

This document is the **Accepted Manuscript version** of a Published Work that appeared in final form in **Journal of Chemical Information and Modeling** **2018 58 (7), 1384-1396**, copyright © 2018, American Chemical Society after peer review and technical editing by the publisher. To access the final edited and published work see <https://doi.org/10.1021/acs.jcim.8b00286>

**Perturbation-Theory and Machine Learning (PTML). Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies**

*Lorena Simón-Vidal, Oihane García-Calvo, Uxue Oteo, Sonia Arrasate, Esther Lete, Nuria Sotomayor, and Humberto González-Díaz*

*Journal of Chemical Information and Modeling 2018 58 (7), 1384-1396*

**DOI: 10.1021/acs.jcim.8b00286**

# PTML: Perturbation-Theory and Machine Learning Model for High-Throughput Screening of Parham Reactions. Experimental and Theoretical Studies

Lorena Simón-Vidal,<sup>a</sup> Oihane García-Calvo,<sup>a</sup> Uxue Oteo,<sup>a</sup> Sonia Arrasate,<sup>a</sup> Esther Lete,<sup>a</sup> Nuria Sotomayor,<sup>a,\*</sup> and Humberto González-Díaz<sup>a,b,\*</sup>

<sup>a</sup>Departamento de Química Orgánica II, Facultad de Ciencia y Tecnología, Universidad del País Vasco / Euskal Herriko Unibertsitatea UPV/EHU. Apdo. 644. 48080 Bilbao (Spain);

<sup>b</sup>IKERBASQUE, Basque Foundation for Science, 48011, Bilbao (Spain)

ABSTRACT. Machine Learning (ML) algorithms are gaining importance in the processing of chemical information and modelling of chemical reactivity problems. In this work, we have developed a PTML model combining Perturbation-Theory (PT) and ML algorithms for predicting the yield of a given reaction. For this purpose, we have selected Parham cyclization, which is a general and powerful tool for the synthesis of heterocyclic and carbocyclic compounds. This reaction has both structural (substitution pattern on the substrate, internal electrophile, ring size, *etc.*) and operational variables (organolithium reagent, solvent, temperature, time, *etc.*), so predicting the effect of changes on substrate design (internal electrophile, halide, *etc.*) or reaction conditions on the yield is an important task that could help to optimize the reaction design. The PTML model developed uses PT operators to account for

1  
2  
3 perturbations in experimental conditions and/or structural variables of all the molecules involved  
4  
5 in a query reaction compared to a reaction of reference. Thus, a dataset of >100 reactions has  
6  
7 been collected for different substrates and internal electrophiles, under different reaction  
8  
9 conditions, with a wide range of yields (0 – 98%). The best PTML model found using General  
10  
11 Linear Regression (GLR) has  $R = 0.88$  in training and  $R = 0.83$  in external validation series for  
12  
13 10000 pairs of query and reference reactions. The PTML model has a final  $R = 0.95$  for all  
14  
15 reactions using multiple reactions of reference. We also report a comparative study of linear vs.  
16  
17 non-linear PTML models based on Artificial Neural Networks (ANN) algorithms. PTML-ANN  
18  
19 models (LNN, MLP, RBF) with  $R \approx 0.1 - 0.8$  do not outperform the first PMTL model. This  
20  
21 result confirms the validity of the linearity of the model. Next, we carried out an experimental  
22  
23 and theoretical study of non-reported Parham reactions to illustrate the practical use of the PTML  
24  
25 model. A 500000-point simulation and a Hammett analysis of the reactivity space of Parham  
26  
27 reactions are also reported.  
28  
29  
30  
31

## 32 33 34 **1. INTRODUCTION**

35  
36  
37 The optimization of chemical reactions is an important goal in organic synthesis towards the  
38  
39 production of new catalysts, drugs, and materials. A common situation in organic chemistry is  
40  
41 the existence of non-optimal reactions with a promising but still low reaction yield under a given  
42  
43 set of experimental conditions. In this context, a large number of reactions formed by  
44  
45 combinations of solvents, additives, catalysts, temperature, time, and other operational variables  
46  
47 have to be studied in order to optimize the synthetic procedure. This reaction space is so vast that  
48  
49 it is very difficult and/or costly to scan with experimental techniques, so computational models  
50  
51 can be employed to predict the yield of a reaction for related substrates. In a recent work,  
52  
53 Marcou<sup>1</sup> highlighted the importance of expert systems for prediction of chemical reactivity in  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 organic synthesis. Thus, computational methods could be efficiently used to establish a  
4 relationship between the reagent structure and the required reaction conditions. This would allow  
5 synthetic chemists to use less time and resources in the optimization of reaction conditions for a  
6 given transformation.<sup>1</sup> Warr has published an important review on computational approaches to  
7 chemical reactivity.<sup>2</sup> In addition, Sigman has demonstrated that computational chemistry models  
8 relying upon calculation of molecular descriptors are useful in organic synthesis.<sup>3-10</sup> However,  
9 most of the known computational methods do not use at the same time the information about the  
10 new reaction and the reaction of reference. Therefore, computational modeling of reactivity is  
11 still a major challenge.  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24

