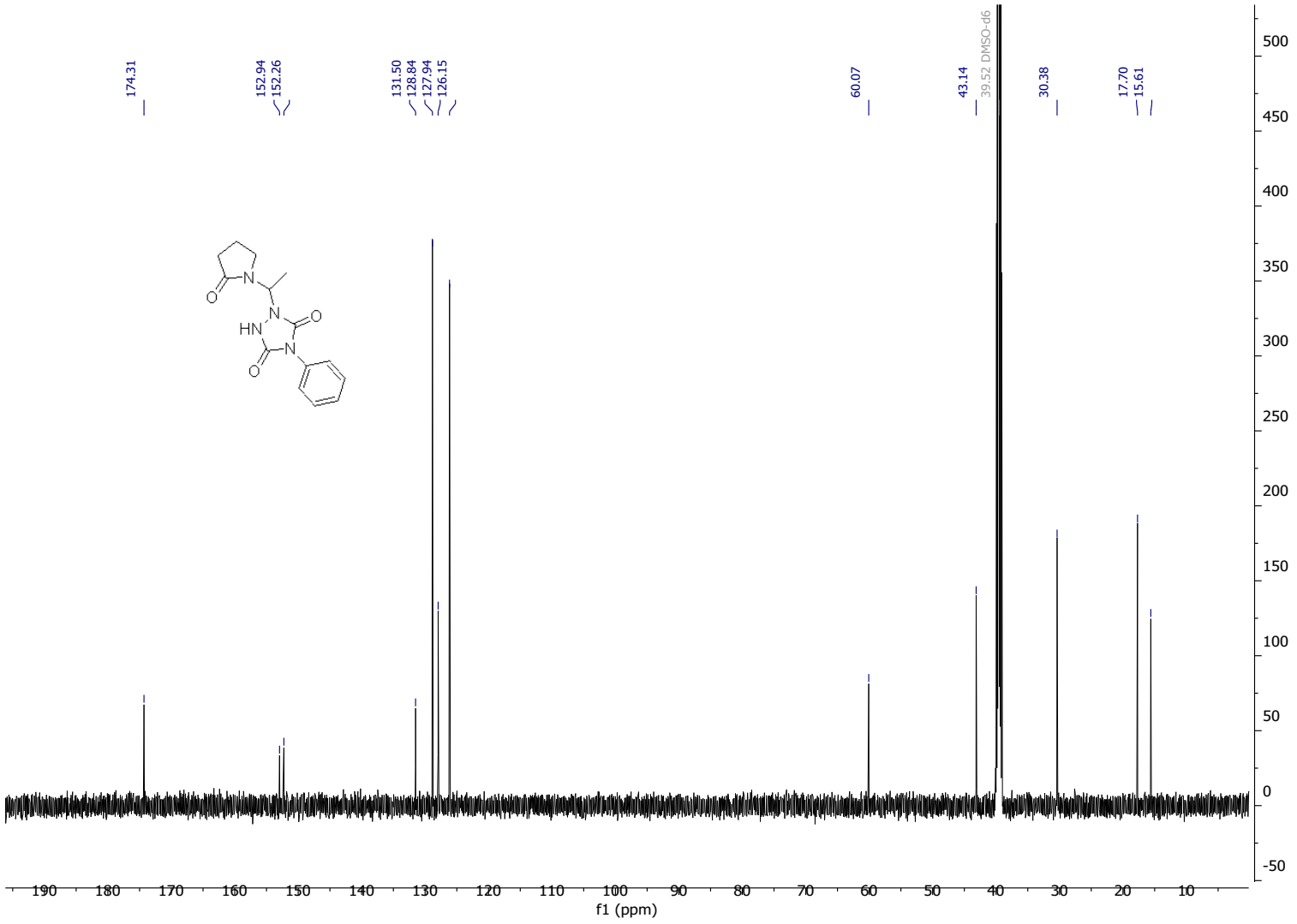
Spectrum S19: ^{1}H-NMR of compound **85**



