chemistry

MDPI

Article

Machine Learning and Approximated Estimation Approaches for Process Design in Drug Synthesis

Andrea Repetto [1], Gianguido Ramis [2] and Ilenia Rossetti [1,*]

[1] Chemical Plants and Industrial Chemistry Group, Department of Chimica, Università Degli Studi di Milano, CNR-SCITEC and INSTM Unit Milano-Università, Via C. Golgi 19, 20133 Milano, Italy
[2] Department of Ingegnaria Civile, Chimica ed Ambientale, Università Degli Studi di Genova and INSTM Unit Genova, Via All'Opera Pia 15A, 16145 Genova, Italy; gianguidoramis@unige.it
* Correspondence: ilenia.rossetti@unimi.it

Abstract

The continuous-flow technologies in organic synthesis for the production of active pharmaceutical ingredients (APIs) are nowadays more and more applied. In-silico process design is a powerful tool able to support organic synthesis in the field of scale-up and process development. Process design feasibility and reliability depend on the availability of a well-defined chemical reaction kinetic scheme, information which is usually derived from experimental datasets collected on purpose. The latter approach is time-consuming and demanding in terms of resources. Different possibilities are here proposed to valorize widely available experimental data from explorative works with different approaches, depending on the nature, richness, and structure of the datasets. The kinetic parameters (i.e., reaction order, kinetic constant, and activation energy) of some interesting organic reactions have been approximately estimated by applying different computational methodologies, thanks to built-in experimental databases. The numerical algebra approach dealing with linear and non-linear regression analysis for the kinetic parameters has been initially considered and related to the database information for oseltamivir synthesis. The Bayesian statistic was applied to the ibuprofen case through the application of the Markov Chain Monte Carlo (MCMC) method for reaction order estimation. At last, a Machine Learning (ML) approach has been applied to the Rolipram and Pregabalin case study. The in-house developed T-ReX experimental kinetic constant database was exploited, with application of the *k*-Nearest neighbor algorithm for classification and regular expression pattern recognition. Advantages and limitations of the three approaches are discussed.

Keywords: flow chemistry; in-silico process design; numerical algebra; bayesian statistic; Markov Chain-Monte Carlo method; artificial intelligence (AI); machine learning (ML); ibuprofen; oseltamivir; Pregabalin; rolipram

Academic Editors: Jun Li, Tong Zhu and Valentine P. Ananikov

Received: 11 January 2026
Revised: 16 February 2026
Accepted: 26 February 2026
Published: 3 March 2026

1. Introduction

Reactions in continuous mode constitute a well-consolidated approach when dealing with petrochemistry, refinery, and heterogeneous catalysis fields. By contrast, fine chemicals and, more generally, organic synthesis, typically rely on a batch reaction mode [1]. However, in the last decade, the flow chemistry topic has received a rapid ascent within the organic synthesis field, and a huge amount of papers have been published in several topic-related journals, to underline how the importance of this new approach in this field is comparable to a gold rush [1–3]. The advantages of the continuous-flow organic synthesis are very

broad (e.g., from selectivity or yield improvement, to transport phenomena boosting and mitigation of safety issues). Continuous processes are often more efficient and require less manpower due to highly automated procedures. Furthermore, they are typically characterized by lower costs, reduced wastes, with an improved environmental footprint, decreased time-to-market for new drugs and products [1,3]. Particularly referring to the pharmaceutical industry, the time-to-market of a successful recipe is even more important than for base chemicals, due to the volatility of the product's life. A general improvement of safety and sustainability of the fine chemicals and pharmaceutical industries may take place, to deal with regulation and safety issues based on the present batch technology. Moreover, a wider window of operating conditions is made possible by flowing systems [1,3]. The main drawbacks of pharmaceutical batch processing are related to long production time issues and supply chain potential issues. Under flow conditions, the reaction and the subsequent purification steps take place at the same time within the same telescoped reactor process. An integrated renovation perspective in terms of continuous manufacturing is able to lead to a modernization of APIs (Active Pharmaceutical Ingredients) production [4].

Process simulation and cost estimation are reliable methodologies to support the feasibility assessment of pharmaceutical continuous processes. Mathematical programming is helpful to support the field of pharmaceutical process systems engineering and continuous process manufacturing [5], where there is an explosive interest in continuous-flow synthesis using micro- or meso-reactors, with a focus on the synthetic details. Wide datasets are available that report APIs synthesis under variable conditions to report the optimization of the reaction from a strictly synthetic point of view. Only in some very inspiring cases are chemical reaction engineering, reactor engineering, and transport phenomena issues taken into account during explorative research in organic synthesis. This mismatch seems critical for scale-up, optimization, and technology transfer [1,3]. What appears a pity is that huge datasets are available that compare conversions and yields, for many reactions, without a generalizing scale-up scope. On the other hand, such sparse data are not collected with the guiding criteria of kinetic modeling and process design, so that a detrimental cultural mismatch seems evident in the field, that retards feasibility estimates until the collection of kinetic data on purpose is planned and executed. This means delays, costs, and large manpower investments, maybe ending in a late, infeasible conclusion.

Coupling chemical engineering and process design to organic synthesis should be taken into consideration carefully to open a very promising, multidisciplinary field. On this basis, transport phenomena and kinetic models have to be applied early in drug development. The holistic organic synthesis view should meet process development [6]. Moreover, high-quality experimental data sets coming from the literature contain a large volume of data, but are often stored in different unstructured formats. This hinders their further treatment and reuse, especially for process design purposes.

While compiling data from the literature has enabled the use of larger data sets, a challenge with this approach is the bias of publications toward positive results, such that only reactions with high yields or selectivity are reported. However, negative results provide important insight into a chemical system and are necessary to build predictive models [7]. Nowadays, the advent of artificial intelligence (AI) has made a hit in organic synthesis [8]. Optimization in data-driven modeling outperforms human decision making in both average optimization efficiency (number of experiments) and consistency (variance of outcome against initially available data) [9]. The final aim is to reach the so-called autonomous discovery, and heavy pioneering academic works have been published that could be undoubtedly considered a milestone in the new upcoming frontier of chemistry [10,11].

The big question is: is it possible to take advantage of the rich information available publicly on APIs synthesis routes with the purpose of a preliminary estimation of kinetic

parameters for basic reactor sizing and costing, and ultimately for a preliminary feasibility estimate? Nowadays, the kinetic parameters estimation refers mainly to the context of theoretical chemistry for the ab initio estimation of kinetic constants [12,13] and to enzymatic reactions [14]. An interesting approach concerning the estimation of kinetic parameters is related to the Bayesian experimental design (BED) tool for guiding experiments, founded on the principle of expected information gain (i.e., which experiment design will inform the most about the model that can be predicted before experiments in a laboratory are conducted) [15].

Therefore, despite the growing availability of experimental data in organic synthesis, most published datasets are generated for reaction optimization rather than for kinetic modeling or process design. This creates a structural mismatch between the type of information reported in the literature and the quantitative requirements of in-silico process development. The present work does not aim to propose a new regression algorithm nor a single optimized modeling framework. Instead, it provides a methodological and comparative evaluation of different data-driven strategies—numerical regression, Bayesian inference, and machine learning approaches—when applied to non-ideal, literature-derived datasets. By deliberately testing these approaches under realistic data constraints, this study seeks to assess their robustness, diagnostic value, and practical relevance for preliminary process design. Within this perspective, the development and application of the structured T-ReX database represents a constructive attempt to move from unstructured synthetic knowledge toward process-relevant kinetic inference. The central contribution of this work, therefore, lies in clarifying both the limitations and the opportunities of data-driven kinetics in the context of pharmaceutical process development.

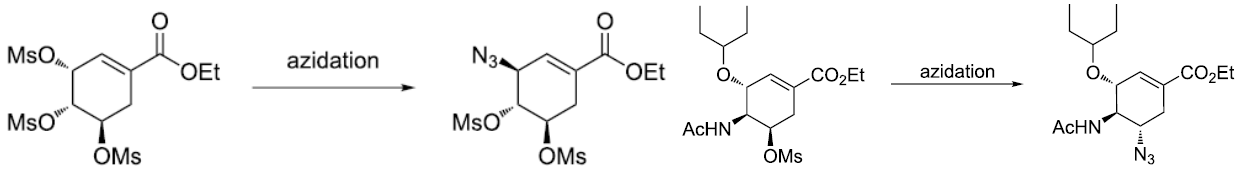
2. Materials and Methods

2.1. Selected APIs

Oseltamivir is one of the most effective antiviral drugs to treat influenza, especially in Africa [16]. Ibuprofen is a high-volume, nonsteroidal anti-inflammatory drug [17]. (±) Pregabalin is an amino acid derivative that is widely used as a therapeutic agent for nervous system disorders such as epilepsy, anxiety disorder, neuropathic pain, fibromyalgia, and as off-label generalized anxiety disorder [18,19]. The anti-inflammatory drug rolipram is a member of the γ-aminobutyric acid (GABA) family. Moreover, rolipram is known as a possible antidepressant and has been reported to have anti-inflammatory, immunosuppressive, and antitumor effects [18]. Rolipram has also been proposed as a treatment for multiple sclerosis and has been suggested to have antipsychotic effects [17].

2.1.1. Oseltamivir

The continuous-flow synthesis of oseltamivir is made up of several steps [16]. The kinetic parameters investigation has been focused here on the two intermediate azidation steps (Figure 1). The azide chemistry, in the context of scale-up, is a highly exothermic chemical process raising many safety concerns that must be handled carefully [20]. The development of a large-scale and safe process for (−)-oseltamivir phosphate synthesis is imperative [16]. For this reason, the continuous-flow synthesis offers a better handling of hazardous azide intermediates thanks to in situ consumption, preventing their accumulation [21].



(**a**) (**b**)

Figure 1. Reaction schemes of the two azidation steps investigated in the continuous-flow synthesis of oseltamivir. (**a**) C-3 mesyl shikimate azidation using different azidating agents (NaN_3, DPPA, TMSA). (**b**) C-5 acetamide azidation using sodium azide in excess. These reactions were selected as representative case studies for testing numerical kinetic parameter extraction from literature-derived conversion datasets.

2.1.2. Ibuprofen

The three-step continuous-flow synthesis of ibuprofen has been investigated previously [22]. The chemical kinetic investigation has focused on the first reaction of the telescoping synthesis. The first reaction concerns the Friedel–Crafts acylation for which mixing isobutylbenzene (IBB) and propionic acid is mediated by triflic acid (TfOH), and the second reaction is a 1,2-aryl migration by phenyliodine(III) diacetate (PIDA) in the presence of trimethyl orthoformate (TMOF) [22]. The classical variation of the three-step continuous-flow synthesis of ibuprofen concerning theFriedel–Crafts acylation exploiting $AlCl_3$ [23] was not investigated in terms of kinetic parameters estimation, since insufficient experimental data were available from the literature.

2.1.3. Pregabalin and Rolipram

The Pregabalin and Rolipram synthesis share the same organic chemistry cornerstone reactions: Knovenagel, 1,4-Michael addition, Horner-Emmons reaction, and Michael addition [18,19,24]. Moreover, both molecules could be synthesized exploiting the same continuous-flow set-up in a modular way, achieving a powerful and desirable versatility in terms of interchangeability and meeting the requirements needed [24]. The kinetic data estimation has been considered homogeneous for both molecules, since the reactions are shared with each other, and the approach used here for kinetic data mining could be considered belonging to the same wider dataset [25].

2.2. Kinetic Data Extraction

All the fundamentals and theoretical background of the methods are detailed in the Supporting Information file, Section S1.

The mathematical methods applied to the different datasets were related to numerical algebra, the use of Bayesian statistics, and the use of Machine Learning. The decision to adopt three different approaches was influenced both by the nature of the data available and by the purpose of comparing the applicability of the approaches to the different cases. The requirement to have a proper organic synthesis dataset focused on kinetic information is an important aspect within the subsequent steps of in-silico process design and emerges as a general hint to the scientific community.

3. Results and Discussion

3.1. Numerical Algebra

The numerical algebraic approach has been applied to the two azide intermediate syntheses concerning the oseltamivir case study, specifically in the analysis of extrapolated kinetic data from an experimental dataset deriving from the azidation reaction conversion plot [26]. The evaluation of experimental results has been drawn directly from a continuous-

flow process instead of the classical batch approach. Theoretical details are reported in the Supporting Information file (Section S1).

