Supporting Information: Computation-Guided Exploration of The Reaction Parameter Space of N,N-Dimethylformamide Hydrolysis

Ignas Pakamorė[†], Ross S. Forgan[†]

[†]*WestCHEM School of Chemistry, University of Glasgow, Joseph Black Building, University Avenue, Glasgow G12 8QQ, UK*

Table of Contents

S1. Optimisation Algorithm Assessment

The search algorithm's performance underwent testing across three distinct functions: Ackley, Mataya's, and Himmelblau's (Figure 3 in the main manuscript). Ackley and Matyas's functions feature a singular minimum (as described in Equations S1 and S2), whereas Himmelblau's function encompasses four distinct minima points (as outlined in Equation S3). Notably, the Ackley function comprises a pronounced global minimum along with various local minima, offering a robust evaluation of the impact of the step size χ on convergence towards the minima points. The choice of the Himmelblau function serves to investigate convergence behaviour in scenarios where multiple global minima coexist.

$$f(x) = -a\exp\left(-b\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}\right) - \exp\left(\frac{1}{N}\sum_{i=1}^{N}\cos\left(c\,x_i\right)\right) + a + \exp(1) \#(S1)$$

Search boundaries: $-32 \le x_i \le 32$

Minimum at $f(0,0) = 0$

$$f(x,y) = 0.26\left(x^2 + y^2\right) - 0.48xy \#(S2)$$

Search boundaries: $-10 \le x_i \le 10$

Minimum at $f(0,0) = 0$

$$f(x,y) = \left(x^2 + y - 11\right)^2 + \left(x + y^2 - 7\right)^2 \#(S3)$$

Search boundaries: $-6 \le x_i \le 6$

Minimum at $f(x,y) = \begin{cases} f(3,2) = 0 \\ f(-2.805, 3.131) = 0 \\ f(-3.779, -3.283) = 0 \\ f(3.584, -1.848) = 0 \end{cases}$

S2. ChemSPX Package

The ChemSPX code introduces three distinct methods for chemical reaction parameter space exploration: *full space*, *restricted space*, and *sub-space* (Figure S1). Initially, a comprehensive exploration can be conducted within the specified boundaries. In cases where the region of interest is confined within the parameter space, the boundaries can be narrowed accordingly. Additionally, the algorithm facilitates exploration within a subspace surrounding a specific data point. In this scenario, samples are drawn within the proximity of a hyperrectangle, defined by the χ parameter, around the chosen data point (Figure S1).

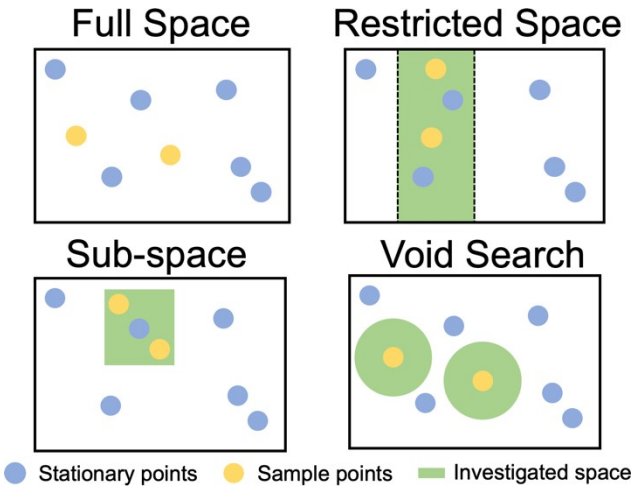

Figure S1. The four main parameter space exploration methods implemented in ChemSPX: full space, restricted space, sub-space, and void search.

S3. ChemSPX Code Default Parameters

Genetic algorithm (https://github.com/PasaOpasen/geneticalgorithm2):

- Mutation probability: *0.1*
- Crossover probability: *0.5*
- Parent population: *0.3*
- Elite ratio: *0.01*
- Maximum number of iterations without improvement: *50*
- Crossover type: *uniform*

S4. ChemSPX – Initial Reaction Parameter Sampling Methods

In the ChemSPX code, initial (pre-optimized) parameter vector sampling is accomplished through three methods: Latin hypercube, equilibrated Latin hypercube, and the void search algorithm (discussed in the manuscript). The Latin hypercube sampling, rooted in the Latin square design, ensures a more uniform sampling across n-dimensional space compared to random approaches. This method divides input distributions into N intervals with equal probabilities, randomly drawing a single sample from each interval, resulting in a Latin hypercube (Figure S2b). This approach contrasts with the less uniform sampling of the random (brute-force) method (Figure S2a).

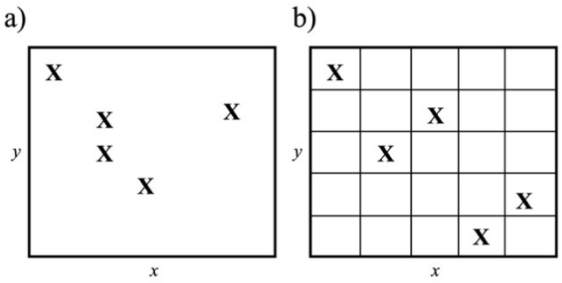

Figure S2. a) Brute force (random) sampling and **b)** Latin hypercube sampling sample distribution in 2-dimensional space.