25 In this context, Machine Learning (ML) methods may play an important role for the prediction  
26 of physicochemical properties of organic compounds.<sup>11-20</sup> ML methods have been used also to  
27 predict chemical reactivity.<sup>21-31</sup> ML methods infer the reactivity of the new or query molecules  
28 ( $m_q$ ) using as input structural variables  $V(m_q)$ , known as molecular descriptors. These input  
29 variables  $V(m_q)$  may be calculated using Quantum Chemistry and/or other methods.<sup>24-26</sup> In fact,  
30 Skoraczynski *et al.* discussed very recently the necessity of new classes of descriptors for the  
31 prediction of chemical reactivity.<sup>27</sup> ML methods generally involve three steps. The first step is  
32 the compilation of a dataset of reactions with known values of output variable and conditions of  
33 reaction [ $V(c_q)$ ]. The yield of reaction [ $Yld(\%)$ ] is probably the most common output variable  
34 used in ML predictive studies, but not the only one. The second step is the calculation of the  
35 molecular descriptors [ $V(m_q)$ ] of all the molecules involved. Sometimes, ML methods consider  
36 also the numerical values [ $V(c_q)$ ] of the experimental conditions of query reaction [ $c_q$ ]  
37 (temperature, time, solvent, additives, *etc.*) as input variables. The third step is the use of a ML  
38 method to fit a quantitative relationship between the output variable [ $Yld(\%)$ ] and the input  
39 variables [ $V(m_q)$  and  $V(c_q)$ ]. Finally, the model can be used to predict the values of the output  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

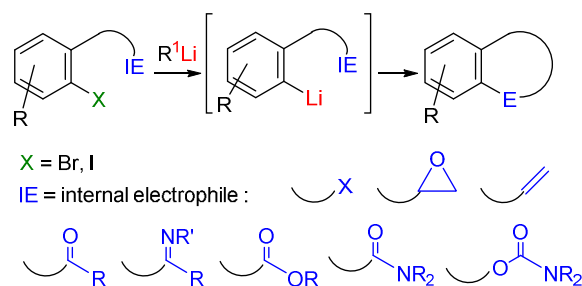
variable for new sets of reactants/products and/or reaction conditions. The output of a ML model is the predicted value of yield [ $\text{Yld}(\%)_{\text{pred}}$ ], not to be confused with the observed value of yield for a query reaction [ $\text{Yld}(\%)_{\text{query}}$ ]. In fact,  $\text{Yld}(\%)_{\text{pred}}$  is an estimation of  $\text{Yld}(\%)_{\text{query}}$  and the difference between these values  $\text{Yld}(\%)_{\text{res}} = \text{Yld}(\%)_{\text{query}} - \text{Yld}(\%)_{\text{pred}}$  is known as the residual value. In equation (1), an example of linear ML additive model is shown.

$$\text{Yld}(\%)_{\text{pred}} = \sum_{q=1}^{\text{qmax}} a_q \cdot V(m_q) + \sum_{r=1}^{\text{rmax}} b_r \cdot V(c_r) + e_0 \quad (1)$$

As has been mentioned, most of the ML models use the previous knowledge (data set of known reactions) to fit the coefficients of the model. Nevertheless, they do not include specific examples of that previous knowledge as direct input variables of the model.<sup>22-31</sup> Conversely, experimentalists in organic synthesis commonly use the information of known reactions as a starting point to infer the possible result for new but similar reactions. This kind of problem is ideal to be approached with a Perturbation Theory (PT) method. PT methods start with a known solution to a known problem and seek a solution to a new, but similar, problem by adding perturbation terms to the known solution. It means that the yield of a query reaction [ $\text{Yld}(\%)_{\text{query}}$ ] can be inferred beginning with the value of yield of a reaction of reference [ $\text{Yld}(\%)_{\text{ref}}$ ] and adding the effect of structural perturbations and/or perturbations in the experimental conditions. In this situation, the values of the reaction of references [ $\text{Yld}(\%)_{\text{ref}}$ ,  $V(m_q)$  and  $V(c_r)$ ] are used to predict the value of the new product of interest, using a similar reaction with structural variables [ $V(m_q)$ ] and conditions [ $V(c_r)$ ]. When a ML method is used to seek the coefficients of the PT model, this can be regarded as a PTML model. In this context, our group has formulated a general-purpose PTML approach to structure-property problems with perturbations in multiple experimental conditions.<sup>32</sup> These are perturbations involving changes in both the chemical structure of reactants and/or input conditions of reaction (solvent, catalyst, temperature, reaction