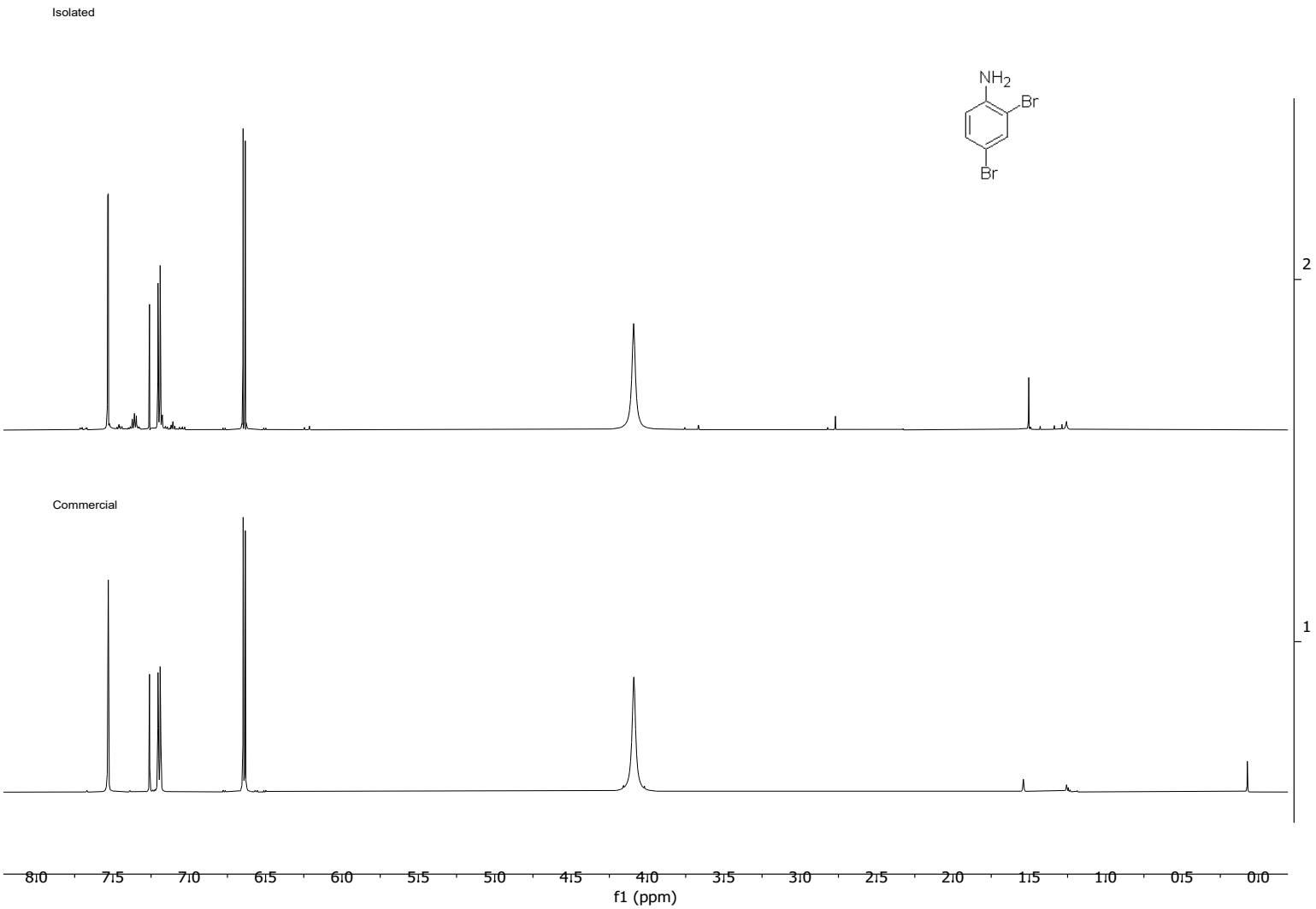
Spectrum S20: ^{13}C-NMR of compound **85**