The equilibrated Latin hypercube algorithm (LHSEQ) implemented in ChemSPX employs the inverse distance function ϕ (equation 3 in the main manuscript) to increase the spatial distance between data points. In other words, the algorithm aims to further improve the sampling obtained from LHS. Equilibration is achieved through Genetic Algorithm (GA), along with customizable parameters for the inverse distance equation. This approach results in a grid-like arrangement of data points in n-dimensional space, with spacing controlled by the ϕ function.

To illustrate LHSEQ performance, a 2-dimensional Latin hypercube with 10 sample points was generated. Equilibration was performed using the GA algorithm with 120 steps, χ value set to 0.1, and default settings for other GA parameters. The inverse distance function exponent b was set to 1, and the number of nearest neighbours varied from 1 to 10 (all).

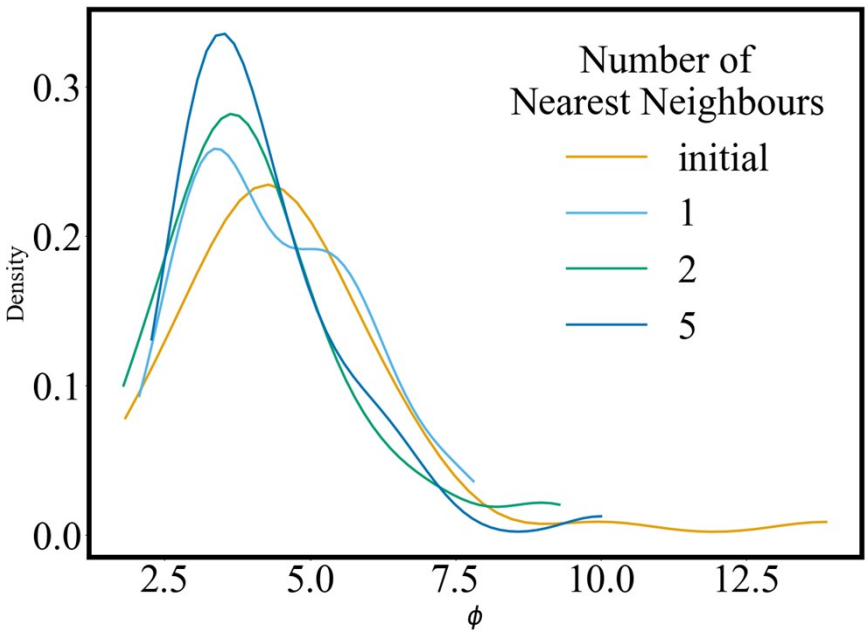

Figure S3. The underlying density distribution of the inverse distance function ϕ of three datasets obtained using different nearest neighbour number settings. The 'initial' dataset represents pre-optimised vectors obtained using LHS sampling, where nearest neighbours are set to 5.

Following Figure S3, it is observed that the number of points in distant proximity increases as the number of nearest neighbours rises. This is expressed by the increase in density distribution of the inverse distance function ϕ at low values. Compared to the initial LHS sampling (Figure S3 'initial') the obtained datasets have a much more improved ϕ distribution

Both Latin hypercube (LHS) and equilibrated-Latin hypercube (LHSEQ) methods allow efficient sampling of data points in n-dimensional space. Compared to the brute-force approach, the LHS method draws data points with larger spatial distances and reduces the probability of cluster formation. In addition, the LHSEQ approach proves to be even more efficient in drawing samples with a grid-like arrangement. Nevertheless, in certain cases, data sets can contain large sparse regions that need a targeted methodology to draw samples within those parts of the defined space.

S5. Parameter Space Exploration Ratio

The fraction of explored (or occupied) parameter space is approximated using the Monte Carlo (MC) integration algorithm[1]. The computation involves sampling of N randomly generated points x ($x \in R^n$) within a defined simulation box (Figure S4). The simulation box dimensions are equal to the defined boundaries of the parameter space.

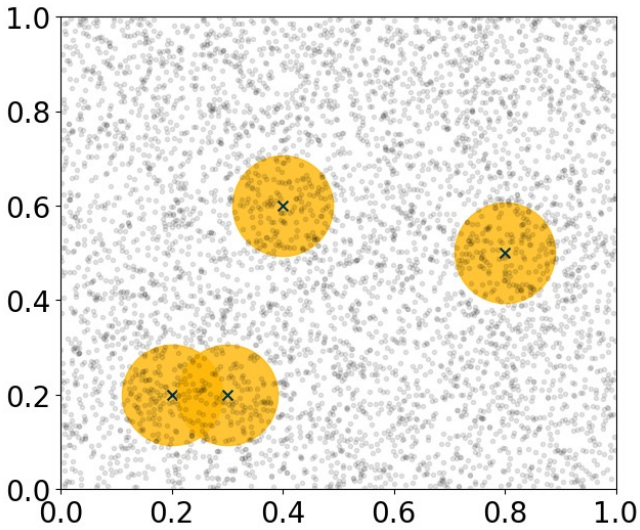

Figure S4. Example of a 2-dimension MC simulation box. Herein, reaction vectors are denoted by the symbol x, and the corresponding n-spheres (in this case circle) area in yellow. The grey points within the box or circles represent MC samples.