1  
2  
3 time, *etc.*). Thus, we have developed PTML models for a very large set of in-out perturbations in  
4 the reaction conditions for intra-molecular carbolithiations.<sup>33</sup> In recent works, we have also  
5 developed new PTML models to predict the enantioselectivity in Heck-Heck cascade reactions<sup>34</sup>  
6 and intermolecular  $\alpha$ -amidoalkylation reactions.<sup>35</sup> In these last examples, the enantiomeric excess  
7 [ee(%)] and not the yield were the output variables.  
8  
9  
10  
11  
12  
13  
14

15 In this work we have selected the Parham reaction,<sup>36</sup> which consists of the cyclization of an  
16 aryllithium intermediate (ArLi) generated by halogen/lithium exchange with internal  
17 electrophiles (IE) to form a cyclic compound (Scheme 1). This is a general reaction, that has  
18 been widely applied for the formation of both carbocycles and heterocycles, and plays a crucial  
19 role in natural product synthesis. Different parameters have to be considered in the design of a  
20 given Parham-type reaction. The structural variables, such as the substitution pattern on the  
21 aromatic ring, the halide atom, the size of the ring formed, or the internal electrophile employed  
22 have an important impact in the reaction outcome. Aryllithiums derived from aromatic or  
23 heteroaromatic precursors (bromides and iodides, X = Br, I) are readily available with variable  
24 substitution patterns, although the effect on the aromatic ring substitution on the reactivity of the  
25 aryllithium has not been clearly established. Lithium – halogen exchange reaction (LHE) is very  
26 fast, so the intermediate aryllithium can be prepared in the presence of different types of internal  
27 electrophiles, which are reactive enough to participate in a subsequent cyclization reaction. Thus,  
28 halides, epoxides, ketones, imines, alkenes alkynes, amides, esters or carbamates have been  
29 efficiently used in this type of reaction.<sup>37</sup>  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