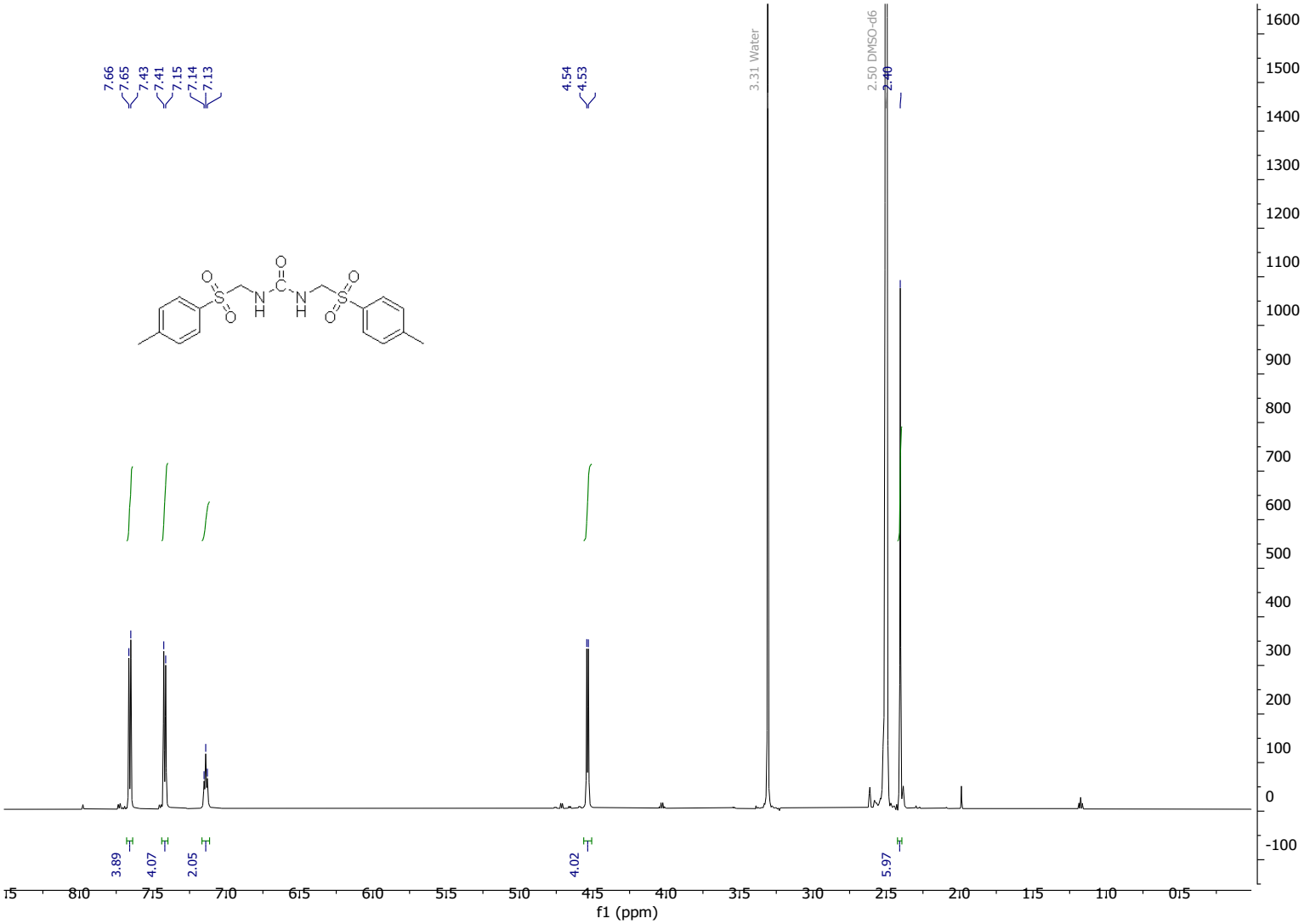
Spectrum S21: ^{1}H-NMR of compound **86**