The reaction vector data points within this simulation box are treated as n-dimensional hyperspheres (S^n) with radius r, which determines the hypersphere size and is defined by equation S4. Here the z_i denotes the hypersphere centre (reaction vector) and a_i the coordinates of the point on the sphere (parameter limit of change).

$$r = \sqrt{\sum_{i=1}^{i+1} (a_i - z_i)^2} \#(S4)$$

Radius value is manually calculated based on the limit of change, which is the smallest parameter value variation that does not significantly affect the reaction outcome. In this case,

[1] This algorithm is integrated into the ChemSPX package ParSpaceCoverage.py module.

the limits were set to: 0.1 (H_2O [mL]), 0.0001 (Acid [mol]), 0.1 (Acid pK_a), 10 (Temperature [°C]), 0.5 (Time [h]), yielding radius:

$$r = \sqrt{0.1^2 + 0.0001^2 + 0.1^2 + 10^2 + 0.5^2} \approx 10.$$

In this calculation, DMF volume (8 mL) is excluded as it is kept constant. The ratio of exploration is determined by dividing the number of MC points falling within the hyperspheres by the total number of MC points (equation S5). Herein, the total number of MC iterations was set to $1 \cdot 10^6$.

$$Fraction\ of\ explored\ parameter\ space = \frac{\sum x \in S^n}{N} \#(S5)$$

S6. Experimental

Chemicals

All chemicals were purchased from Alfa Aesar, Acros Organics, Fisher Scientific, Sigma-Aldrich, Honeywell, and Cambridge Isotope Laboratories and used without further purification. All organic acid catalysts used in the experiments were concentrated and have not been further diluted. In the case of mineral acids, 37 % HCl, 75 % HNO_3 and conc. H_2SO_4 were employed.

Initial experiments

The initial experiments were carried out in 25 mL hydrothermal synthesis reactors at 100 and 120 ^oC for 24, 38, 27 and 168 h. The DMF volume was kept constant (8 mL) while the HCl amount was varied from 0.1 to 1 mL (0.0012 - 0.012 mmol).

Parameter Space Sampling Reactions

All reactions were carried out in two configurations: 25 mL hydrothermal synthesis reactors that were used for higher temperature range and 20 mL scintillation vials used for lower temperature range reactions. High-temperature range reactions were carried out in convectional ovens at 100, 120 and 150 $^\circ C$ with low-temperature range reactions done using a hotplate set up at 40, 60 and 80 ^oC. Room temperature (25 ^oC) experiments were conducted using scintillation vials with no temperature control. The reagents were measured using a measuring cylinder for DMF and micropipettes for other reagents (except for chloroacetic acid crystals of which accurate mass was weighed). All formulations can be found summarised in the attached .csv file.

S7. NMR Sample Analysis

All NMR spectra were measured on a Bruker AVIII 400 MHz spectrometer at room temperature (University of Glasgow). For the quantification of formic acid in the analyte, samples were dissolved in deuterated chloroform that contained a known amount (% *w/w*) of tetramethylsilane (TMS) standard. The formic acid concentration calculations were carried out using accurate masses of sample and TMS, using the equation S6:

$$FA(\%w/w) = \frac{n(TMS) \cdot I(FA) \cdot M_r(FA)}{12 \cdot m(sample)} \cdot 100.\#(S6)$$

Masses were acquired using four-figure balances. Measurements were done in triplicate to obtain standard deviation. The NMR detection limit in this work was found to be 0.00542 ±0.0036% (*w/w*).

S8. Initial DMF Hydrolysis Experiments

NMR spectroscopy was a primary choice for quantitative and qualitative analysis of the samples as investigated species are poor chromophores, hence other available spectroscopy methods could not be used. Both ^{1}H and ^{13}C NMR analysis methods afford the detection of formic acid in the reaction mixture (Figures S5 a) and b)).