This approach prescribes that the process designer retrieve from the literature homogeneous subsets of data, collected under the same conditions, to allow regression through the variation of 1 or more variables at a time. Some papers report clearly and systematically such data (as the example reported here). The main detrimental aspect is that the experimental data refer to high conversions, and this is not optimal for kinetic study investigation, since at medium-high values of conversion, the reaction is almost done, and key kinetic information is partially lost. Furthermore, no preliminary test to ensure kinetic control (i.e., excluding transport limitations) is usually reported. This is an intrinsic bias that imposes a careful assessment of the results for their physical meaning.

Kinetic data regression often relies on linear regression, and the integral method is normally used to find the kinetic constant k when the reaction order is known [27]. Here, regression has been carried out considering the differentiation method, in order to find the reaction orders α, β and the kinetic constant k as parameter to be approximately estimated (see Equation (1)). This avoided the arrangement of the data in a forced linear correlation and forced a priori postulated (linear) model to fit data that might not have a linear nature (integral method) [28].

In some entries, the reagents were mixed in a stechiometric or quasi-stechiometric ratio. Such conditions prevent the use of some well-known kinetic equations. Moreover, the customary method for analyzing a pre-defined kinetic order (i.e., 1st or 2nd order reaction, etc.) leads to a deficiency in terms of representative goodness of output values since they are not equally reliable. The concentration is measured with a constant error, whereas the error in the linearization procedure does not have a constant value (propagation of errors) across the range, and this means that the data are heteroskedastic [28].

Non-linear regression has also been taken into consideration in order to evaluate and highlight a possible improvement in finding the kinetic parameters.

These two different approaches remarked their difference, and their careful management allowed for building up a stronger robustness and wider replicability. Non-linear regression revealed a better choice. The iterative approach has allowed us to obtain the same kinetic parameters, but with a different perspective in terms of error evaluation (non-linear least-squares), leading to a different awareness for data analysis [28].

The oseltamivir experimental dataset for the C-3 Mesyl shikimate azidation is reported in Supporting Information—Table S1. The C-5 acetamide azidation dataset deriving from the original conversion plots is reported in Supporting Information Section S2.2—Tables S2–S5.

The estimated kinetic parameters and complete list of algorithms used are collected in Table S6, while Tables S7–S13 report the calculated values of the pre-exponential factor (A) and activation energy (E), the latter retrieved by Arrhenius regression of groups of data at variable temperature, keeping fixed the other conditions (Equation (1)).

Two kinetic models have been compared:

$$r_A = -\frac{dC_A}{dt} = kC_A{}^{\alpha}; \qquad r_A = -\frac{dC_A}{dt} = kC_A{}^{\alpha} \cdot C_B{}^{\beta}; \qquad k = Ae^{\frac{-E}{RT}} \tag{1}$$

The former is able to represent the continuous-flow C-3 mesyl shikimate azidation using different azidating agents, since it takes into account that the shikimate and the azidating reagent have approximately the same concentration in the available dataset. Moreover, in order to simplify the model and lead to a first hit approach, the presence of the base (where present) has been neglected.

The second model proposed is able to represent the continuous-flow C-5 intermediate acetamide azidation using sodium azide under different temperature conditions. In this

case, the azidating agent is present in excess (3eq). This is one of the biases usually found in datasets not collected for kinetic modeling purposes, since a systematic and incremental variation of all the variables is usually not reported.

The two models are non-linear, and a partial linearization has been applied. Partial linearization is a common approach to move into a linear domain by a suitable transformation of the model. However, use of a non-linear transformation requires caution. The influences of the data values will change, as will the error structure of the model and the interpretation of any inferential result [28,29]. This method of data analysis is also useful to determine the best values of the rate parameters from a series of measurements when three or more parameters are involved (e.g., reaction order, frequency factor, and activation energy) [27].

Since the linear regression is evaluated through a system of linear equations, an iterative algorithm methodology for determining the solution vector and the corresponding parameter(s) has been applied. The Jacobi and Gauss-Seidel algorithms could be used to evaluate the convergence, robustness, and effective applicability. These methods have been applied to the C-5 intermediate acetamide azidation dataset in opposition to the direct solving method (e.g., Gauss elimination), in order to address the case study in a more robust way, with the aim of finding a more reliable and consistent solution.

The C-5 acetamide azidation dataset offers the widest range of data (i.e., temperatures and conversion values): four out of five groups of data have almost all of the initial data points below the 70% conversion.

The application of these two methods to the C-3 mesyl shikimate azidation using different azidating agents has been ruled out due to little practical use. The application of a pure non-linear model (linearized) has instead been done for this dataset, through the application of the Levenberg–Marquardt algorithm.

In addition to the searched parameters, all the datasets have been evaluated in order to assess the goodness of the dataset itself and a possible correlation. An interesting point could be the application of Bayesian Analysis of composite datasets deriving from multiple sources [30]. For this reason, and according to the theoretical background (see Section S1.1 in Supporting Information), the following key indicators have been calculated: the coefficient of determination, the condition number, and the matrix convergence. It was not always possible to calculate all the parameters as reported in Table S6.

The results were quite internally consistent, with some algorithms applied to different datasets, demonstrating reliability also for different conditions. Table S6 entries 23–26 for the Jacobi method returned the same order of magnitude of the parameters as well as the positive/negative sign of the parameters. Other methods did not ensure this consistency because the results varied widely.

An interesting case reported in Table S6 is that for the application of a row-reduced method such as the Gauss algorithm with pivoting (entry 13) and for the application of the Levenberg–Marquardt algorithm (entry 32) for the C-5 acetamide azidation at 80 °C, returned both the same results (i.e., kinetic constant, orders of reaction), but they diverged for the r^2 value. Additionally, the Levenberg–Marquardt algorithm also returned a warning message (vide infra).

Another detrimental aspect concerned the condition number of the original and/or the X^TX matrix. All these matrices were *ill*-conditioned. At this point, and according to the theory, this is the worst scenario possible since all the data estimated from the application of these algorithms cannot be considered as robustly estimated solution parameters for a certain chemical kinetic model dataset. The values of the condition number were indeed very high (e.g., $\kappa(A) > 10^2$).

The final convergence parameter value (applicable in the iterative method) failed. Almost all of the algorithms applied to the system reported in Table S6 did not converge (sufficient and necessary condition). The convergence was guaranteed just in the case of entries 29 and 31 reported in Table S6 and related to the application of the Gauss-Seidel algorithm, despite the condition number being very high ($\kappa(A) > 10^{10}$).

Referring to the C-3 mesyl shikimate azidation dataset, the kinetic constant values were of decimal or unit magnitude, whereas the reaction order parameter α was close to 1. This final value was quite unexpected, since according to the literature, the reaction follows a bimolecular S_N2 mechanism [31] with the participation of both reagents (added in equal amounts in the reaction dataset). Despite the values of r^2 were sufficiently good, the condition number matrix was not satisfactory ($\kappa(A) \leq 10^2$).

Considering the C-5 acetamide azidation dataset, the most acceptable results were the ones deriving from the application of the simplest kinetic model, $r = kC_A{}^2$, as for the C-3 mesyl shikimate azidation dataset. It returned the highest, but still acceptable, values of condition number. However, when applying the second-order model $r = kC_AC_B$, a lot of issues became apparent. Of particular interest is the application of the iterative Jacobi algorithm. The Jacobi method is used to find the eigenvalues and eigenvectors of a symmetric matrix. If a matrix is not symmetric, then the eigenvectors are not guaranteed to be orthogonal, which is a key assumption of the Jacobi method. Additionally, the eigenvalues of a non-symmetric matrix are not guaranteed to be real numbers, while this is required in the Jacobi method. Therefore, this method does not work with a non-quadratic matrix, and thus the dataset has been relaxed to a matrix of dimension 3×3 in order to have the same dimension as the initial guess vector (transposed vector). The refining data points process has been settled in order to take into consideration the early stage of reaction—the beginning of reactions themselves (i.e., low conversion). In this way, it was possible to address both the chemical kinetic reaction and computational method (quadratic matrix) demands.

To overcome this issue, it is advisable to apply the singular value decomposition (SVD) approach. The order of reaction should also be considered from the chemical point of view, since the most acceptable data points were the ones retrieved at low temperature, and the order was suggested between 1 and 2. On the contrary, when regressing data collected at higher temperatures, the order of reaction reached unreliable values (i.e., 5 and 7).

A warning message was returned during the application of the Levenberg–Marquardt algorithm in all cases, indicating that the model was overparameterized and there was no solution, implying the use of a simpler model or getting more data.

Arrhenius regression was done for every regressed solution, for the moment remaining blind to the physical meaningfulness, and the activation energy and frequency factor have been calculated based on the k values reported in Table S6. The most significant examples are reported in Table 1 for the C-3 azidation with NaN_3 (Table S6, rows 1–3) and Table 2 for the C-5 azidation with NaN_3 (Table S6, rows 8–12). All the other cases are reported in Supporting Information Section S2.3—Tables S7–S13. Examples of Arrhenius regression are reported in Figures 2 and 3.

Table 1. Arrhenius regression plot for the C-3 azidation with NaN_3 (Table S6, rows 1–3). Activation energy and frequency factor calculated for k values reported in Table 1.

T(K)	273	283	298
Kinetic constant (k)	1.72×10^{-1}	2.56×10^{0}	4.30×10^{-1}
Frequency factor A (1/min)		1.26×10^{3}	
Activation energy E (cal/mol)		4343.81	

Table 2. Arrhenius regression plot for the C-5 azidation with NaN_3 (Table S6, rows 8–12). Activation energy and frequency factor calculated for *k* values reported in Table S6.

T(K)	353	373	463
Kinetic const. (*k*)	7.62×10^4	6.22×10^3	1.48×10^7
Frequency factor A (1/min)		1.01×10^{16}	
Activation energy E (cal/mol)		19,181.56	

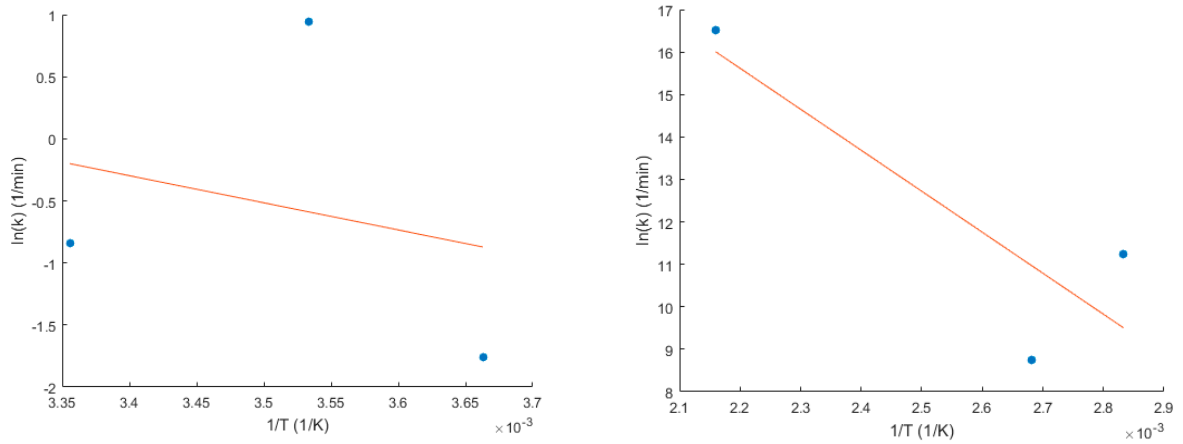


Figure 2. Representative Arrhenius plots obtained from regressed kinetic constants for selected datasets. (**a**) C-3 azidation with NaN_3 (Table S6, rows 1–3). (**b**) Refined C-5 azidation dataset with NaN_3 (Table S6, rows 8–12). Linear regression was performed after selective exclusion of numerically unstable entries. The slope provides the apparent activation energy, while the intercept corresponds to the pre-exponential factor. The variability reflects dataset sensitivity to regression conditioning. The regression results are summarized in Tables 1 and 2. Blue dots: data points; orange lines: Arrhenius regression.