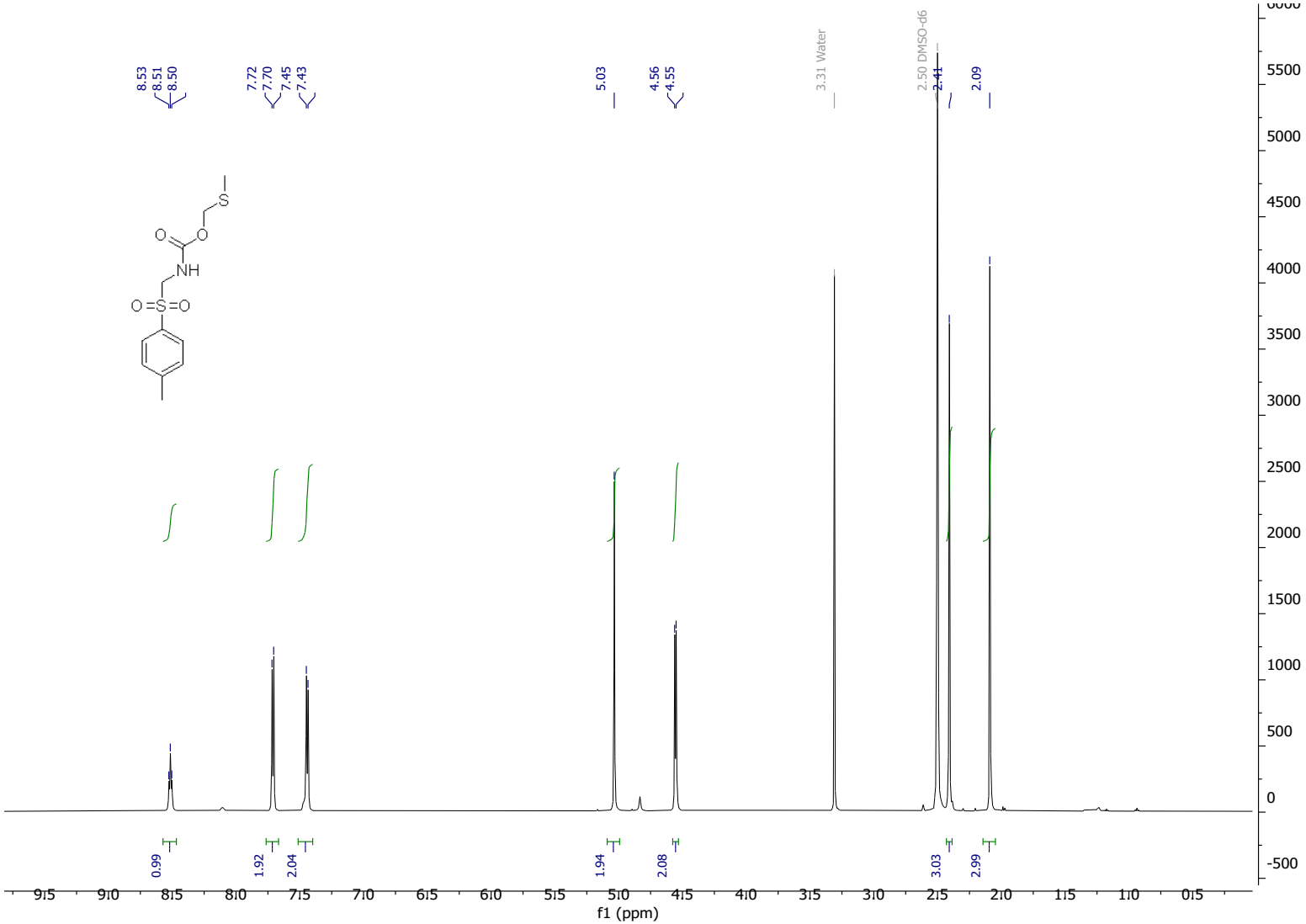
Spectrum S22: ^{13}C-NMR of compound **86**