13 **Scheme 1.** General scheme for the Parham reaction

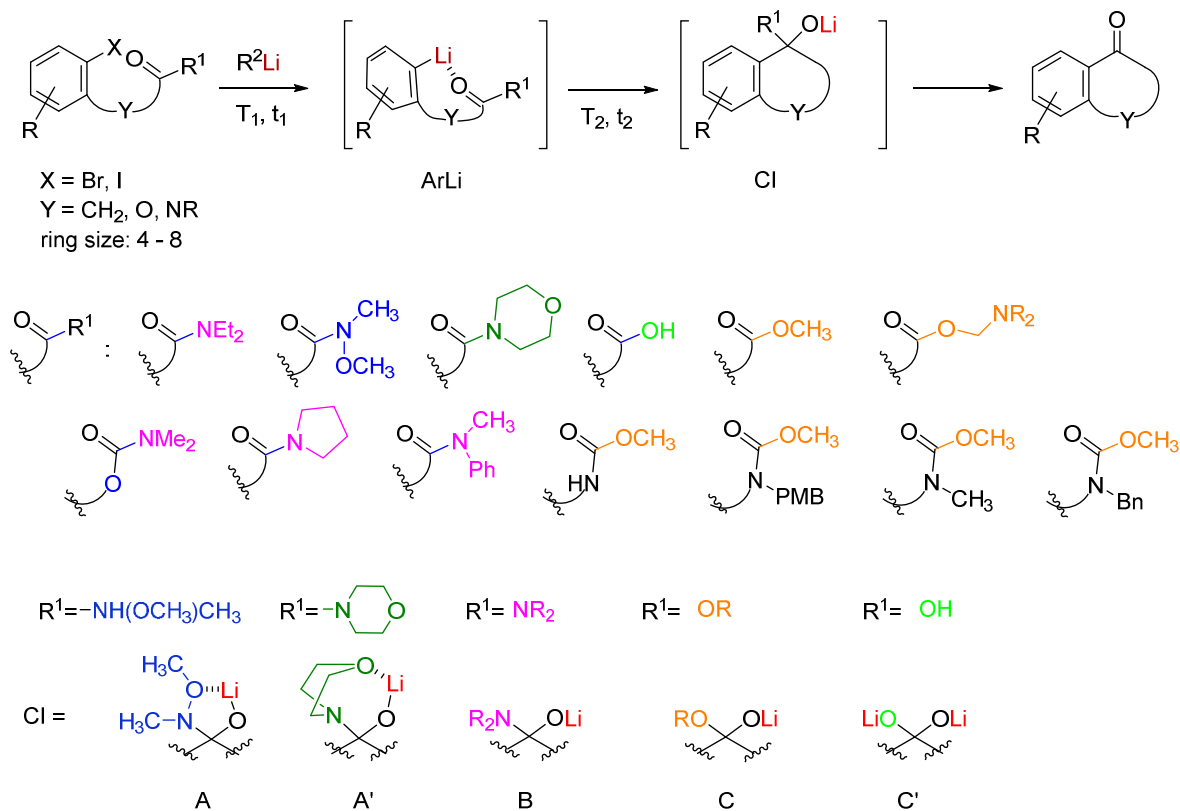
14  
15  
16 When the internal electrophile is a carboxylic acid derivative, the reaction may be regarded as  
17 a carbanionic Friedel-Crafts equivalent, lacking the electronic limitations of the classical  
18 reaction.<sup>37</sup> The use of esters as internal electrophiles in Parham cyclizations could have an  
19 important drawback. In fact, although it is possible to perform a lithium-halogen exchange  
20 reaction, the intermediate generated by acylation of the aryllithium is not stable in the reaction  
21 medium. Thus, RLi addition to the generated carbonyl group affording alcohols is an important  
22 side reaction. This can be avoided using other types of derivatives and, to this end, amides and  
23 carbamates constitute some of the most effectively used internal electrophiles in Parham  
24 cyclizations. On the other hand, it is not clear how the substitution pattern of the aromatic ring or  
25 the size of the ring formed affects the overall process. Apart from the structural parameters, the  
26 operational reaction parameters, such as the temperatures, solvent, organolithium (RLi) used, etc.  
27 have a clear impact on the reaction outcome. Thus, the development of a model that considers all  
28 this variables, both structural and operational, could be helpful for the selection of reaction  
29 conditions, including the most efficient IE, for the synthesis of a given target compound,  
30 reducing the experimental screening.  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49

50 To this end, in this work, we have focused on Parham reactions that employ carboxylic acid  
51 derivatives as internal electrophiles. Thus, the general scheme of the reactions studied involves  
52 the intramolecular cyclization of aryllithium compounds generated by lithium-halogen exchange  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 with different types of carboxylic acid derivatives (amides, esters, carbamates) (Scheme 2),  
4 considering both structural and operational variables. To cover a wide reaction space, examples  
5 that use 13 different classes of carboxylic acid derivatives as internal electrophiles for the  
6 formation of 4 to 8-membered carbo- and heterocycles (Scheme 2) have been selected. Thus,  
7 reaction of the aryl halide (substrate) with an organolithium reagent (RLi) at a given temperature  
8 ( $T_1$ ) for a given time ( $t_1$ ) affords an aryllithium intermediate (ArLi). The temperature may or may  
9 not be changed ( $T_2$ ) for a given time ( $t_2$ ) to afford the cyclized intermediate (CI). The reaction  
10 outcome depends on the nature of the internal electrophile used, considering both the reactivity  
11 (electrophilicity of the carbonyl group) and the stability of the intermediate formed after the  
12 cyclization (CI in Scheme 2). Thus, carboxylic acids and esters would form unstable tetrahedral  
13 intermediates of type C. On the other hand, amides would form stable tetrahedral intermediates  
14 (A or B) in the reaction medium, favoring the overall reaction yield. In the case of using Weinreb  
15 amides (A) or morpholine amides (A'), the presence of an oxygen atom provides additional  
16 stabilization of the tetrahedral intermediate *via* coordination and chelate formation, thus favoring  
17 the reaction.<sup>38,39</sup>

18  
19 In this work, we describe the development of the first PTML model taking into consideration  
20 the structure of the reactants, products, and intermediates of the reaction as well as the reaction  
21 conditions for a query reaction and a reaction of reference at the same time. We compared  
22 different linear and non-linear alternative models using GLR (General Linear Regression) and  
23 ANN (Artificial Neural Networks) algorithms. In addition, a theoretical and experimental study  
24 of new Parham reactions has been carried out to illustrate the development of the model and the  
25 practical applications, including simulations of the space of reaction and Hammett analysis.