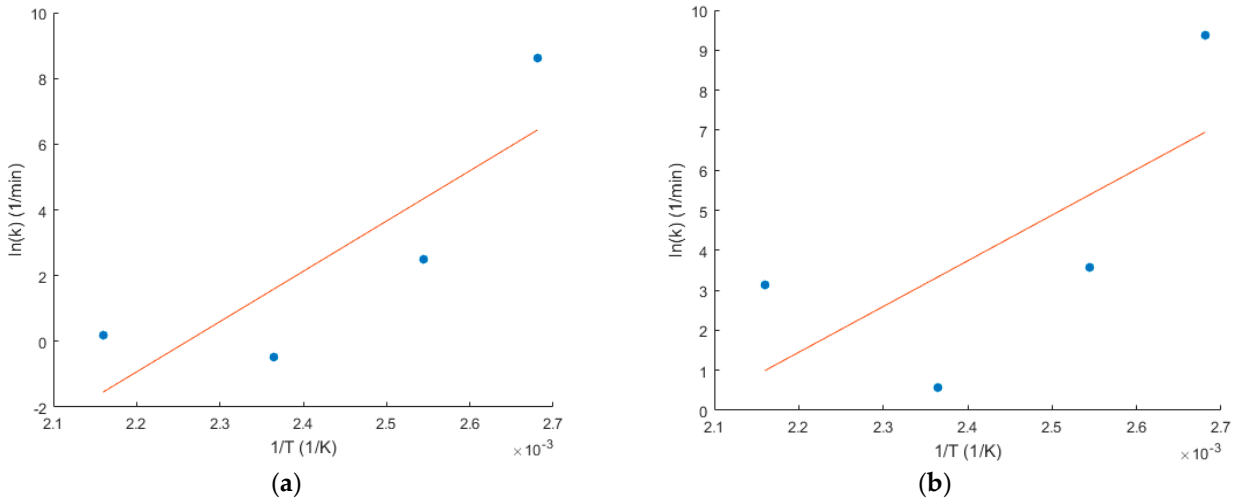


Figure 3. Arrhenius plots illustrating regression instability in higher-order kinetic models. (**a**) C-5 azidation using the second-order model $r = kC_A{}^{\alpha}C_B{}^{\beta}$ (Table S6, rows 13–17). (**b**) Same reaction fitted via Levenberg–Marquardt (rows 32–36). Positive slopes correspond to negative apparent activation energies, indicating non-physical temperature dependence likely arising from ill-conditioned regression and non-kinetic experimental regimes. Blue dots: data points; orange lines: Arrhenius regression.

The discussion is indeed finalized to prove that using the whole dataset does not lead to reliable conclusions, thus manipulation of the dataset would be needed, with consequent unreliable results and conclusions. The refinement of the dataset was not performed to "improve the fit," but to explicitly diagnose numerical instability and parameter non-identifiability. Data points were excluded only when they were demonstrably associated with extreme condition numbers, non-convergence of iterative solvers, or physically inconsistent regression outputs. The objective was not to bias the result toward positive activation energies, but to evaluate how regression stability depends on dataset conditioning.

Regarding the use of the Jacobi algorithm, its inclusion was exploratory and comparative, intended to test the sensitivity of iterative solvers to ill-conditioned matrices. We agree that reducing the dataset to a 3×3 system is not a statistically rigorous regression strategy; this step was meant to illustrate computational constraints rather than to establish definitive kinetic parameters.

The estimated values of activation energy were too low to have physical meaning. This is due to inaccurate regression, and a possible further explanation can reside in some diffusive limitations, which do not allow for drawing even qualitative conclusions on the temperature effect. Indeed, during a kinetic testing campaign, the preliminary checks should be devoted to carefully eliminating any possible interference from diffusion and mass/heat transfer. This was not possible here since, as stated, the scope was to use the huge dataset of organic synthesis data to check their possible application for kinetic data extraction [32]. However, alternative explanations (e.g., mechanistic shifts or equilibrium effects) are possible to justify negative activation energies, besides diffusion limitations.

In complex systems with multiple competing processes, diffusional limitations can induce deviations from the expected temperature dependence (non-Arrhenius behavior) [33]. The main issue was related to negative values of activation energy, which is contrary to normal expectations, since an increment of reaction rate with temperature is conventionally expected [34]. They could be considered, in principle, a mistake deriving from the computational approach used or from the intrinsic nature of the experimental results and their manipulation. Moreover, in case of a negative temperature dependence of the reaction rate, great caution must be exercised when their explanation is applied to chemical reactions under harsh conditions [35], and statistical thermodynamics functions may be ill-defined [36].

Since the negative activation energy values have been found for the dataset elements having a matrix condition number that is very high, a dataset refining procedure was done, leading to positive values only. The approach was as follows:

- Arrhenius activation parameters regression was done using kinetic constants coming only from the dataset points having condition number strictly equal to $\kappa(A)10^2$
- Refining the kinetic constant dataset from the outlier values.

As a first hint, the refining process was based on ruling out the data that led to a high condition number.

The nature of the dataset handled (i.e., a small experimental data source with alleged error bias affected) did not allow the application of a robust statistical method for outlier identification. Hence, they have been identified as the variable couple leading to reduced quality of linear regression fitting.

It should be remarked that the datasets were included based on three primary criteria: availability of time-resolved concentration or conversion data, consistent operating conditions within homogeneous subsets, and sufficient variation of at least one independent variable. High-conversion points were retained initially to reflect the real structure of exploratory synthetic datasets, but their impact on parameter identifiability was subsequently assessed through condition number analysis and regression stability.

Outliers were not removed arbitrarily; instead, entries associated with extreme condition numbers, non-convergence of iterative algorithms, or physically unrealistic parameter values (e.g., negative apparent activation energies under otherwise monotonic trends) were flagged as numerically unstable. Dataset refinement consisted of excluding only those subsets demonstrably responsible for ill-conditioning or regression breakdown. We will revise the manuscript to explicitly summarize these criteria and to clarify how each methodological approach (linear regression, Bayesian inference, ML) responds differently to data sparsity, heteroskedasticity, and structural bias.

The values of Arrhenius activation energy showed a positive value, even if the regression was not very accurate, as shown in Figure 2, for C-3 azidation with NaN_3 (rows 1–3, Table S6) and the refined C-5 azidation with NaN_3 (rows 8–12, Table S6). The Arrhenius plots returning negative activation energy are instead exemplified in Figure 3, concerning the C-5 azidation with NaN_3 (rows 13–17 and 32–36, Table S6).

In conclusion, it is reasonable to assume that the negative activation energy values were more related to an inappropriate intrinsic nature of the experimental data collected than to the computational algorithm. This is a pivotal aspect that should be considered for future development: the proper experimental data collection at the time of laboratory-scale trials plays a key role in the subsequent data handling.

Reasonable kinetic parameters were retrieved for the C-5 azidation reaction, as reported in Table 2. These may be used (with caution) to make a preliminary feasibility assessment for a continuous-flow plant for the production of the selected API. On the contrary, too low A and E have been found in the best cases for the C-3 reaction (Table 1).

To summarize, the numerical algebra approach was applied to the oseltamivir azidation steps using literature data derived from continuous-flow conversion plots. The objective was not to perform a rigorous kinetic study, but to evaluate whether exploratory synthetic datasets can be exploited for preliminary parameter estimation in process design.

Linearization (with both linear and non-linear regression) allowed rapid parameter estimation but amplified heteroskedasticity and error propagation effects. Non-linear regression, particularly using the Levenberg–Marquardt algorithm, provided more consistent solutions, although convergence warnings frequently appeared, indicating overparameterization or rank deficiency.

A major limitation emerged from the intrinsic structure of the datasets. Most data points correspond to high conversion values, where kinetic sensitivity is reduced, and no information is available regarding transport limitations. As a result, several regression matrices were ill-conditioned, and iterative methods did not systematically converge. Full numerical details and algorithmic comparisons are reported in the Supporting Information.

For the C-3 azidation, reaction orders close to unity were obtained, despite literature evidence suggesting bimolecular behavior. For the C-5 reaction, higher-order models generated unstable and physically unrealistic parameters. Arrhenius analysis frequently led to low or even negative apparent activation energies. After refining the dataset by excluding entries associated with extreme condition numbers, positive activation energies were obtained, although still affected by significant uncertainty.

These findings highlight that conventional regression techniques, even when computationally robust, cannot compensate for datasets not originally designed for kinetic modeling. The exercise, therefore, serves as a methodological stress test, demonstrating that process-relevant kinetic parameters require appropriately structured experimental campaigns. Thus, the numerical inconsistencies observed should not be interpreted as methodological failure, but as quantitative evidence of the structural mismatch between synthetic optimization datasets and process design requirements.

3.2. Bayesian Statistics

The kinetic parameter estimation related to the ibuprofen case study has been reported in some selected papers [37–39]. In our previous work [39], a simple algebraic approach was applied, taking into consideration two different second-order kinetic models to estimate the desired kinetic parameters to be used in a preliminary Aspen Plus flowsheet (as in Section 3.1). Here, the intention is to broaden the investigation on the available dataset, exploring the potential of statistical methods for kinetic modeling.

The original experimental dataset was readapted from ref. [22] is summarized in Supporting information Section S3.1—Tables S14 and S15, whereas in Table S16, it is possible to find the two different kinetic models previously considered in [39]. Also in this case, the original experimental data were not derived from a kinetic study: they are a collection of data retrieved under different reaction conditions to compare yields and conversions for the purposes of empirical optimization in organic synthesis. Therefore, all the discussed biases also apply here.

This case history will exemplify the use of Bayesian statistics and, in particular, the application of the Markov Chain Monte Carlo algorithm (MCMC). The theoretical basis is discussed in Supplementary Information Section S2.1.

The MCMC algorithm has been used to test some hypotheses related to the problem class of optimization. This choice was proposed to highlight the possibility of disengaging from a cultural scientific bias that can affect the way in which the data are handled and processed [6,7,28].

The first step was to perform a least-squares fit of the kinetic parameters and generate confidence intervals (95% confidence interval half-widths) for the non-linear model reported in Equation (2) from single response data, using the approximate posterior density function (3):

$$r_A = -\frac{dC_A}{dt} = kC_A{}^{\alpha} \cdot C_B{}^{\beta} \tag{2}$$

$$\pi(\theta, \sigma|y) \propto \sigma^{-(N+1)} exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta_M)^T \left[X^T X|_{\theta_M}\right](\theta - \theta_M)\right\} exp\left\{-\frac{\nu s^2}{2\sigma^2}\right\} \tag{3}$$

Equation (2) describes the case study capturing the chemistry behind the subsequent computational approach.

Non-linear regression was done by returning the vector containing the estimated kinetic parameters $\theta = [k\ \alpha\ \beta]^T$, which in turn was followed by the generation of the confidence intervals of the non-linear regression parameters.

The results obtained have been tested more rigorously using MCMC simulation with the exact marginal posterior and confidence (credible) intervals. In this way, the 1-D marginal posterior density $p(\theta_1|y)$ and $p(\theta_1|y)$ has been evaluated. The 1-D marginal posterior density gives information about the probability of obtaining a certain θ_1 or θ_2 knowing the experimental value y.

$$\pi(\theta|y) = \int_0^{\infty} \pi(\theta, \sigma|y) d\sigma \propto \left[1 + \frac{1}{\nu s^2}(\theta - \theta_M)^T \left[X^T X\,|_{\theta_M}\right](\theta - \theta_M)\right]^{-\frac{N}{2}} \tag{4}$$

The 1-D marginal posterior density requires applying Equation (3) and integrating the mutually exclusive parameters one at a time. In this way, the other variables in the subset of variables are retained. The use of the 2-D marginal posterior density $p(\theta_1, \theta_2|y)$ It is also possible since it gives important information about the probability of obtaining a certain θ_1 AND θ_2 value, knowing (conditioned |) the experimental y (predictor). The 2-D marginal posterior density has not been computed in this case study.