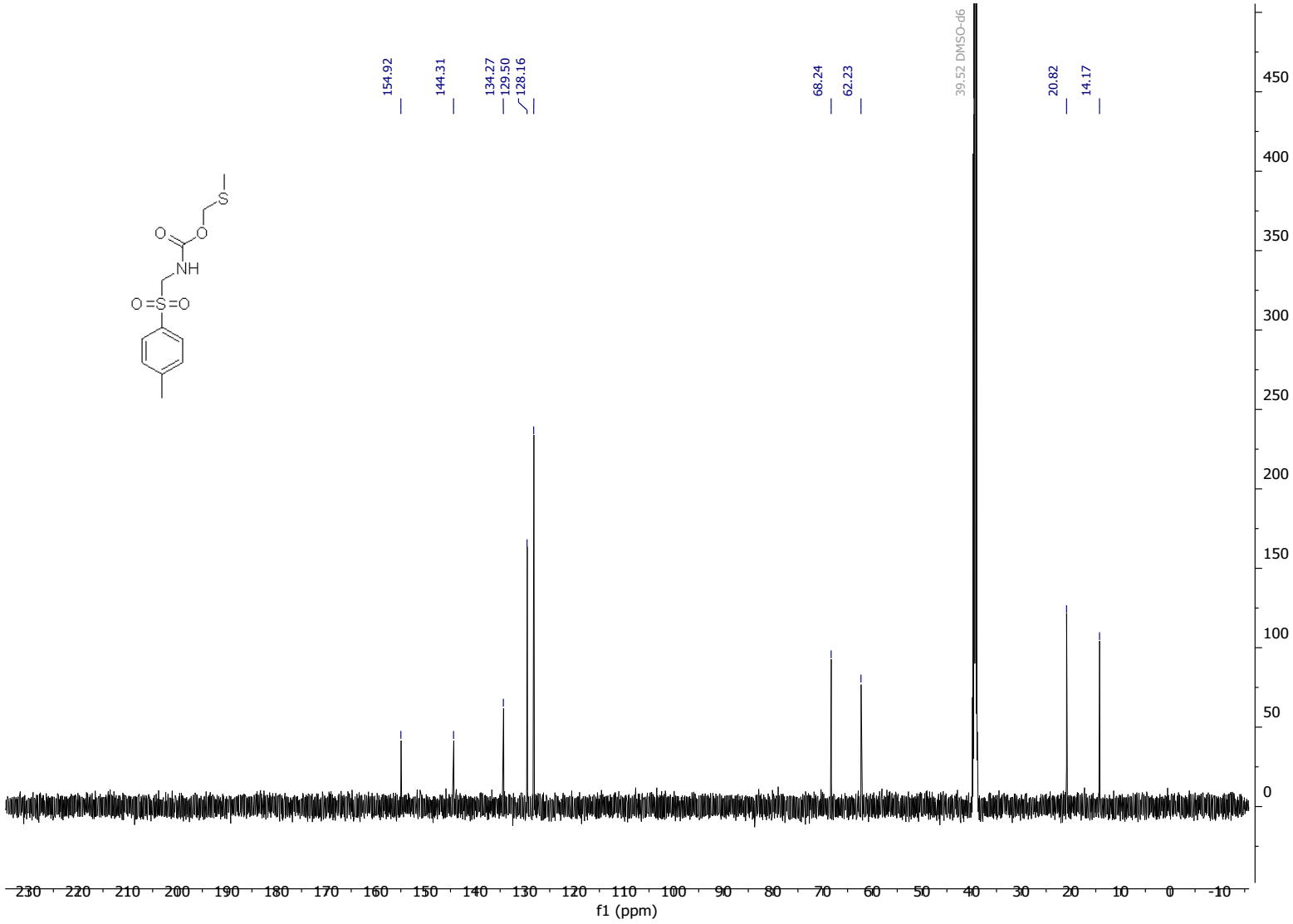
Spectrum S23: ^{1}H-NMR of compound **83,** compared with the commercially available 2,4-dibromoaniline