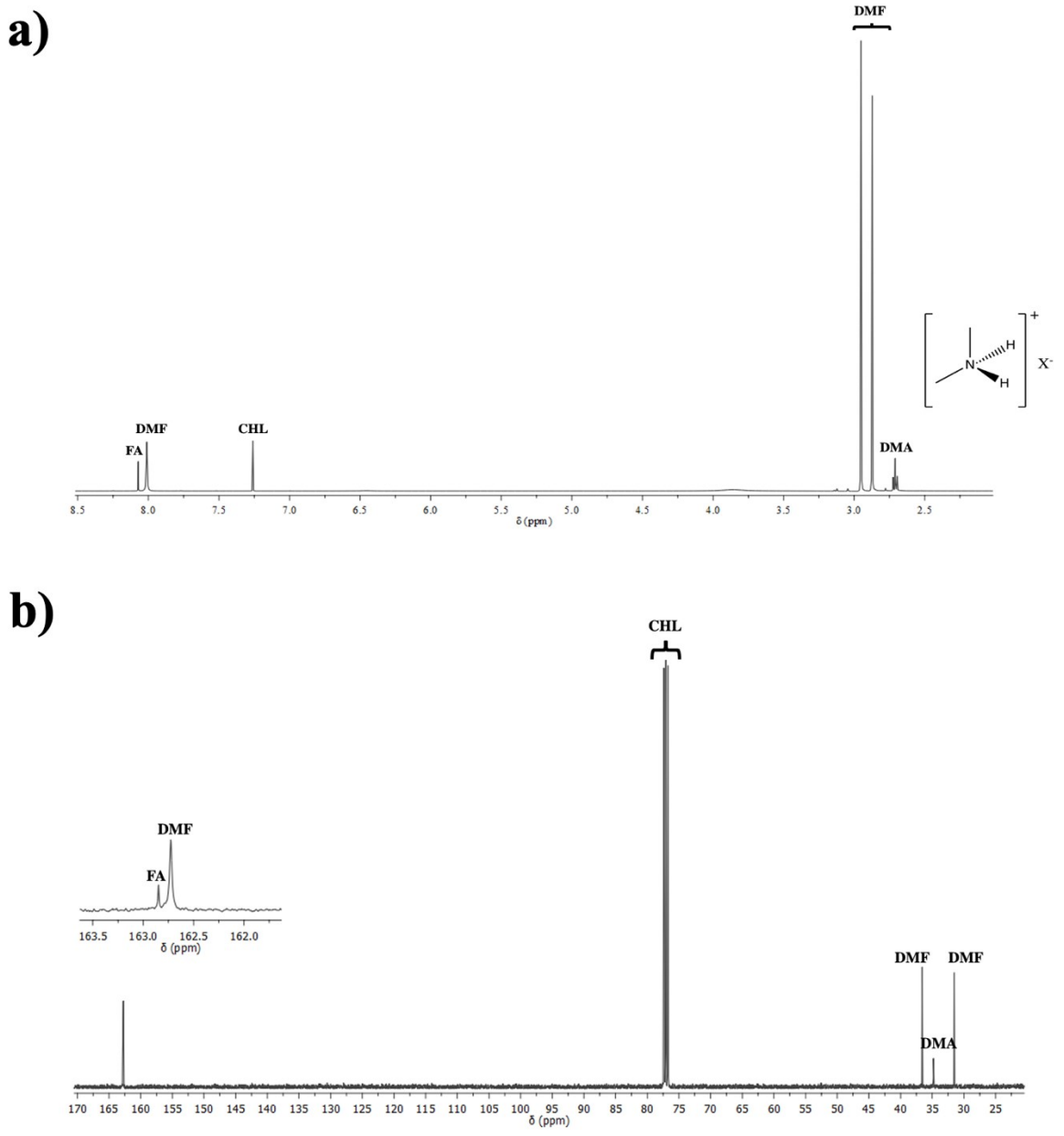

Figure S5. a) ^{1}H and **b)** ^{13}C NMR spectrograms of DMF hydrolysis sample containing 0.012 mol of HCl after 72 h. Peaks in the spectrograms are labelled by compounds: CHL-*chloroform*; DMF-*N,N-dimethylformamide*; FA-*formic acid*; DMA-*dimethylamine.*

The investigation focused on varying the amount of HCl, ranging from 0.0012 mol (0.1 mL) to 0.012 mol (1 mL), with respect to 0.1038 mol (8 mL) of DMF. Reagent-grade DMF (≥ 99%) was utilized to emulate common MOF synthesis conditions. The concentrated HCl acid used in this study is 37%, indicating that approximately 63% of the solution comprises water, facilitating the hydrolysis reaction; therefore, no additional water was introduced. Furthermore, experiments were carried out at typical MOF crystallization temperatures: 100 and 120 °C. The obtained results revealed the expected concentration and time dependence of the formation of formic acid (%FA) during the hydrolysis reaction. The augmentation in the moles of HCl led to a substantial increase in %FA, ranging from approximately 0.01 to 0.5% (*w/w*) (equivalent to 0.002 - 0.1 mmol) per gram of the sample (Figure S6). In both temperature experiments, there is a nearly linear relationship between the investigated acid amounts and %FA. Notably, there was no significant difference observed between the two temperatures, indicating comparable outcomes.

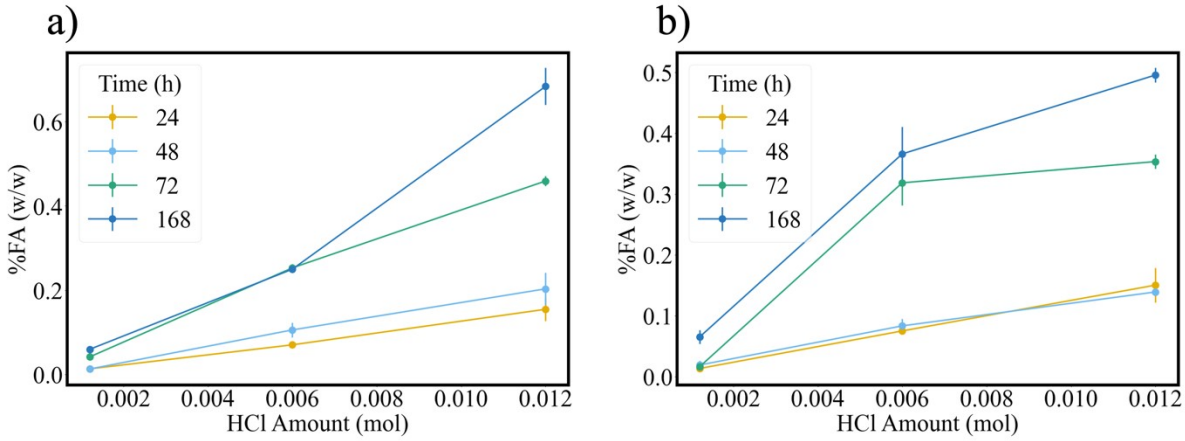

Figure S6. The relationship between the amount of formic acid generated and HCl amount across different time intervals. Graphs **a)** and **b)** show the results at 100 and 120 °C. Error bars represent the standard deviation of 3 repeat reactions.