A reduced ibuprofen dataset was used, since not all of the original entries were useful for kinetic data extrapolation. The non-linear regression has been applied to the dataset R1 reported in Supporting Information Section S3.2—Table S16. The results were not consistent with elementary reactions. Moreover, Jac parameters, representing the Jacobian of the model functions (i.e., the linearized design matrix), led to a warning message, indicating that the problem was overparameterized and model parameters were not identifiable. The overparameterization may take place on small datasets such as those managed by us. The intrinsic nature of 'primitive data not developed for built-in kinetic studies' drawn from literature, limited us in resolving the issue—this led us once again to declare the importance of having a very large dataset built in a proper way. The Matlab script returned several θ solution vectors, including the following one:

$$\theta = [3 \quad 185 \quad -209]^T \tag{5}$$

This θ vector has been chosen from all the other possible vectors since deriving from the fullest and most complete dataset (Supporting Information Section S3.2—Table S16). The data computed were the ones valid for Reactor R1, entries from 1 to 3, and it returned a set of kinetic parameters suitable to be used as retained (marginal) variables. The parameters α and β were the parameters subjected to MCMC for testing the hypothesis, though, looking at the numerical values, it was not meaningful from the physical point of view—the unrealistic reaction order values are related to the kinetic model adopted and reasonably able to describe the classic ibuprofen reaction [5]. This was, anyway, a good candidate for testing this statistical method due to the blindness to exact solutions that accompanies this kind of approach.

According to these computational results, the MCMC simulation for the testing hypothesis through the 1-D marginal posterior density has been performed using the parameter θ values as a new initial guess. Several attempts have been tested, e.g., varying the initial guess of the parameters. These tests were necessary to rationalize the output $p(\theta_i|y)$.

The MCMC simulation for 1-D marginal posterior density using the vector as an initial guess $\theta_0 = [1.0\ 1.0]$ and fixing the kinetic constant parameter at a value of 3, led to the results reported in Figure 4 (for the complete results, refer to Supporting Information Section S3.2—Figure S14).

The outcome was not satisfactory. When trying to force constraints to the parameters search space (partial reaction orders ca. 1), the x axis interval was set from 0.8 to 1.2. This aspect highlights how forcing a priori an interval for the search for the solution may lead to poor results. Application of space constraints to the $\mathbb{R}$ set may affect the solution of a system; the system solution may not be found in that reduced $\mathbb{R}$ subset, and for this reason, it is advisable to consider a wider interval. On the other hand, a too large interval may be dispersive in terms of computational efficiency as well as in terms of chemically meaningfulness of the solution space.

Moreover, the curve shapes were not regular and varied widely. This last feature could be related to a sub-optimal coding structure, as well as to the nature of the data or to the default values [1.0 1.0] chosen as an initial guess. Anyway, the values of 1D-marginal probability (intensity of the main peaks) on the y axis were too low to consider the outcome θ_1 and θ_2 as at least approximately suitable to represent the proposed kinetic model.

Several other attempts have been made, and the results are summarized in Figure 5 (full details can be found in Supporting Information Section S3.2—Figure S15).

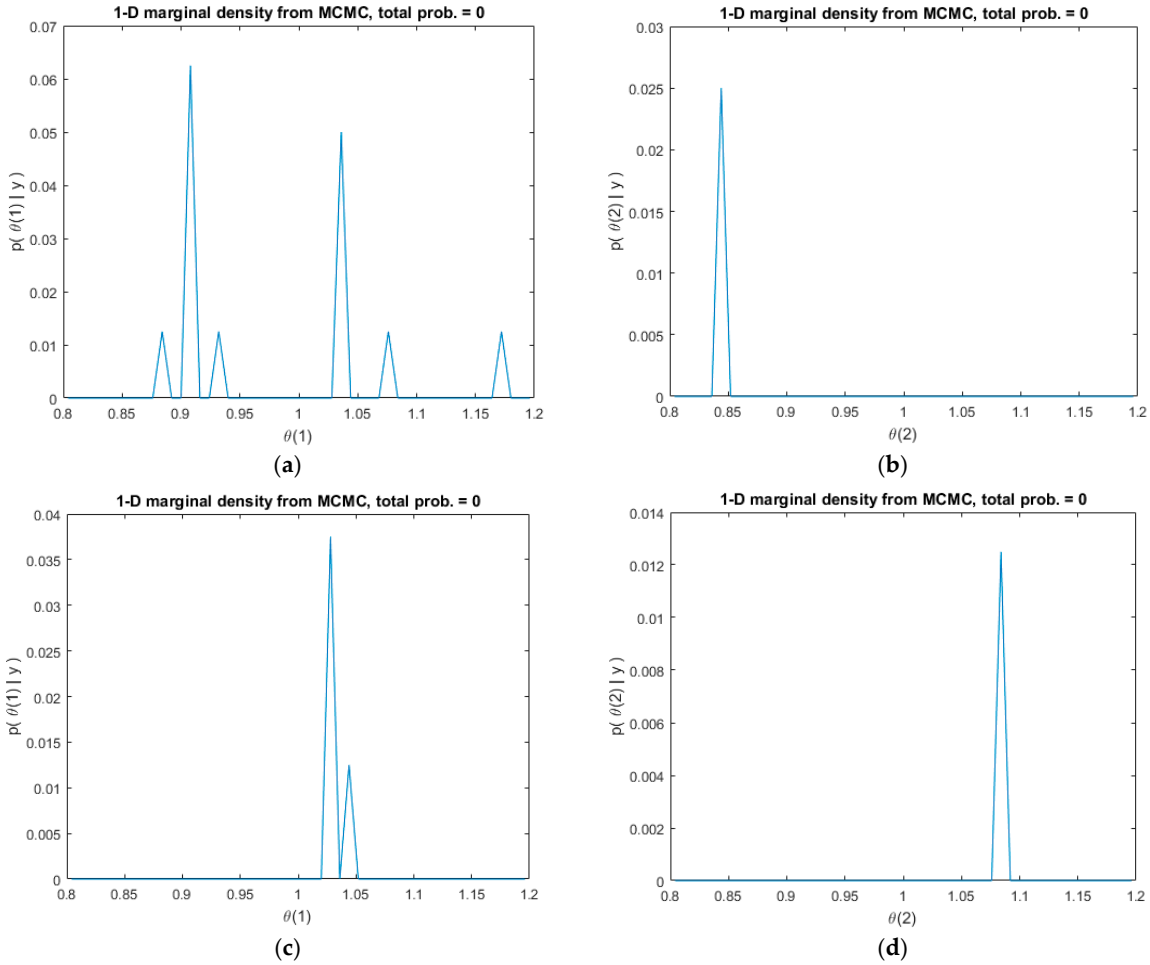


Figure 4. 1-D marginal posterior densities obtained from the first MCMC trial for the ibuprofen case study, using as initial guess the vector θ_0 = [1.0 1.0]. (**a**,**c**) Marginal distribution of reaction order α. (**b**,**d**) Marginal distribution of reaction order β. The constrained search interval is reported on the *x*-axis. Irregular curve shapes and low probability intensities indicate weak parameter identifiability and insufficient data information content.

The x axis interval considered in the second trial actually included the α and β values obtained in the first approximate posterior density computational solution [−209 185]. The interval considered in MCMC has been settled according to the ones obtained previously, so the subsequent density curve was reasonably assumed in a certain expected (desired/not desired) region of the $\mathbb{R}$ space. The curve shapes were more reliable, although the probability values in terms of $p(\theta_i|y)$ were still small, far from reaching the maximum value (1)—this means that the posterior density is less plausible given the data and prior. A lack of posterior density should be sought in the quality of the data.

Additional comments should be related to the values assumed by the parameters. θ_1 and θ_2. As shown in Figure 5, there are some values commonly shared and interchangeable with each other. This means that the $p(\theta_1|y)$ or $p(\theta_2|y)$ is the same and θ_1 and θ_2 could assume the same values, but not at the same time (see Supporting Information Section S3.3 for more plots describing the several trials evaluated). To further analyze this aspect, the 2D-marginal posterior density should be taken into consideration, because, as already introduced, it considers the probability of obtaining a certain θ_1 and θ_2 value, knowing (conditioned |) the experimental y (predictor). So, in the 2D-marginal posterior density, the fixed θ_1 and θ_2 are able to explain the probability of having both such exact values, having the experimental datum y. However, since the 1-D marginal posterior density evaluation

was insufficiently meaningful, in terms of results and reliability of the computational part, the 2D-marginal posterior density has not been computed.

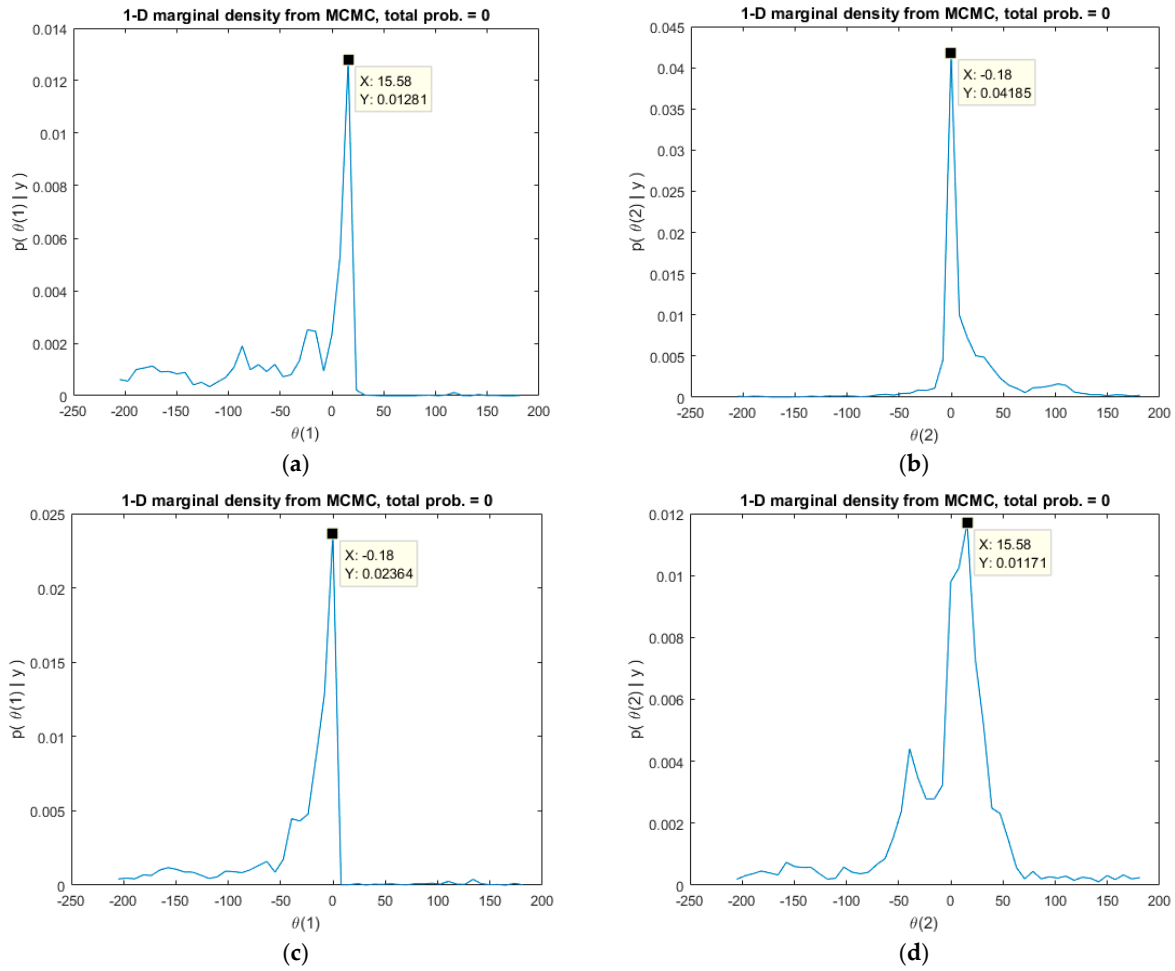


Figure 5. 1-D marginal posterior densities from the second MCMC trial with expanded parameter search intervals. (**a**,**c**) Marginal posterior for α. (**b**,**d**) Marginal posterior for β. Broader parameter domains partially stabilize the density shape, but low posterior probabilities confirm limited identifiability and dataset sparsity.

To summarize, Bayesian statistics applied to the ibuprofen dataset was not successful in retrieving a reasonable set of kinetic parameters, at least not more than simplified numerical regression. This is ascribed to the nature of the dataset, likely insufficiently rich for a genuine statistical approach, besides the already mentioned limits of using data not derived for this purpose. The intention was not to present a fully validated Bayesian kinetic model, but rather to explore the behavior of probabilistic inference when applied to non-ideal literature datasets. The inconclusive outcome primarily arises from structural limitations of the dataset: the number of experimental points is small, the explored variable space is narrow, and the data were not originally generated for kinetic parameter identification.