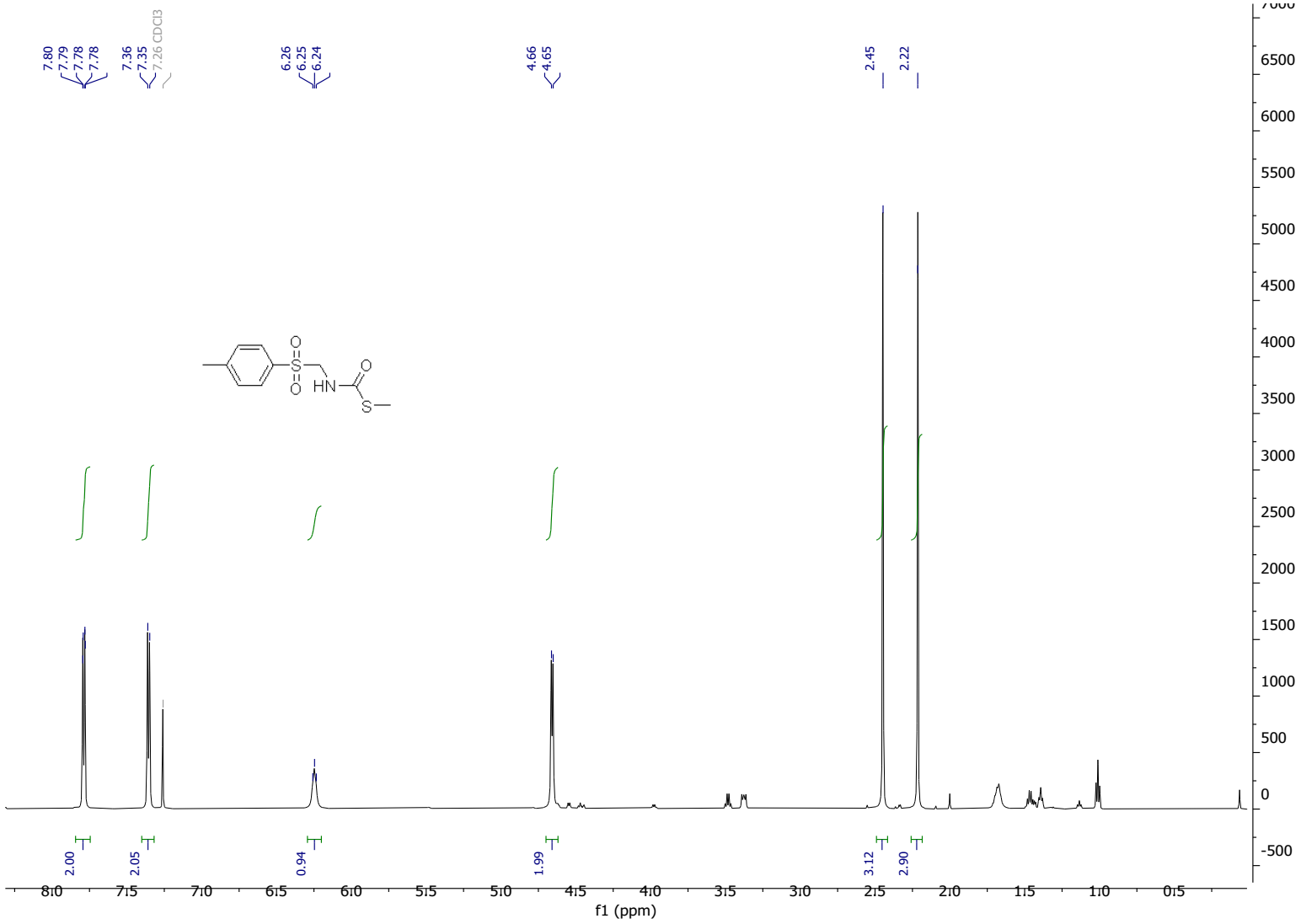
Spectrum S24: ^{1}H-NMR of compound **64**