**Scheme 2.** General types of Parham cyclizations collected in the dataset, with intermediates considered (ArLi and CI)

## 2. MATERIALS AND METHODS

### 2.1. Computational Methods.

**Parham reaction dataset.** To the best of our knowledge, there are no previous reports of datasets of Parham reactions for ML studies. Thus, a large dataset with >100 chemical reactions, carried out experimentally by our group and others, was collected from public literature,<sup>38-54</sup> as well as new experimental results not reported before. Overall, the dataset includes 117 reactions with 93 different substrates, including 13 types of internal electrophiles, 3 RLi reagents, and 64 products, with a wide range of yields (See Supporting Information for details). These 117 reactions include 107 reactions (n = 1 – 107) collected from the literature and 10 new reactions (n = 108 – 117)

reported here. These reactions have been carried out in many different conditions [c<sub>i</sub>] including different values of c<sub>0</sub> = LHE temperature (T<sub>1</sub>) and c<sub>1</sub> = time (t<sub>1</sub>), c<sub>2</sub> = reaction temperature (T<sub>2</sub>) and c<sub>3</sub> = time (t<sub>2</sub>), and c<sub>4</sub> = RLi equivalents (equiv). Next, structural and/or physicochemical variables V(m<sub>q</sub>) or molecular descriptors for the substrates, products, and proposed intermediates of reaction for LHE and for the cyclization step were calculated. A total of 10 000 pairs of query vs. reference reactions were sampled from this dataset. Each member of the pair was selected at random.

**PTML model.** Different schemes (multiplicative, additive) may be used for the construction of the functions. In so doing, different initial models (H<sub>0</sub> hypothesis) may be selected and tested. In this work, we propose an additive hypothesis H<sub>0</sub>. Thus, we consider that the initial value of ΔYld(%) for a new or query reaction is the value of the reaction of reference [Yld(%)<sub>ref</sub>] (value to be perturbed). Next, we can predict the value of the query reaction [Yld(%)<sub>new</sub>] by adding to Yld(%)<sub>ref</sub> the corrections due to structural perturbations ΔV<sub>k</sub>(m<sub>q</sub>)<sub>g</sub> and/or operational perturbations ΔV(c<sub>r</sub>) (changes in the experimental conditions). The formula of the PTML model used is shown on Equation 2.

$$\begin{aligned}
 \text{Yld}(\%)_{\text{pred}} &= \text{Yld}(\%)_{\text{ref}} + \sum_{q=1}^{q_{\text{max}}} \sum_{g=1}^{g_{\text{max}}} a_{q,g} \cdot \Delta V_k(m_q, 'm_q)_g \quad (2) \\
 &+ \sum_{r=1}^{r_{\text{max}}} b_r \cdot \Delta V(c_r, 'c_r) + e_0 + e_0 \\
 \text{Yld}(\%)_{\text{pred}} &= \text{Yld}(\%)_{\text{ref}} + \sum_{q=1}^{q_{\text{max}}} \sum_{g=1}^{g_{\text{max}}} a_{q,g} \cdot [V_k(m_q) - V_k('m_q)]_g \\
 &+ \sum_{r=1}^{r_{\text{max}}} b_r \cdot [V(c_r) - V('c_r)] + e_0
 \end{aligned}$$

The model uses as input the observed values of Yld(%)<sub>ref</sub> for the reaction of reference and two sets of PT operators ΔV(c<sub>r</sub>, 'c<sub>r</sub>) and ΔV(m<sub>q</sub>, 'm<sub>q</sub>). The PT operators are used to quantify the perturbations in the reaction conditions ΔV(c<sub>q</sub>, 'c<sub>r</sub>) or in the molecular structure ΔV(m<sub>q</sub>, m<sub>r</sub>),

1  
2  
3 respectively. The operators of the type  $\Delta V(c_q, 'c_r) = [V(c_r) - V('c_r)]$  are used to quantify  
4 perturbations in the reaction condition of query reaction  $c_q$  compared to the experimental  
5 conditions of the reaction of reference  $c_r$ .  
6  
7  
8

9  
10 On the other hand, the operators of the type  $\Delta V_k(m_q, 'm_q)_g = [V_k(m_q) - V_k('m_q)]_g$  quantify  
11 structural perturbations for five different classes of molecules with different roles in the reaction.  
12 These classes of molecules are the same for query and reference reactions ( $q = 'q$ ). The different  
13 types of molecules are:  $m_0 =$  Substrate (S),  $m_1 =$  Product (P),  $m_2 =$  RLi reagent,  $m_3 =$  IE,  $m_4 =$   
14 ArLi Intermediate, and  $m_5 =$  Cl.  
15  
16  
17  
18  
19