These features lead to weak parameter identifiability and practical non-identifiability, which in turn produce flat or irregular posterior distributions. Model overparameterization further amplifies this issue, as the available data do not sufficiently constrain the reaction orders. The MCMC algorithm itself did not fail computationally; rather, it correctly reflected the lack of information content in the likelihood function. In this sense, the Bayesian framework functioned diagnostically, revealing the intrinsic limitations of the dataset.

Overall, this Bayesian analysis serves to demonstrate how advanced statistical tools cannot compensate for insufficient or poorly structured experimental data, thereby reinforcing the central aim of this study.

3.3. Artificial Intelligence and Machine Learning

The third approach used for kinetic parameters estimation has been based on machine learning. Part of a manually-pre-compiled dataset of kinetic constants was used as a training algorithm. This dataset has been built in-house, collecting a total of more than 500 kinetic constant values concerning the reaction of interest for the synthesis of Pregabalin and Rolipram. In particular, the focus has been devoted to the intermediate steps discussed above (i.e., Knovenagel, 1,4-Michael addition, Horner-Emmons reaction, and Michael addition), since these types of reactions are the bottleneck to achieving the final product. Data mining consisted of the collection of all the kinetic constants related to the selected reactions from the Herbert Mayr database [25,40], which was developed for the possibility of the existence of a general nucleophilicity and electrophilicity reactivity scale [41–44] (Supporting Information, Section S4.1). Among the two electrophiles and nucleophiles macro categories, several different classes of molecules have been chosen according to the intermediates present in the Pregabalin and Rolipram synthesis (e.g., phosphorous ylide, Michael acceptor, malonate derivatives, etc.).

A comprehensive list of molecules considered, extracted from our developed database, is reported in Supporting Information Section S4.2. For each class of molecules, several references have been considered in order to retrieve the kinetic constants [45–56]. All the kinetic parameters (e.g., kinetic constant, electrophile/nucleophile parameters) mined have been collected into our in-house developed database called T-ReX. For a quite exhaustive example of this database, refer to Supporting Information, Table S17. The construction of the database had to take into consideration some features in order to optimize the progressive implementation and make the approach simpler and more usable.

The temperature was 20 °C in all reaction cases considered, so this variable has been removed. The solvent has also been removed from the T-ReX database as a first instance, in order to reduce the complexity of the computational aspects, reducing the number of correlations and classifiers. The nature of the solvent (e.g., non-polar/polar) plays a pivotal role in polar reactions (e.g., S_N2), and appropriate consideration should be taken into account in the future. In the present case, the choice was to focus on the correlation between the electrophile/nucleophile character of the selected molecule pair and the resulting kinetic constant. The solvents reported in literature (i.e., DMSO, DCM, H_2O, MeOH/CH_3CN) did not give us additional helpful information in handling the data. The parameter's data values collected are naturally derived from the use of a certain solvent—in other words, the solvent accounting is redundant. The present work is not related to mechanistic insight for the development of a kinetic model, but rather highlights the possibility of drawing useful kinetic information from non-tailored experimental kinetics results. Solvent effects can significantly influence polar reaction kinetics, and here, the deliberate omission was a dimensionality-reduction choice aimed at isolating intrinsic electrophile–nucleophile reactivity in a first-order approximation. Of course, this limits direct physical interpretability of the predicted rate constants, so that the reported values should be interpreted as solvent-implicit estimates tied to the original experimental conditions.

The robustness of this database has been checked thanks to the cross-match of a certain molecule in different references (e.g., a benzhydrylium ion has been found in several publications as a substrate to evaluate its reactivity as an electrophile with different nucleophiles). No redundant investigation was included.

The T-ReX database is an advancement with respect to the original set of data, which constitutes an unstructured database where large volumes of data are stored in an unstructured format [7] or at least as a semi-structured data lake in which a certain dataset structure provides a better representation of the chemical space for the reaction [40].

On the contrary, the T-ReX database is structured and thus suitable for a supervised machine learning approach. It includes more than five hundred kinetic constants (entries), assigning the class to each electrophile/nucleophile involved in a certain reaction. This approach proved winning, especially in the use of k-Nearest Neighbor algorithm for easy classification and pattern recognition.

The parameter able to differentiate the two different categories of compounds was selected as the Linear Free-Energy Relationship (LFERs) for Substituent Effects, in order to quantify the correlation between substituent groups and resulting chemical properties (as electrophile or nucleophile).

In many cases, structure-reactivity relationships can be expressed quantitatively in ways that are useful both for the interpretation of reaction mechanisms and for the prediction of reaction rates [57]. The most widely applied of these relationships is the Hammett equation, which correlates rates and equilibria for many reactions [57]. The σ parameters able to describe the influence of substituents on a molecule, were unfortunately not useful and reproducible in the present case, since they are strictly correlated to the substituted phenyl group. The Hammet σ approach has been ruled out, and the focus was driven toward two more accessible and closer case-study parameters: the experimental E and N parameters found in Herbert Mayr's linear relationship among the electrophiles/nucleophiles and the resulting kinetic constant of their reaction as reported below:

$$log\, k_{20\,^{\circ}C} = s_N(N + E) \quad (6)$$

The choice to consider just E and N relies on the fact that with just these two parameters, it was possible to perform pattern recognition using the k-Nearest Neighbor algorithm.

The covered nucleophilicity range is $-8.80 \leq N \leq 31.92$, whereas the electrophilicity range is $-29.60 \leq E \leq 8.02$. These are the values reported in [25] and covering all of the 1273 nucleophiles and 1344 electrophiles. In the present case, the interval was set narrower and covers a symmetrical overlapped interval of $-30.0 \leq \mathrm{N/E} \leq 30.0$. Herbert Mayr's equation could be considered as belonging to a primordial work in the development of data-driven science in organic chemistry, and these strategies are still rooted in LFERs, of which the Hammett relationship (and the modified Taft equation) is the paradigmatic example [7].

The final step was to develop a suitable approach for the evaluation of the dataset and application of a machine learning technique for kinetic constant extraction. The Principal Component Analysis (PCA) approach (Supporting information Section S1) has also been applied as a bridge between the classical methods and the ones belonging to the machine learning field, among which the *k*-Nearest Neighbor (*k*-NN) was selected.

The regular expressions have been applied to the SMILE notation. The SMILE string notation has been chosen according to natural language processing tasks and widely used in the literature [8,10,11]. The regular expression is able to address the readiness of the SMILE string in case-sensitive mode, avoiding misleading interpretation (e.g., uppercase/lowercase topic).

The results derived from the application of the PCA technique to the T-ReX database are listed below and plotted in Figure 6.

Principal Component (PC) list	
PC1	$k_{2\_experimental}$
PC2	$k_{2\_calculated}$
PC3	s_nucleophile
PC4	N_nucleophile
PC5	E_electrophile

(**a**)

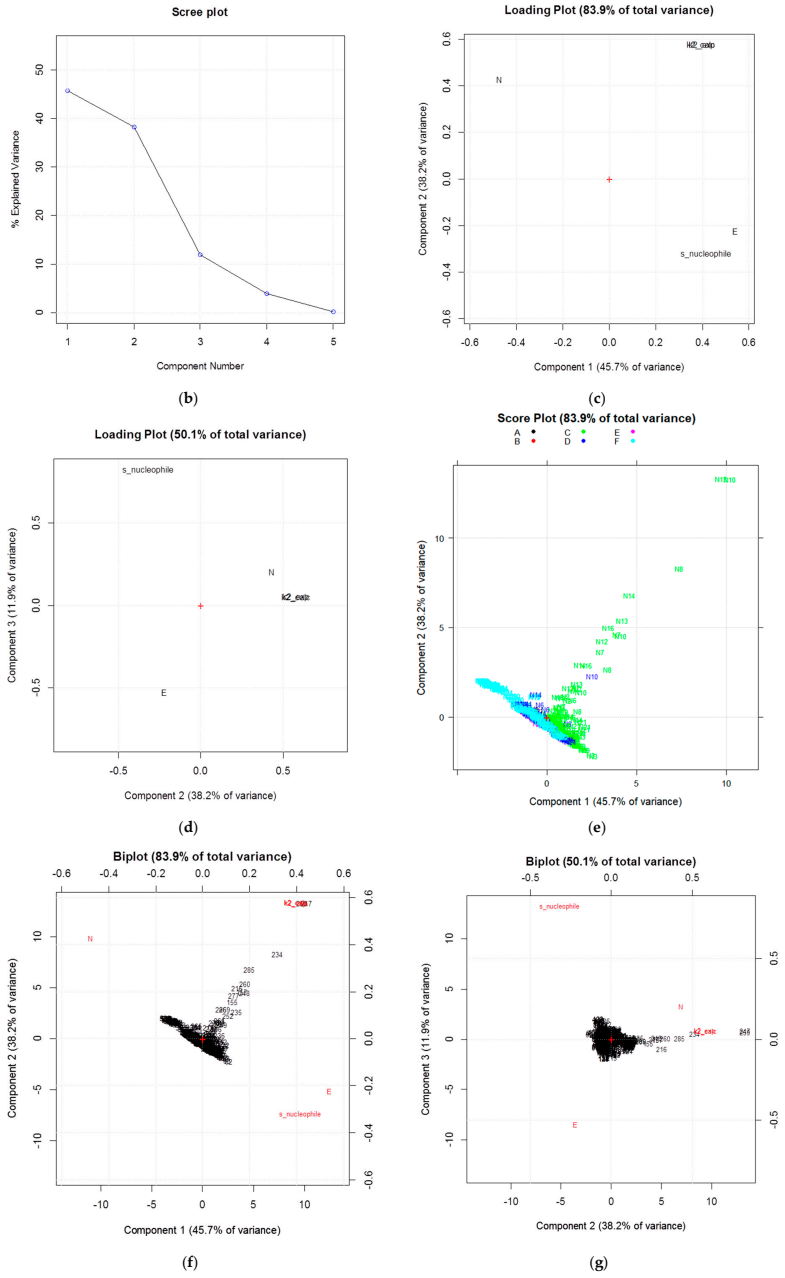


Figure 6. Principal Component Analysis (PCA) of the T-ReX kinetic database. (**a**) Principal component structure. (**b**) Scree plot showing eigenvalue distribution and variance explained. (**c**) PC1 vs. PC2 loading plot (83.9% cumulative variance). (**d**) PC2 vs. PC3 loading plot. (**e**) PC1 vs. PC2 score plot highlighting class clustering of electrophiles and nucleophiles. The green points (electrophile) were predominantly influenced by s_N and negatively influenced by the electrophile parameter E. The top right corner was influenced by k_{exp} and k_{calc}. The PC1 vs. PC2 score plot highlights which samples are majorly influenced and by which other variables. (**f**) PC1 vs. PC2 biplot integrating sample distribution and variable influence highlights the influence of original variables (in red) on the samples (black). (**g**) PC2 vs. PC3 biplot highlights the influence of original variables (in red) on the samples (black). The analysis illustrates that most variance is captured by kinetic and reactivity parameters, supporting descriptor selection for *k*-NN classification.

The first diagram is the Scree Plot, also called the Elbow Plot, and gives the principal components (Figure 6b), which are listed for convenience in Figure 6a. According to the scree test, the elbow of the graph where the eigenvalues (K1 or Kaiser criterion) seem to level off is found, and factors or components to the left of this point should be retained as significant. According to these definitions, the variance is explained by 5 components, but mainly depends on the components one and two, which, following the dataset structure, represent the kinetic parameters.

The loading plot reports the spatial disposition of the five components, evaluating the goodness of the spatial positioning of the variables. Moreover, the component PC1 (x axis) and PC2 (y axis) were able to explain the 45.7% and 38.2% of the variance, respectively, for a total 83.9% of the variance (Figure 6c). This means that these two components have a high weight in explaining data dispersion, since the idea is to have as much variance explained, with as few principal components as possible.

The loading plot reports the spatial location of all five components, evaluating the goodness of the spatial distribution of the variables. Moreover, the component PC2 (x axis) and PC3 (y axis) were able to explain respectively the 38.2% and the 11.9% of the variance, respectively, for a total 50.1% (Figure 6d).