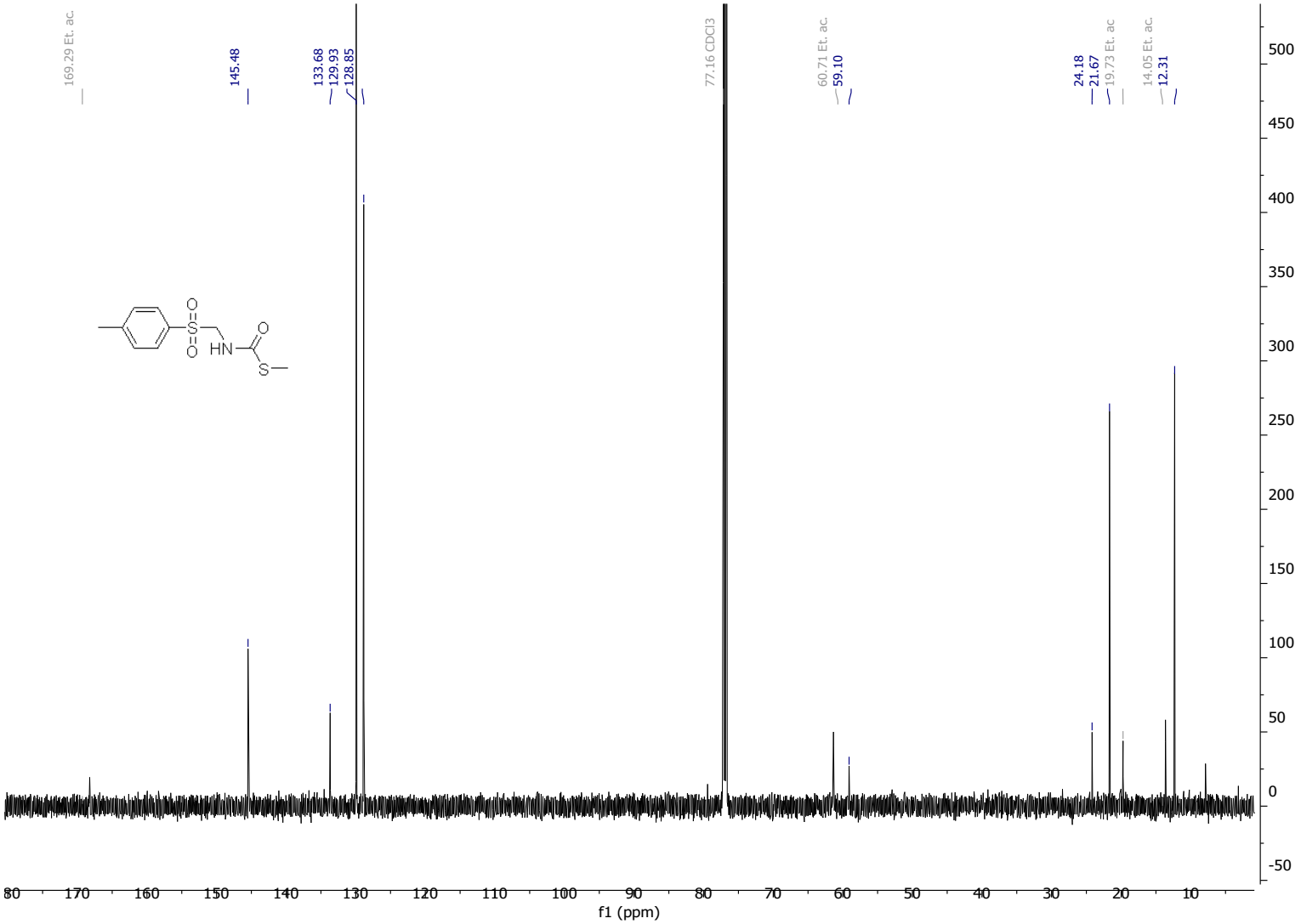
Spectrum S25: ^{1}H-NMR of compound **65**



Spectrum S26: ^{13}C-NMR of compound **65**



Spectrum S27: ^{1}H-NMR of compound **81**



Spectrum S28: ^{13}C-NMR of compound **8**

[32] D.P. Kingma, J. Ba, CoRR, abs/1412.6980, 2014.

[33] https://github.com/tensorflow/tensorflow (accessed 11/03/19).

[34] Sorenson, W. R. Reaction of an Isocyanate and a Carboxylic Acid in Dimethyl Sulfoxide. J. *Org. Chem*. **24**, 978–980 doi: 10.1021/jo01089a024 (1959).

[35] Chen, Q., Huggins, M. T., Lightner, D. A., Norona, W. & McDonagh, A. F. Synthesis of a 10-oxo-bilirubin: Effects of the oxo group on conformation, transhepatic transport, and glucuronidation. *J. Am. Chem. Soc.* **121**, 9253–9264 doi: 10.1021/ja991814m (1999)

[36] B. E. Hoogenboom; O. H. Oldenziel; Leusen, A. M. v., p-TOLYLSULFONYLMETHYL ISOCYANIDE. *Org. Synth.* **57**, 102 doi: 10.15227/orgsyn.057.0102 (1977).