20 The structural variables or molecular descriptors used in this study were the average values of  
21 electronegativities  $V_k(m_q)_g = \chi_k(m_q)_g$  for different sets or groups of atoms ( $g$ ) in the molecule.  
22 This average value of electronegativity runs over all the atoms in  $g$  and all their neighbors placed  
23 at topological distance  $d \leq k$  (from  $k_{\min} = 0$  to  $k_{\max} = 5$ ). The groups of atoms considered in this  
24 study were  $g_1 =$  Total (all atoms in the molecule),  $g_2 =$  Heteroatoms,  $g_3 =$  Halogens,  $g_4 =$   
25 Saturated Carbons, and  $g_5 =$  Unsaturated Carbons. A detailed explanation of all the input  
26 variables used in this model is shown in Table 1. These average values of  $\chi(m_q)_{gk}$  were calculated  
27 using a Markov chain algorithm published by our group in previous works.<sup>55</sup>  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 1.** Definition of variables used as inputs of the model

Experimental conditions ( $\mathbf{c}_q$ )	Perturbation operators	Type of operator
HLE Temperature ( $T_1$ )	$\Delta V(T_{1q}, T_{1r}) = \Delta T_1 = T_{1q} - T_{1r}$	Temperature deviation
HLE time ( $t_1$ )	$\Delta V(t_{1q}, t_{1r}) = \Delta t_1 = t_{1q} - t_{1r}$	Time deviation
Cyclization Temperature ( $T_2$ )	$\Delta V(T_{2q}, T_{2r}) = \Delta T_2 = T_{2q} - T_{2r}$	Temperature deviation
Cyclization time ( $t_2$ )	$\Delta V(t_{2q}, t_{2r}) = \Delta t_2 = t_{2q} - t_{2r}$	Time deviation
RLi Equivalent (RLieq)	$\Delta V(\text{RLi}(\text{eq})) = \text{RLi}(\text{eq})_q - \text{RLi}(\text{eq})_r$	Conc. difference
Molecules ( $m_q$ ) <sup>a</sup>	Perturbation terms	Type of operator <sup>a</sup>
Substrate (S)	$\Delta\chi_k(\text{S}_q, \text{S}_r)_g = [\chi_k(\text{S}_q)_g - \chi_k(\text{S}_r)]_g$	
Product (P)	$\Delta\chi_k(\text{P}_q, \text{P}_r)_g = [\chi_k(\text{P}_q)_g - \chi_k(\text{P}_r)]_g$	Change of the average value of
Internal Electrophile (E)	$\Delta\chi_k(\text{E}_q, \text{E}_r)_g = [\chi_k(\text{E}_q)_g - \chi_k(\text{E}_r)]_g$	Electronegativity $\chi_k(m_q)$ in the structure of the query
Organolithium reagent (RLi)	$\Delta\chi_k(\text{RLi}_q, \text{RLi}_r)_g = [\chi_k(\text{RLi}_q)_g - \chi_k(\text{RLi}_r)]_g$	molecule $m_q = \text{S}_q, \text{P}_q, \text{E}_q,$
Aryllithium Intermediate (ArLi)	$\Delta\chi_k(\text{ArLi}_q, \text{ArLi}_r)_g = [\chi_k(\text{ArLi}_q)_g - \chi_k(\text{ArLi}_r)]_g$	$\text{RLi}_q, \text{ArLi}_q$ or $\text{Cl}_q$ in regard to average value of reference.
Cyclized Intermediate (CI)	$\Delta\chi_k(\text{CI}_q, \text{CI}_r)_g = [\chi_k(\text{CI}_q)_g - \chi_k(\text{CI}_r)]_g$	

<sup>a</sup> Change of the average value of Electronegativity  $\chi_k(m_q)_g$  in the structure the query molecule  $m_q = \text{S}_q, \text{P}_q, \text{E}_q, \text{RLi}_q, \text{ArLi}_q,$  or  $\text{Cl}_q$  with respect to the average value of Electronegativity  $\chi_k(m_r)_g$  the same molecules in the reaction of reference  $m_r = \text{S}_r, \text{P}_r, \text{E}_r, \text{RLi}_r, \text{ArLi}_r,$  or  $\text{Cl}_r$ . The values of  $\chi_k(m_q)_g$  or  $\chi_k(m_r)_g$  are the average value of the atomic Electronegativities of Pauling ( $\chi$ ) for all the atoms in the group  $g$  and all their neighbors atoms placed at a topological distance  $k \leq 5$ .