The same experimental dataset was employed to evaluate the time dependence in both temperature settings, as illustrated in Figure S7. Consistent with the literature-reported data discussed, the time plots exhibit a sigmoid-shaped distribution, signifying that the %FA production rate decreases with increasing time.[1] In both instances, there is a notable sharp increase in the %FA value after 72 hours, followed by a reduced rate at higher time intervals. This suggests that beyond the 168-hour period, *i.e.*, one week, no significant increase in %FA can be anticipated.

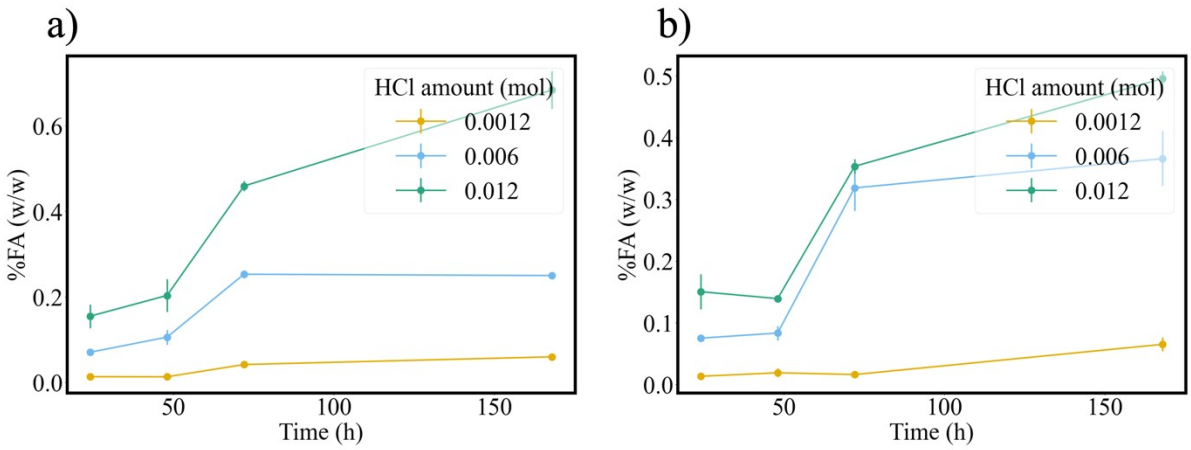

Figure S7. The relationship between the amount of formic acid generated and time for reactions containing distinct amounts of HCl catalyst. Graphs **a)** and **b)** show the results at 100 and 120 °C. Error bars represent the standard deviation of 3 repeated reactions.

The errors in NMR quantification of formic acid stem from analyte sample preparation and measurement. First of all, one possible source of error is the inaccurately obtained weight of the analyte and standard spiked deuterated solvent. More importantly, the accurate NMR analysis of samples with high acid concentration is hindered by FA and DMF peak overlap (Figure S8).

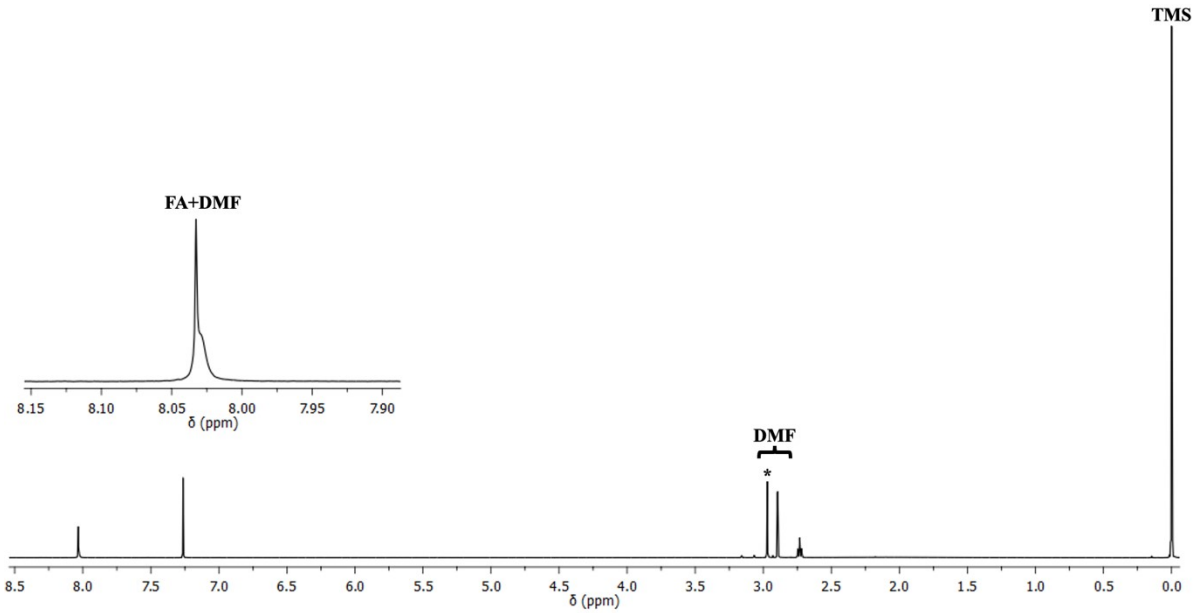

Figure S8. ^{1}H NMR spectrum of a sample with overlapping peaks. The * symbol denotes the methyl group in the DMF molecule.