The PC1 vs. PC2 score plot highlights which samples are majorly influenced and by which other variables. In the present case, the green points (electrophile, i.e., benzhydrylium ions) were predominantly influenced by s_N and negatively influenced by the electrophile parameter E. The top right corner was influenced by k_{exp} and k_{calc}. The other labeled and colored points were all the other classes of electrophiles that showed their relationship with the corresponding nucleophile (Figure 6e).

The biplot for PC1 vs. PC2 could be considered as a merged diagram. It highlights the original samples (in red) and the variables with their own influence on the samples (black) (Figure 6f).

The biplot for PC2 vs. PC3 is analogous to the previous one, but the two principal components are obviously different (Figure 6f).

The k-NN approach has been applied for pattern recognition in order to classify different classes of molecules belonging to two different categories (electrophile/nucleophile) on the basis of different parameter values (E/N) and pK_A. The k-NN algorithm has been chosen for its own simplicity (implementation and explanation), to a lazy learner, high memory storage for the full training set, and is applicable to small datasets. The k-value applied (k = 3) allows the algorithm to look at the three (E/N and pK_A) closest parameters with respect to the new one. In the present study, the k-value selected for the k-NN algorithm was k = 3. This choice reflects a balance between local sensitivity and robustness, given the moderate size of the T-ReX dataset (ca. 500 entries). Lower values (k = 1) led to excessive sensitivity to single-point variability, whereas higher values reduced chemical specificity in classification.

The selection of the k-NN algorithm was mainly driven by the specific nature of our dataset and objectives. The T-ReX database is relatively small (ca. 500 entries), structured, and chemically interpretable; therefore, a non-parametric, instance-based method was preferred over more complex models (e.g., neural networks or ensemble methods) that would require significantly larger datasets to avoid overfitting. Moreover, k-NN provides local, similarity-based predictions that are directly aligned with the chemical intuition underlying Mayr's reactivity parameters (E and N), preserving interpretability—an essential aspect for process design applications.

Regarding performance assessment, a benchmark validation was performed using an independent set of entries of the dataset not included in the training phase, verifying the algorithm's ability to correctly classify electrophile/nucleophile categories and to assign a

realistic kinetic constant range. More extensive validation metrics (e.g., cross-validation or error quantification) could further strengthen the approach and will be considered in future developments.

Solvent and temperature effects exclusion was a deliberate dimensionality-reduction strategy aimed at isolating intrinsic electrophile/nucleophile reactivity as a first-order approximation. We recognize that solvent polarity and temperature can significantly influence polar reaction kinetics; however, incorporating these variables at this stage would have introduced sparsity and multicollinearity issues in the dataset. Future extensions of the T-ReX framework will explicitly include solvent descriptors and temperature dependence to enhance chemical realism and broaden applicability.

The choice of pK_A as shareable classification parameter, both for electrophile/nucleophile have been found useful in better sorting among the different classes of electrophile/nucleophile. Since the pK_A values were not present in almost all the references used; a value has been assigned according to the nature of the substituent (resonance/mesomeric effect) directly bonded to the methylene group of functional groups (e.g., aldehyde, Michael acceptor, etc.), referring to the so-called α-hydrogen. The pK_A values assigned in this way potentially cover a very broad range of molecules. This means that we do not have to limit ourselves to a specific evaluation in the screening phase, but rather speed up the evaluation process so as to make it support a subsequent refinement phase where necessary.

In the abscissa, the merged scale values for E and N parameters ($-30 \leq \mathrm{E/N} \leq +30$) were reported, while in the Y-Axis the pK_A scale values for electrophiles and nucleophiles.

The benchmark validation was done using an independent set of literature values not used for training in order to evaluate the goodness of the classification. This approach highlights how it is possible to assign a molecule to the same category as the ones used in training (e.g., aldehyde, Michael acceptor, etc.), considering a handful of values available. The results confirmed the success of the implementation of this method.

Regular Expression

The kinetic data extraction was possible thanks to a synergistic approach between the *k*-Nearest Neighbor algorithm and the regular expression. The supervised k-NN algorithm showed its ability to classify several different classes of molecules belonging to the two categories of electrophile/nucleophile. The regular expression was able to read the molecules expressed in SMILE notation (not present in the trained database) and to classify them as electrophile or nucleophile. In this way, it was possible to correlate the new molecule(s) to a set of E/N pairs having a certain kinetic constant value.

To better detail, the regular expression algorithm was able to read a certain portion of text and recognize that pattern among other ones: a certain portion of a molecule (e.g., the carbonyl group or the phosphonium ylides group) is literally read by the algorithm. This approach has been addressed by the conversion of molecular formula into SMILE notation [8]. The benchmark process has also been addressed in this way using an already known molecule, confirming the goodness of the algorithm approach.

Once the command-line -egrep has been invoked, the regular expression algorithm was able to read the molecule(s) in SMILE notation, recognizing and highlighting where that pattern (in terms of string) was already present in the list of SMILE molecules format. Hence, it was possible to assign the new molecule to a reduced list of possible candidate classes of molecules, having that specific recursive pattern. The forecasting step takes place at this point: the new molecules expressed in SMILE notation are classified as potential candidates belonging to the binomial category of electrophile/nucleophile—this allows us to foresee which possible E/N combinations are possible and what will be the related kinetic constant.

About these aspects, some considerations are necessary in order to better understand how the regular expression approach has been conceived and applied.

(a) Electrophile.

About the electrophiles, benzaldehyde has been chosen as a benchmark example in order to evaluate the goodness of the algorithm implementation, exploiting the regular expression approach. As it is possible to appreciate in the following bullet list, the benzaldehyde molecule has been translated into SMILE notation, and the original pattern (=O)=C has been converted into the string format by invoking the -egrep command. The algorithm was able to recognize the specific pattern declared initially, and it was able to identify all the molecules having that certain pattern, including the aldehyde molecule.

- smile notation \texttt{O=Cc1ccccc1}–Benzaldehyde
- original_pattern = r″(=O)=C″
- original_pattern = r″[\\(]=O[\\)]=C″} -egrep
- O=Cc1ccccc1}

(b) Nucleophile

About the nucleophile, triethyl phosphonoacetate has been chosen as a benchmark example in order to evaluate the goodness of the algorithm implementation, exploiting the regular expression approach. As it is possible to appreciate in the following bullet list, the Phosphoryl-Stabilized Carbanion molecule has been translated into SMILE notation, and the original pattern (=O)C has been converted into the string format by the -egrep command. The merged approach was able to recognize the specific pattern declared initially and identify all the molecules having that certain pattern, including the phosphonoacetate molecule.

- smile notation for \texttt{CCOC(=O)CP(=O)(OCC)OCC}
- original_pattern = r″CCOC(=O)CP(=O)(OCC)OCC″
- original_pattern = r″[\\(]=O[\\)]=C″} - \texttt{egrep}
- CCOC (=O)CP (=O)(OCC)OCC

(c) Supervised algorithm implementation

It is important to highlight how the pattern recognition toward the regular expression was able to recognize all the molecules having a certain pattern within their own structure in SMILES notation. This leads to possible side-effects since, as listed below, certain molecules could in principle have the same pattern but cannot be considered as electrophile/nucleophile for the purpose of APIs synthesis (e.g., CO_2). Moreover, another aspect that must be taken into account is the way in which these regular expressions read the molecule. The molecules translated into SMILE notation must have the same encoding structure to be read in a unique way by the regular expression algorithm. A SMILE database must be built following unambiguous and commonly accepted encoding rules [8,10,11]. On the contrary, the application of different encoder tools leads to a miscellaneous approach and failure in terms of recognition of characters reported in the two different translated SMILE (e.g., acrylophenone, benzaldehyde (subst.), ethyl acetoacetate vs. triethyl phosphonoacetate).

A general approach for SMILE encoding should be addressed within the scientific community. Listed below are some issues that could take place and that suggest a better understanding of the reason why the *k*-NN algorithm is defined as belonging to supervised learning.

- O=C=O as CO_2 not as carbonyl compounds
- CCOC(=O)CP(=O)(OCC)OCC} triethyl phosphonoacetate

- CCOC(=O)CP(=O)(OCC)OCC} triethyl phosphonoacetate vs. C=CC (=O)c1ccccc1 for acrylophenone (E)
- O=Cc1cccc(Cl)c1 for subst. benzaldehyde (E)
- CCOC(=O)CC(C)=O for ethyl acetoacetate (N)

As appreciated, the application of a supervised algorithm is useful in this case, since it gives the possibility to consider or rule out a set of molecules that cannot belong to the electrophile/nucleophile class for the scope aimed. In this way, machine learning becomes an important tool that helps the chemist in reducing a half-thousand kinetic constant values dataset, supporting the final decision on which electrophile/nucleophile couple of molecules are the best ones to use and test in the laboratory for further experimental validation of kinetic studies, or directly implement that specific kinetic constant in process simulation software as first approximation value.

Moreover, an example is shown in Figure 7, concerning the application of regular expressions in order to differentiate some functional groups (F.G.) among different classes of molecules.

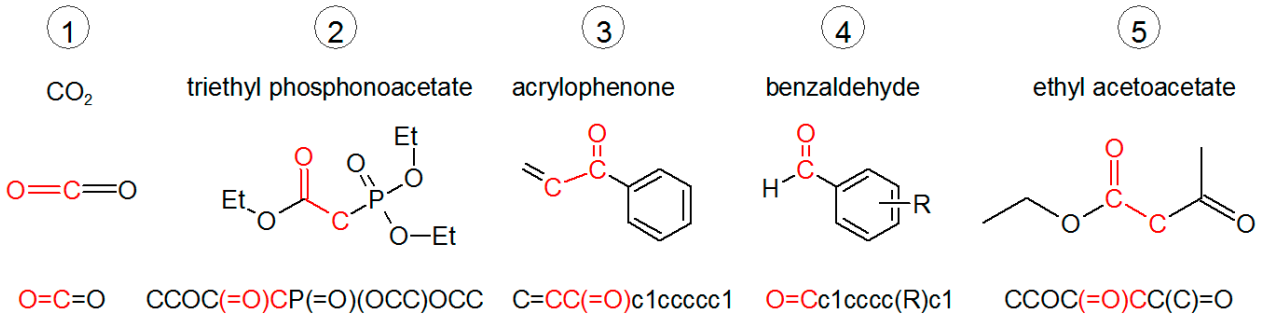


Figure 7. Examples of SMILES encoding and functional group recognition via regular expressions. The figure illustrates how carbonyl and methylene motifs are identified through case-sensitive pattern matching, enabling classification of electrophile and nucleophile candidates. The approach supports automated assignment of new molecules to chemically coherent classes prior to *k*-NN-based kinetic constant estimation.

In the example proposed, it was possible to differentiate the carbonyl F.G. The carbonyl compound belonging to carbon dioxide is the same in terms of SMILE notation of substituted benzaldehyde. Furthermore, the outcome must be checked in order to rule out false positive results.

Another remark is related to the comparison of triethyl phosphonoacetate and the acrylophenone: in principle, the carbonyl F.G. is bonded with a methylene carbon atom, and no difference would be recognizable from the organic chemistry point of view. Actually, thanks to the application of regular expression and, in particular, invoking the -egrep command-line, it is possible to differentiate these two carbonyl compounds so that also the molecules themselves can be differentiated: the triethyl phosphonoacetate works as nucleophile at the methylene carbon (i.e., in Horne-Hemmons reaction as nucleophile source), whereas the acrylophenone works as electrophile (i.e., Michael acceptor). Moreover, the regular expression -egrep command line is capable of differentiating uppercase from lowercase: (=O)C vs. (=O)c (i.e., case sensitive) as a great advantage.

Analogously to the previous point, another aspect is the ability of the -egrep command line to recognize the parentheses present in the pattern searched (here not reported): (=O)=C vs. O=C (i.e., case sensitive).

A last point is related to the way in which the SMILE notation is read. The correct way in which the molecule is encoded and the way in which the pattern is declared in the programming script must be the same. C(=O) is different by definition from (=O)C due

to the order of pattern declared, as well as C(=O)C in ethyl acetoacetate, highlighting the methylene carbon instead of C(=O)C since this last one would define just the carbonyl group (avoiding the methylene).