[37] Cappon, J. J. *et al.* Synthesis of L-histidine specifically labelled with stable isotopes. *Recl. des Trav. Chim. des Pays-Bas* **113**, 318–328 doi: 10.1002/recl.19941130603 (1994).

[38] Chen, C. Y., Bocian, D. F. & Lindsey, J. S. Synthesis of 24 bacteriochlorin isotopologues, each containing a symmetrical pair of 13C or 15N atoms in the inner core of the macrocycle. *J. Org. Chem.* **79**, 1001–1016 doi: 10.1021/jo402488n (2014).

[39] Gossauer, A.; Suhl, K., Totalsynthese des Verrucarins E sowie ihre Anwendung zur Herstellung eines ^{13}C-markierten Derivates desselben. *Helvetica Chimica Acta***, 59** (5), 1698-1704 doi: 10.1002/hlca.19760590530 (1976)

[40] Fenselau, A. H.; Moffatt, J. G., Sulfoxide-Carbodiimide Reactions. III.1Mechanism of the Oxidation Reaction. *J. Am. Chem. Soc.* **88**, 1762-1765 doi: 10.1021/ja00960a033 (1966)

[41] Le, H. V. & Ganem, B. Trifluoroacetic anhydride-catalyzed oxidation of isonitriles by DMSO: A rapid, convenient synthesis of isocyanates. Org. Lett. 13, 2584–2585 doi: 10.1021/ol200695y (2011)

[42] Kemp, W., Nuclear Magnetic Resonance in Chemistry: A Multinuclear Introduction, 75 (Macmillan education ltd, 1986)

[43][43] Sattar, M., Rathore, V., Prasad, C. D. & Kumar, S. Transition-metal-free Chemoselective Oxidative C−C Coupling of the sp3 C−H Bond of Oxindoles with Arenes and Addition to Alkene: Synthesis of 3-Aryl Oxindoles, and Benzofuro- and Indoloindoles. *Chem. - An Asian J.* **12**, 734–743 doi:10.1002/asia.201601647 (2017) (supporting information)

[44] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 1–13 doi: 10.1186/s13321-015-0069-3 (2015).

[45] Jiang, H. *et al.* Direct C-H functionalization of enamides and enecarbamates by using visible-light photoredox catalysis. *Chem. - A Eur. J.* **18**, 15158–15166 doi: 10.1002/chem.201201716 (2012).

[46] Tomioka, Y., Nagahiro, C., Nomura, Y. & Maruoka, H. Synthesis and 1,3-Dipolar Cycloaddition Reactions of N-Aryl-C,C-dimethoxycarbonylnitrones. *J. Heterocyclic Chem.* **40**, 121 doi: 10.1002/chin.200326130 (2003).

[47] Senogles, E. & Thomas, R. A. The kinetics and mechanism of the acid-catalysed hydrolysis of N-vinyl-pyrrolidin-2-one. *J. Chem. Soc. Perkin Trans.* ***2*** 825–828 doi:10.1039/P29800000825 (1980)

[48] Sheldrick, G. M. SHELXT - Integrated space-group and crystal-structure determination. *Acta Crystallogr. Sect. A Found. Crystallogr.* **71**, 3–8 doi: doi: 10.1107/S2053273314026370 (2015).

[49] Sheldrick, G. M. Crystal structure refinement with SHELXL. *Acta Crystallogr. Sect. C Struct. Chem.* **71**, 3–8 doi: 10.1107/S2053229614024218 (2015).

[50] Farrugia, L. J. WinGX suite for small-molecule single-crystal crystallography. *J. Appl. Crystallogr.* **32**, 837–838 doi: 10.1107/S0021889899006020 (1999).

[51] Blessing, R. H. An empirical correction for absorption anisotropy *Acta Crystallogr.* **51**, 33-8 doi: 10.1107/s0108767394005726 (1995)