It should be noted that the model in this form is not adequate for regression studies, as there could be many repeated values of  $\text{Yld}(\%)_{\text{pred}}$  vs. different values of  $\text{Yld}(\%)_{\text{ref}}$ , which may lead to distortions in the normal distribution of the data. Thus, the following form of the Equation 3 was used for the regression analysis we used.

$$\begin{aligned} \Delta Yld(\%) &= \sum_{q=1}^{q_{\max}} \sum_{g=1}^{g_{\max}} a_{q,g} \cdot \Delta V_k(m_q, 'm_q)_g + \sum_{r=1}^{r_{\max}} b_r \cdot \Delta V(c_r, 'c_r) + e_0 \\ &= \sum_{q=1}^{q_{\max}} \sum_{g=1}^{g_{\max}} a_{q,g} \cdot [V_k(m_q)_g - V_k('m_q)_g] + \sum_{r=1}^{r_{\max}} b_r \cdot [V_k(c_r) - V_k('c_r)] + e_0 \end{aligned} \quad (3)$$

We sought a linear PTML model with this form using a GLR algorithm implemented in the software STATISTICA. We also explored other linear and non-linear alternative PTML models using ANN algorithms (see last section of results and discussion). The ANN models trained have Linear Neural Network (LNN), Multiple Layer Perceptron (MLP), and Radial Basis Function (RBF). We used the values of Squared Regression coefficient ( $R^2$ ), Regression coefficient (R), Error Mean, Standard Deviation (S.D.), Average (Avg.), Maximum (Max.), and Minimum (Min.), and other statistics to study the dataset and compare the models.<sup>56</sup>

**2.2. Experimental Methods. Typical Procedure for the Parham Cyclization.** Synthesis of (S)-7,8-dimethoxy-1,2,3,10a-tetrahydropyrrolo[1,2-*b*]isoquinolin-10(5*H*)-one (**P64**) (Table 5, n = 111). To a solution of *N,N*-diethyl-1-(4,5-dimethoxy-*o*-yodobenzyl)pyrrolidine-2-carboxamide (**S92**) (0.09 g, 0.20 mmol) and TMEDA (0.071 mL, 0.46 mmol) in dry THF (10 mL), *n*-BuLi (0.34 mL of a 1.3 M solution in hexane, 0.44 mmol) was added at -78 °C, and the resulting mixture was stirred at this temperature for 1 h. The reaction was quenched by the addition of sat.  $\text{NH}_4\text{Cl}$  (5 mL). The organic layer was separated, and the aqueous phase was extracted with AcOEt (3 × 5 mL). The combined organic extracts were dried ( $\text{Na}_2\text{SO}_4$ ) and concentrated in vacuo. Flash column chromatography (silica gel, 50% AcOEt:MeOH) afforded pyrroloisoquinolone **P64** as white powder (0.02 g, 49%): mp (AcOEt/MeOH) 163-165 °C;  $[\alpha]_{20}^D = -14.1$  (c = 1,  $\text{CH}_2\text{Cl}_2$ ); IR (KBr) 1673  $\text{cm}^{-1}$ ;  $^1\text{H}$  NMR ( $\text{CDCl}_3$ ) 1.81-1.90 (m, 2H), 2.04-2.23 (m, 2H), 2.50 (q,  $J = 8.6$  Hz, 1H), 2.93 (td,  $J = 8.6, 1.8$  Hz, 1H), 3.17-3.24 (m, 1H), 3.69 (dd,  $J = 15.0, 1.3$  Hz, 1H), 3.92 (s, 3H), 3.93 (s, 3H), 4.13 (d,  $J = 15.0$  Hz, 1H), 6.67 (s, 1H), 7.51 (s, 1H);  $^{13}\text{C}$  NMR ( $\text{CDCl}_3$ ) 21.4, 25.0, 54.0, 54.3, 56.0, 56.1, 69.0, 108.1, 124.4, 137.2, 148.2,

153.5, 195.0. MS (CI)  $m/z$  (rel intensity) 248 ( $MH^+$ , 100), 247 (34), 246 (12), 231 (7), 219 (11), 178 (12), 151 (6). HRMS (CI-TOF) Calcd. for  $C_{14}H_{18}NO_3$  [ $MH$ ] $^+$  248.1287; found: 248.1293. Anal. Calcd. for  $C_{14}H_{17}NO_3$ : C, 68.00; H, 6.93; N, 5.66. Found: C, 68.34; H, 6.75; N, 5.75.