For the quantitative analysis of samples, this problem has been solved by using the equation S7:

$$I(FA) = I(FA + DMF) - \frac{I(DMF*)}{3} \#(S7)$$

In this case, the integral of one methyl unit DMF (DMF*) is obtained and then subtracted from the merged peak of FA+DMF. The obtained integral I(FA) is further used in equation S6 to calculate %FA (see SI7).

S9. Correlation Matrix and Pair Correlation Plots

Pearson's correlation matrix (Figure S9) aligns with the trends discussed in the manuscript data analysis section. Notably, water and acid emerge as pivotal factors in DMF hydrolysis, showcasing notable correlation coefficients in the dataset, both falling within the range of approximately 0.4. The acid pK_a R coefficient exhibits a negative correlation, indicating an inverse relationship between pK_a and %FA. Conversely, temperature and time reveal no significant correlation, reflecting the non-linear thermodynamic and kinetic behaviour of the reaction. The mixing of different chemical species in the dataset hinders the identification of general trends for temperature and time due to the nature of the chemicals. Despite relatively low R coefficients, the analysis highlights the impact and correlation of parameters in DMF hydrolysis reactions.

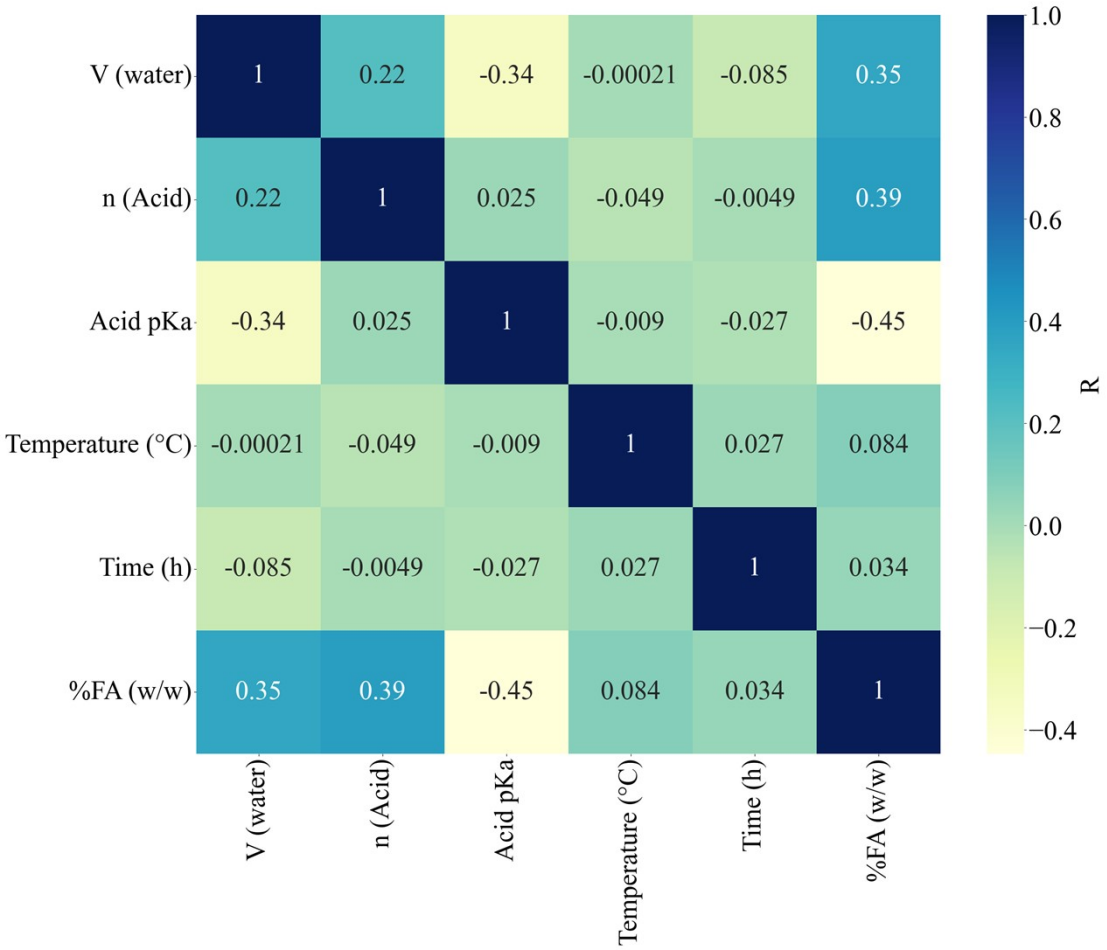

Figure S9. Pearson's correlation matrix of all data (152 samples).