The attention is selectively focused on the carbonyl group due to its main role as F.G. in defining the electrophile compounds and on the methylene carbon for the nucleophile ones. Other classes of compounds would be in principle investigated (e.g., enammine vs. immine) in order to better point out the potential of the application of regular expression to the chemical SMILE notation.

Therefore, the following hints can be summarized.

1. The assigned molecule must be checked and confirmed to effectively belong to that class (and also category), i.e., strictly supervised learning algorithms should be preferred, such as the *k*-NN.
2. The reduced dataset is now constituted of the new molecule and all the other possible corresponding polar partners (i.e., E/N or N/E) with whom the new molecule could, in principle, react. The pattern feature, as an expression of a specific organic F.G. able to describe a certain class of molecules (e.g., imine group, carbonyl group, etc.), allows for the identification in a unique way of the class to which the new molecule belongs and links itself to all the corresponding polar partner class of molecules. This mutual relationship allows us to identify all the possible combinations available.

The most important result is that for each combination in terms of electrophile/nucleophile reagent couple among all the possible ones, the kinetic constant is made available through the application of a k-NN machine learning technique, trained with the T-ReX database: i.e., by correlating a new molecule labeled as SMILE pattern to an estimated kinetic constant value. This means that a subset of kinetic constant parameters has been found as possible values for that certain specific combination of polar partners and the corresponding reaction. This approach, though time-consuming in the definition of homogeneous labeled datasets from literature available information, is promising to build first estimates of kinetic constants for preliminary process design and feasibility assessment. Even considering a large approximation in kinetics and the relative design, this can anyway save huge experimental efforts and the relative investment, ruling out ineffective routes and leaving the experimental validation needs only to the most promising alternatives after a preliminary screening.

Overall, SMILES representations are not unique and may vary depending on encoding. Nevertheless, in this workflow, molecules were standardized prior to pattern matching using a consistent canonical SMILES generation procedure to reduce ambiguity. The regular expression step was not intended to replace graph-based cheminformatics methods, but to provide a lightweight, interpretable pre-classification filter within a supervised framework. More robust graph-based approaches (e.g., canonicalization and substructure search via cheminformatics toolkits) represent a natural extension of the method.

4. Conclusions

In order to retrieve approximate estimates of kinetic equations and the relative parameters for preliminary process design, different approaches have been applied to literature data. Anticipating a first sight basic economic assessment of fine chemicals and drugs synthesis at the early explorative stage could save experimental efforts and investments, focusing the resources only on the most promising candidate routes. However, to exploit basic process design principles, the preliminary design of main equipment is needed, in turn requiring at least an approximate formulation of the kinetic equation and the relative parameters.

Unfortunately, these entry data are collected only on the most promising reactions after many trial and error experimental efforts, so that rigorously collected kinetic datasets are

scarce in the literature. On the contrary, a huge body of information is publicly available, which reports explorative testing of many reactions. Therefore, it is possible to take advantage of this asset by adapting different modeling approaches to the specific structure of the database.

Three examples have been reported, i.e., linear/non-linear regression, applied to a small, manually sorted dataset that reports the dependence of an observable with respect to different sensitive variables. The case of Oseltamvir synthesis was used for this purpose. Satisfactory results can be achieved for limited datasets provided that the data are unbiased (free from transport limitations and sufficiently sparse) and in sufficiently large amounts.

The results coming from numerical analysis of the oseltamivir dataset for azidation kinetic parameters investigation cannot be considered consistent with a reasonably expected elementary reaction, and the dataset cannot be safely used for further applications, such as the implementation in a process design simulation project. The experimental dataset is not fully suitable for kinetic investigation (conversion values are still too high, as well as almost all temperatures). Even by manually selecting subsets of data and in spite of a rather homogenous origin of the raw data, the kinetic parameters were, in most cases, unreliable, and their physical meaning was not always supported.

It would be essential to repeat the synthesis steps with a kinetic testing approach, including the exploration of the whole variables space and performing the check for intrinsic kinetic regime (excluding the presence of diffusional limitations). Noteworthy would be to apply the so-called High Throughput Experimentation (HTE) [10,11,58] to the reaction of interest and to collect, evaluate, and store all the parameter data in a structured database.

Considering the Ibuprofen case study with the application of Bayesian statistical tools, the results obtained were also not coherent with chemical kinetic expectations. However, also from a mathematical point of view, the mere values of 1D-marginal posterior density were not acceptable (Figure 5) since the numerical values were really low ($\leq 10^{-2}$). The poor dataset, both in terms of quality and quantity of data, suggests increasing the dataset entries through a new experimental data collection suitable both for optimization of organic synthesis (high yield) and kinetic investigation (whole variables space), so as to improve the sampling procedure by Markov Chain-Monte Carlo, too. This last aspect could take into consideration a reconsideration of the computational procedure adopted, exploiting different approaches as well as comparing other algorithms in order to improve the robustness of data analysis.

Concerning the Pregabalin and Rolipram case and the reactions considered, two approaches have been exploited. About the application of PCA, it was not able to suggest useful information about the influence of a possible correlation among the parameters reported in Mayr's equation, with the exception of some samples that were influenced by σ_N and E, all the other samples were not affected by the other variables. This last part is well explained by the biplot diagram, where the samples are all enclosed in a restricted space (see Figure 6) and almost all of them were not successfully influenced by the variables (i.e., s_N, N and E).

On the contrary, the application of a cutting-edge machine learning algorithm returned satisfactory results in terms of forecasting the kinetic constant values range. This approach outlines how the kinetic constant—as well as the other parameters considered, i.e., s_N N and E should be considered as derivatives—no direct determination of parameters, avoiding the potential risk of a recursive relationship already built.

The forecasting ability to assign a certain new molecule to a reduced potential candidate set from the wider one leads the process to be considered reliable both in terms of quality and quantity: the kinetic constant_ range is important in the first instance to be applied in the subsequent process development stage.

The T-ReX in-house developed database, based on the original Mayr's one, has given the possibility to explore more than 500 kinetic constants derived from several electrophile/nucleophile combinations. The *k*-Nearest Neighbor (*k*-NN) algorithm proved to be appropriate for the classification of the molecules, whereas regular expression proved to be a powerful technique in reading a molecule. A unified approach of regular expression with a pattern classification system allowed for forecasting the kinetic constant. The prediction of kinetic constants between an E/N couple became possible, limited only by the database used for training. Other approaches already used for different purposes (i.e., retrosynthetic organic compounds pathway) can be investigated and applied to this context, such as the methodology based on the graph theory (GCNNs) [58,59] for searching for more robust approaches. Further evolution can be the application of artificial neural networks (ANNs) [60,61], though the simplest *k*-Nearest Neighbor already demonstrated satisfactory results. Our intent was not to rediscover Mayr relationships, but to demonstrate how structured reactivity descriptors can be operationalized within a machine learning workflow for rapid kinetic range estimation in process design contexts. The predictive step concerns classification and interpolation within a chemically coherent but non-exhaustive subset, rather than analytical recovery of the Mayr equation itself.

An appropriate improvement in the field of data mining would be a cutting-edge step forward in this context [10,11].

At last, this kinetic data handling management could be considered as the starting point for further kinetic investigation concerning the more exotic organic synthesis reaction variation (i.e., Mukayama aldol reaction, asymmetric Michael addition, organo-catalysis approach [62], etc.), leading to an improvement in terms of scale-up design for the synthesis and production of other important APIs.

In conclusion, this study demonstrates that the main bottleneck in data-driven kinetic modeling for pharmaceutical process design is not the lack of computational tools, but the structural inadequacy of typical literature-derived datasets. Linear and non-linear regression, as well as Bayesian inference, reveal significant limitations when applied to high-conversion, non-kinetic-oriented experimental data, highlighting issues of ill-conditioning and parameter non-identifiability. Rather than representing methodological failure, these outcomes provide quantitative evidence of the mismatch between synthetic optimization data and process design requirements. In contrast, the structured T-ReX database and its integration with machine learning techniques illustrate a viable pathway toward transforming dispersed chemical knowledge into actionable kinetic estimates. The comparative analysis presented here reframes data-driven kinetics as a problem of data engineering and epistemic alignment, emphasizing that reliable process-relevant modeling must begin with appropriately structured experimental information. In this sense, the contribution of the present work is methodological and conceptual: it clarifies the boundaries of current approaches while outlining directions for more robust integration between organic synthesis and process systems engineering, which should merge as early as possible to rule out unfeasible routes and speed up drug development.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/chemistry8030032/s1. References [16,21,22,28–31,63–70] are cited in the supplementary materials.

Author Contributions: Conceptualization, A.R. and I.R.; methodology, A.R.; software, A.R.; validation, G.R.; investigation, A.R.; resources, G.R. and I.R.; writing—original draft preparation, A.R.; writing—review and editing, G.R. and I.R.; supervision, I.R.; project administration, G.R. and I.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The valuable help of Ing. Antonio Tripodi and all people who have supported the project with the translation of the ideas into computational development in terms of algorithm are gratefully acknowledged.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

APIs	Active Pharmaceutical Ingredients
SVD	singular value decomposition
GNA	Gauss-Newton algorithm
LMA	Levenberg–Marquardt
MCMC	Markov Chain Monte Carlo
PCA	Principal component analysis
k-NN	*k*-nearest neighbors
SVM	supporting vector machine (SVM)
SMILE	Simplified Molecular Input Line Entry System
ANNs	Artificial Neural Networks
GCNNs	Graph Convolutional Neural Networks

References

1. Rossetti, I. Continuous flow (micro-)reactors for heterogeneously catalyzed reactions: Main design and modelling issues. *Catal. Today* **2018**, *308*, 20–31. [CrossRef]
2. Rossetti, I. Modelling of continuous reactors for flow chemistry. *Chim. Oggi/Chem. Today* **2017**, *35*, 8–11.
3. Rossetti, I.; Compagnoni, M. Chemical reaction engineering, process design and scale-up issues at the frontier of synthesis: Flow chemistry. *Chem. Eng. J.* **2016**, *296*, 56–70. [CrossRef]
4. Adamo, A.; Beingessner, R.L.; Behnam, M.; Chen, J.; Jamison, T.F.; Jensen, K.F.; Monbaliu, J.-C.M.; Myerson, A.S.; Revalor, E.M.; Snead, D.R.; et al. On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system. *Science* **2016**, *352*, 61–67. [CrossRef]
5. Jolliffe, H.G.; Gerogiorgis, D.I. Technoeconomic Optimization of a Conceptual Flowsheet for Continuous Separation of an Analgaesic Active Pharmaceutical Ingredient (API). *Ind. Eng. Chem. Res.* **2017**, *56*, 4357–4376. [CrossRef]
6. Nielsen, M.K.; Ahneman, D.T.; Riera, O.; Doyle, A.G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004–5008. [CrossRef] [PubMed]
7. Williams, W.L.; Zeng, L.; Gensch, T.; Sigman, M.S.; Doyle, A.G.; Anslyn, E.V. The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Cent. Sci.* **2021**, *7*, 1622–1637. [CrossRef]
8. Coley, C.W.; Green, W.H.; Jensen, K.F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281–1289. [CrossRef]
9. Shields, B.J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J.I.M.; Janey, J.M.; Adams, R.P.; Doyle, A.G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89–96. [CrossRef]
10. Coley, C.W.; Eyke, N.S.; Jensen, K.F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem. Int. Ed.* **2020**, *59*, 23414–23436. [CrossRef]
11. Coley, C.W.; Eyke, N.S.; Jensen, K.F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem. Int. Ed.* **2020**, *59*, 22858–22893. [CrossRef] [PubMed]
12. Komp, E.; Janulaitis, N.; Valleau, S. Progress towards machine learning reaction rate constants. *Phys. Chem. Chem. Phys.* **2022**, *24*, 2692–2705. [CrossRef] [PubMed]
13. Komp, E.; Valleau, S. Addition to "Machine Learning Quantum Reaction Rate Constants.". *J. Phys. Chem. A* **2021**, *125*, 9259–9260. [CrossRef] [PubMed]
14. Baltussen, M.G.; van de Wiel, J.; Fernández Regueiro, C.L.; Jakštaitė, M.; Huck, W.T.S. A Bayesian Approach to Extracting Kinetic Information from Artificial Enzymatic Networks. *Anal. Chem.* **2022**, *94*, 7311–7318. [CrossRef]
15. Walker, E.A.; Ravisankar, K. Bayesian Design of Experiments: Implementation, Validation and Application to Chemical Kinetics. *arXiv* **2019**, arXiv:1909.03861. [CrossRef]