Figure S10. Pair plot of all experimental data, presenting the pairwise correlation among the examined DMF hydrolysis reaction parameters. The diagonal subplots feature univariate histograms, illustrating the distribution of variable values in the dataset. The bivariate correlation between %FA generated, and reaction parameters (bottom) reveals the correlation between the investigated parameters, offering insights into the directionality of the correlation.

In alignment with the correlation coefficient matrix, the pairwise relationship plot of the parameters vividly illustrates the observed correlations among %FA, water, acid moles, and pK_a values (Figure S10). In this case, it is confirmed that time and temperature parameter distribution cannot be fit to express correlation with generated %FA in the overall data set.

The diagonal histogram plots portray the individual distribution of sampled parameters and %FA. These distributions are uneven, with water particularly affected by data corrections (see main manuscript). Additionally, the dataset is impacted by the inclusion of non-equilibrated samples (initial HCl data), leading to an uneven distribution of parameter values.

S10. Machine Learning

Five machine learning algorithms underwent testing to determine the most effective model. The analysis revealed that tree-based algorithms exhibited superior performance, with LightGBM, a gradient boosting-based approach, emerging as the top-performing option (Table S1). All models were tested using the leave-one-out cross-validation method.

Table S1. The performance assessment of selected machine learning models.

Model	R^2	MAE
Linear Regression	0.39	0.16
Support Vector Regressor	0.33	0.16
Random Forest Regressor	0.45	0.14
Extra Tree Regressor	0.36	0.15
LightGBM	**0.73**	**0.07**

The performance of the LightGBM model was evaluated using SHAP analysis. The SHAP value plots for each feature (Figure S11) reveal the expected impact of various parameters on %FA yield and align well with the results of Pearson's correlation analysis (see Section S9). Among all features, acid moles and acid pK_a have the most significant contributions to the model's predictions. The acid moles parameter shows a positive correlation, with its contribution to model predictions increasing as the number of acid moles rises. In contrast, acid pK_a exhibits an inverse trend, where its impact on the predictions diminishes as the pK_a value increases.

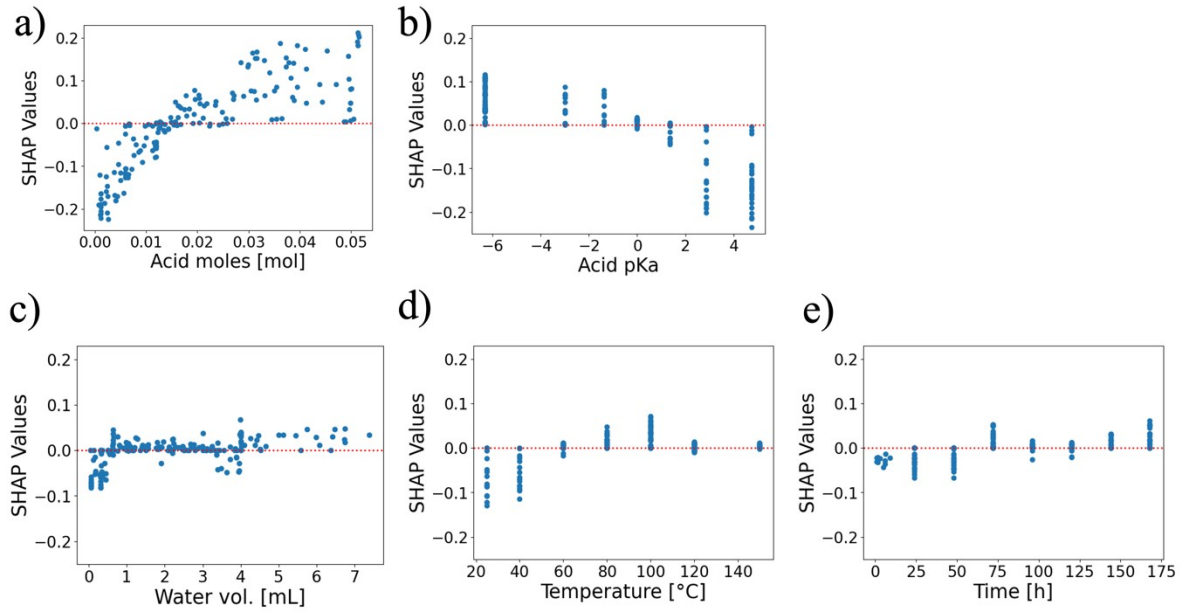

Figure S11. SHAP value plots per feature: **a)** acid moles, **b)** acid pKa, **c)** water volume, **d)** temperature and **d)** time.

To further validate the LightGBM model, the y-scrambling technique was employed to ensure that the model's performance is meaningful and not influenced by random correlations. A total of 100 permutations were conducted to test this hypothesis. The model was trained using a random 80/20 train/test split for each iteration, and the R^2 and MSE values were recorded. Random feature permutations were generated using the NumPy numpy.random.permutation() function. The results were visualised using histogram plots, with the original data highlighted by a red line (Figure S12). The significant separation between the original data and the y-scrambled data indicates that the model is capturing true correlations rather than random noise.

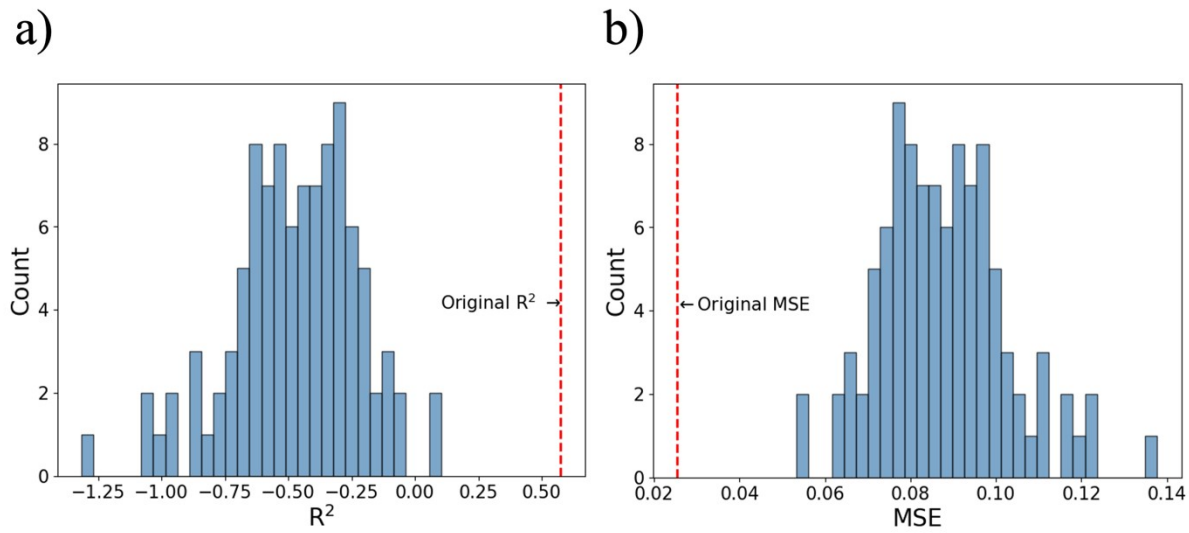

Figure S12. y-scrambling test results. Histograms of the obtained **a)** R^2 and **b)** MSE values. The red line denotes the result using the original data.

The obtained LightGBM model was used to generate maps of %FA distribution across the investigated parameter space (Figure S13). All maps were generated based on the acid mole parameter, which has been identified as a key feature in machine learning %FA predictions. These maps can help in understanding the distribution of %FA in the parameter space and in fine-tuning reaction conditions to control the amount of formic acid produced.

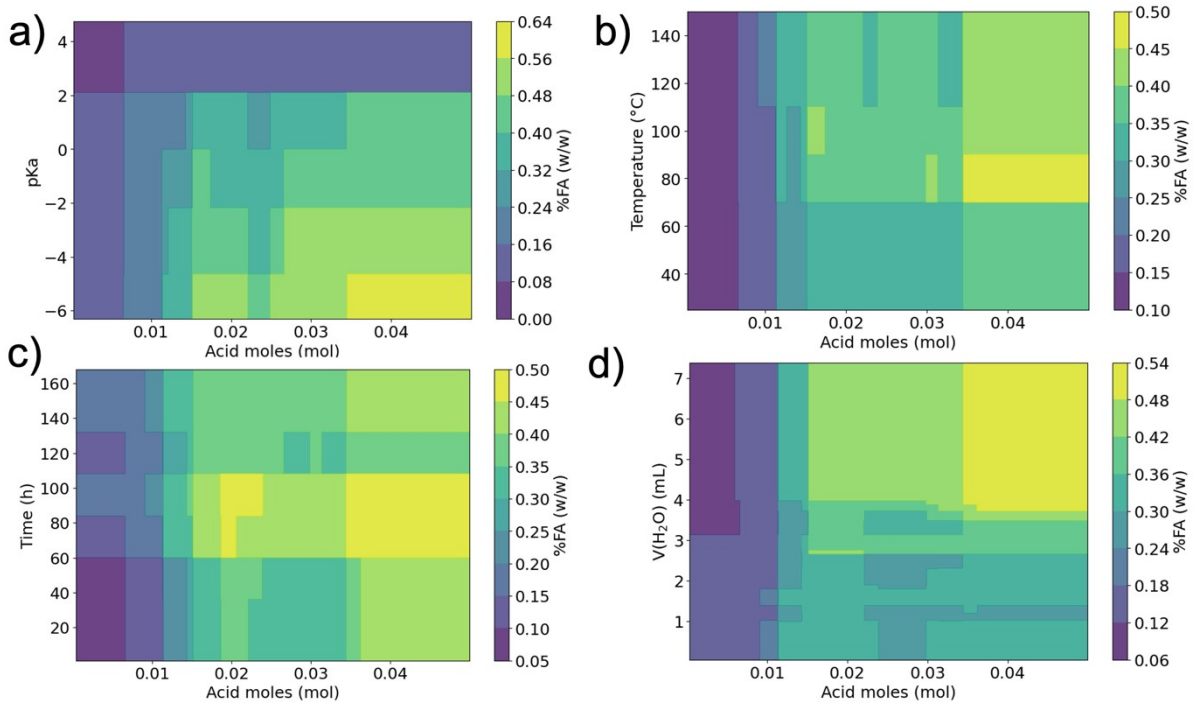

Figure S13. ML predicted %FA distribution maps for acid moles *versus* **a)** pK_a, **b)** temperature, **c)** time, and **d)** water.