16. Sagandira, C.R.; Watts, P. Continuous-Flow Synthesis of (–)-Oseltamivir Phosphate (Tamiflu). *Synlett* **2020**, *31*, 1925–1929. [CrossRef]
17. Kobayashi, S. Flow "Fine" Synthesis: High Yielding and Selective Organic Synthesis by Flow Methods. *Chem.-Asian J.* **2016**, *11*, 425–436. [CrossRef]
18. Tsubogo, T.; Oyamada, H.; Kobayashi, S. Multistep continuous-flow synthesis of (R)- and (S)-rolipram using heterogeneous catalysts. *Nature* **2015**, *520*, 329–332. [CrossRef]
19. Ishitani, H.; Kanai, K.; Saito, Y.; Tsubogo, T.; Kobayashi, S. Synthesis of (±)-Pregabalin by Utilizing a Three-Step Sequential-Flow System with Heterogeneous Catalysts. *Eur. J. Org. Chem.* **2017**, *2017*, 6491–6494. [CrossRef]
20. Sagandira, C.R.; Watts, P. A study on the scale-up of acyl azide synthesis in various continuous flow reactors in homogeneous and biphasic systems. *J. Flow Chem.* **2018**, *8*, 69–79. [CrossRef]
21. Sagandira, C.R.; Watts, P. Safe and highly efficient adaptation of potentially explosive azide chemistry involved in the synthesis of Tamiflu using continuous-flow technology. *Beilstein J. Org. Chem.* **2019**, *15*, 2577–2589. [CrossRef] [PubMed]
22. Bogdan, A.R.; Poe, S.L.; Kubis, D.C.; Broadwater, S.J.; McQuade, D.T. The Continuous-Flow Synthesis of Ibuprofen. *Angew. Chem. Int. Ed.* **2009**, *48*, 8547–8550; 8699–8702. [CrossRef] [PubMed]
23. Snead, D.R.; Jamison, T.F. A Three-Minute Synthesis and Purification of Ibuprofen: Pushing the Limits of Continuous-Flow Processing. *Angew. Chem. Int. Ed.* **2015**, *54*, 983–987. [CrossRef] [PubMed]
24. Ghislieri, D.; Gilmore, K.; Seeberger, P.H. Chemical Assembly Systems: Layered Control for Divergent, Continuous, Multistep Syntheses of Active Pharmaceutical Ingredients. *Angew. Chem. Int. Ed.* **2015**, *54*, 678–682. [CrossRef]
25. Available online: https://www.cup.lmu.de/oc/mayr/reaktionsdatenbank/ (accessed on 1 December 2024).
26. Available online: https://automeris.io/WebPlotDigitizer.html (accessed on 1 December 2024).
27. Fogler, H.S. *Elements of Chemical Reaction Engineering*, 5th ed.; Pearson Education Inc.: Hoboken, NJ, USA, 2016.
28. Perrin, C.L. Linear or Nonlinear Least-Squares Analysis of Kinetic Data? *J. Chem. Educ.* **2017**, *94*, 669–672. [CrossRef]
29. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*, 5th ed.; Wiley: Hoboken, NJ, USA, 2012.
30. Beers, K.J. *Numerical Methods for Chemical Engineering: Applications in MATLAB*; Cambridge University Press: Cambridge, UK, 2006.
31. Bolstad, W.M.; Curran, J.M. *Introduction to Bayesian Statistics*; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2016.
32. McCluskey, A.R. Is There Still a Place for Linearization in the Chemistry Curriculum? *J. Chem. Educ.* **2023**, *100*, 4174–4176. [CrossRef]
33. Froment, G.F.; Bischoff, K.B.; De Wilde, J. *Chemical Reactor Analysis and Design*, 3rd ed.; Wiley: Hoboken, NJ, USA, 2011.
34. Wei, J. Adsorption and cracking of n-alkanes over ZSM-5: Negative activation energy of reaction. *Chem. Eng. Sci.* **1996**, *51*, 2995–2999. [CrossRef]
35. Revell, L.E.; Williamson, B.E. Why Are Some Reactions Slower at Higher Temperatures? *J. Chem. Educ.* **2013**, *90*, 1024–1027. [CrossRef]
36. Ge, Y. Agreement, Complement, and Disagreement to "Why Are Some Reactions Slower at Higher Temperatures?". *J. Chem. Educ.* **2017**, *94*, 821–823. [CrossRef]
37. Jolliffe, H.G.; Gerogiorgis, D.I. Process modelling and simulation for continuous pharmaceutical manufacturing of ibuprofen. *Chem. Eng. Res. Des.* **2015**, *97*, 175–191. [CrossRef]
38. Jolliffe, H.G.; Gerogiorgis, D.I. Plantwide design and economic evaluation of two Continuous Pharmaceutical Manufacturing (CPM) cases: Ibuprofen and artemisinin. *Comput. Chem. Eng.* **2016**, *91*, 269–288. [CrossRef]
39. Tripodi, A.; Martinazzo, R.; Ramis, G.; Rossetti, I. Process Modeling Issues in the Design of a Continuous-Flow Process for the Production of Ibuprofen. *Chem. Eng. Technol.* **2020**, *43*, 2557–2566. [CrossRef]
40. Kearnes, S.M.; Maser, M.R.; Wleklinski, M.; Kast, A.; Doyle, A.G.; Dreher, S.D.; Hawkins, J.M.; Jensen, K.F.; Coley, C.W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143*, 18820–18826. [CrossRef] [PubMed]
41. Mayr, H.; Patz, M. Scales of Nucleophilicity and Electrophilicity: A System for Ordering Polar Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **1994**, *33*, 938–957. [CrossRef]
42. Mayr, H.; Kempf, B.; Ofial, A.R. π-Nucleophilicity in Carbon−Carbon Bond-Forming Reactions. *Acc. Chem. Res.* **2003**, *36*, 66–77. [CrossRef]
43. Mayr, H.; Ofial, A.R. Kinetics of electrophile-nucleophile combinations: A general approach to polar organic reactivity. *Pure Appl. Chem.* **2005**, *77*, 1807–1821. [CrossRef]
44. Mayr, H.; Ofial, A.R. Do general nucleophilicity scales exist? *J. Phys. Org. Chem.* **2008**, *21*, 584–595. [CrossRef]
45. Appel, R.; Mayr, H. Quantification of the Electrophilic Reactivities of Aldehydes, Imines, and Enones. *J. Am. Chem. Soc.* **2011**, *133*, 8240–8251. [CrossRef]
46. Appel, R.; Loos, R.; Mayr, H. Nucleophilicity Parameters for Phosphoryl-Stabilized Carbanions and Phosphorus Ylides: Implications for Wittig and Related Olefination Reactions. *J. Am. Chem. Soc.* **2009**, *131*, 704–714. [CrossRef]

47. Byrne, P.A.; Karaghiosoff, K.; Mayr, H. Ambident Reactivity of Acetyl- and Formyl-Stabilized Phosphonium Ylides. *J. Am. Chem. Soc.* **2016**, *138*, 11272–11281. [CrossRef]
48. Bug, T.; Lemek, T.; Mayr, H. Nucleophilicities of Nitroalkyl Anions. *J. Org. Chem.* **2004**, *69*, 7565–7576. [CrossRef]
49. Phan, T.B.; Mayr, H. Nucleophilicity Parameters for Carbanions in Methanol. *Eur. J. Org. Chem.* **2006**, *2006*, 2530–2537. [CrossRef]
50. Bug, T.; Mayr, H. Nucleophilic Reactivities of Carbanions in Water: The Unique Behavior of the Malodinitrile Anion. *J. Am. Chem. Soc.* **2003**, *125*, 12980–12986. [CrossRef] [PubMed]
51. Lucius, R.; Loos, R.; Mayr, H. Kinetic Studies of Carbocation-Carbanion Combinations: Key to a General Concept of Polar Organic Reactivity. *Angew. Chem. Int. Ed.* **2002**, *41*, 91–95. [CrossRef]
52. Corral-Bautista, F.; Appel, R.; Frickel, J.S.; Mayr, H. Quantification of Ion-Pairing Effects on the Nucleophilic Reactivities of Benzoyl- and Phenyl-Substituted Carbanions in Dimethylsulfoxide. *Chem.-A Eur. J.* **2015**, *21*, 875–884. [CrossRef]
53. Puente, Á.; He, S.; Corral-Bautista, F.; Ofial, A.R.; Mayr, H. Nucleophilic Reactivities of 2-Substituted Malonates. *Eur. J. Org. Chem.* **2016**, *2016*, 1841–1848. [CrossRef]
54. Puente, Á.; Ofial, A.R.; Mayr, H. Nucleophilic Reactivities of Bis-Acceptor-Substituted Benzyl Anions. *Eur. J. Org. Chem.* **2017**, *2017*, 1196–1202. [CrossRef]
55. Zenz, I.; Mayr, H. Electrophilicities of trans-β-Nitrostyrenes. *J. Org. Chem.* **2011**, *76*, 9370–9378. [CrossRef]
56. Allgäuer, D.S.; Jangra, H.; Asahara, H.; Li, Z.; Chen, Q.; Zipse, H.; Ofial, A.R.; Mayr, H. Quantification and Theoretical Analysis of the Electrophilicities of Michael Acceptors. *J. Am. Chem. Soc.* **2017**, *139*, 13318–13329. [CrossRef]
57. Carey, F.A.; Sundberg, R.J. *Advanced Organic Chemistry—Part A Structure and Mechanisms*; Springer Nature: London, UK, 2008.
58. Mennen, S.M.; Alhambra, C.; Allen, C.L.; Barberis, M.; Berritt, S.; Brandt, T.A.; Campbell, A.D.; Castañón, J.; Cherney, A.H.; Christensen, M.; et al. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.* **2019**, *23*, 1213–1242. [CrossRef]
59. Coley, C.W.; Jin, W.; Rogers, L.; Jamison, T.F.; Jaakkola, T.S.; Green, W.H.; Barzilay, R.; Jensen, K.F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377. [CrossRef]
60. Galván, I.M.; Zaldívar, J.M.; Hernández, H.; Molga, E. The use of neural networks for fitting complex kinetic data. *Comput. Chem. Eng.* **1996**, *20*, 1451–1465. [CrossRef]
61. Kovács, B.; Tóth, J. Estimating Reaction Rate Constants with Neural Networks. *World Acad. Sci. Eng. Technol.* **2007**, *26*, 13–17.
62. Porta, R.; Benaglia, M.; Coccia, F.; Rossi, S.; Puglisi, A. Enantioselective Organocatalysis in Microreactors: Continuous Flow Synthesis of a (S)-Pregabalin Precursor and (S)-Warfarin. *Symmetry* **2015**, *7*, 1395–1409. [CrossRef]
63. Bates, D.M.; Watts, D.G. *Nonlinear Regression Analysis and Its Applications*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 1988.
64. Gavin, H.P. The Levenberg-Marquardt Algorithm for Nonlinear Least Squares Curve-Fitting Problems. Available online: https://people.duke.edu/ hpgavin/ExperimentalSystems/lm.pdf (accessed on 1 December 2025).
65. Robert, C.P. *The Bayesian Choice from Decision-Theoretic Foundations to Computational Implementation*; Springer Texts in Statistics; Springer: New York, NY, USA, 2001.
66. Chen, M.-H.; Shao, Q.-M.; Ibrahim, J.G. *Monte Carlo Methods in Bayesian Computation*; Springer Series in Statistic; Springer: New York, NY, USA, 2000.
67. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. *Markov Chain Monte Carlo in Practice*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1996.
68. MacKayy, D. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
69. Raschka, S. Lecture Notes in STAT 479: Machine Learning, Department of Statistics University of Wisconsin Madison. Available online: https://github.com/rasbt/stat479-machine-learning-fs18/tree/master/02_knn (accessed on 1 December 2025).
70. Bogdan, A.R.; Poe, S.L.; Kubis, D.C.; Broadwater, S.J.; McQuade, D.T. The Continuous-Flow Synthesis of Ibuprofen. *Angew. Chem.* **2009**, *121*, 8699–8702. [CrossRef]