### 3. RESULTS AND DISCUSSION

**3.1. Description of the dataset.** As has been stated, many factors affect the yield of Parham reactions. Therefore, the goal was to develop predictive computational model useful to search for optimal reaction conditions taking into consideration all the experimental variables involved. These reactions may have been carried out under different experimental conditions (see Table 2). Thus, the number of perturbations that could be carried out experimentally for reactions in our dataset can be easily calculated using Equation 4.

$$N_{\max} = N_{\text{products}} \cdot \prod_{r=1}^{r=5} \left[ \frac{\text{Max}(c_r) - \text{Min}(c_r)}{\text{Step}(c_r)} \right] \quad (4)$$

In this equation,  $N_{\max}$  stands for the maximum number of reactions,  $N_{\text{products}}$  is the number of different products of reactions in the dataset,  $\text{Max}(c_r)$  and  $\text{Min}(c_r)$  are the maximum and minimum values of the experimental conditions of reaction [ $T_1$ ,  $t_1$ ,  $T_2$ ,  $t_2$ , and  $\text{RLi}(\text{eq})$ ]. The  $\text{Step}(c_r)$  are the minimal variations allowed for the different experimental conditions of reaction  $c_r$ . A very simple calculation according to the previous equation give a total of number of reactions  $N_{\max} = 100\,421\,685$  changing the input parameters of the reactions in our dataset. We used values of  $\text{Step}(c_r) = 5 - 10$  to be conservative. Thus, it is unpractical to verify in the laboratory, an important reason to support the development of computational models of chemical reactivity for Parham reaction, or for many other reactions in organic synthesis. The values of Max., Min., Avg., and S.D. for the data set are reported in Table 2. These values were calculated

for the observed yield of reaction [Yld(%)<sub>obs</sub>] and the experimental parameters ( $c_r$ ) of all the reactions in the dataset.

**Table 2.** Summary of basic statistics for reactions in the dataset

Stat. <sup>a</sup>	Operational conditions of reaction for experimental design in simulation ( $c_r$ ) <sup>b</sup>					
	$c_1 = T_1(^{\circ}\text{C})$	$c_2 = t_1(\text{min})$	$c_3 = T_2(^{\circ}\text{C})$	$c_4 = t_2(\text{min})$	$c_5 = \text{RLi}(\text{eq})$	Yld(%) <sub>obs</sub>
Min.	-105	1	-105	1	1	0
Max.	-60	180	20	960	3	98
Avg.	-82.2	68.4	-57.7	101.8	2	69
S.D.	8.9	69.3	43.5	166.9	0.5	21.7
N <sub>observed</sub>	8	14	9	17	13	95
Stat. <sup>c</sup>	Calculation of Max. number of reactions after changing the conditions ( $c_r$ )					
Step	5	2	5	10	0.5	-
Max - Min	45	179	125	959	2	98
N <sub>max</sub>	9	90	25	96	4	733850775

<sup>a</sup>Stat. = Statistical parameters for the input parameters (conditions of operation) of all the Parham reactions present in our dataset: Min. = minimum value, Max. = maximum value, Avg. = average value, S.D. = Standard deviation, N<sub>total</sub> = Number of reactions present in our dataset. <sup>b</sup> Experimental parameters (see Scheme 2):  $c_1 = T_1(^{\circ}\text{C})$ : Temperature for the lithium-halogen exchange reaction (LHE),  $c_2 = t_1(\text{min})$  = reaction time for the LHE,  $c_3 = T_2(^{\circ}\text{C})$  = Temperature of cyclization reaction step,  $c_4 = t_2(\text{min})$  = reaction time cyclization step,  $c_5 = \text{RLi}(\text{eq})$  = amount of the organolithium reagent (RLi) expressed in equivalents. (See Supporting Information for the details of all reactions in the dataset). <sup>c</sup> Step = integer number to express the minimal change allowed in one experimental condition, N<sub>max</sub> = Maximum number of reactions experimentally reachable if all the experimentally possible variations for the reactions in our dataset were carried out.

**3.2. PTML model for Parham reactions.** Next, a general PTML model for Parham reactions was developed. The overall p-level of the model is  $p < 0.05$  and all the variables of the model, but one, are statistically significant according to student test (see values of t and p-level in Table 3). The equation of this linear PTML model is shown in Equation 5.

$$\begin{aligned} \Delta Yld(\%)_{\text{pr}} = & 11.970 \cdot \Delta X_0(S)_{11Br} - 29.302 \cdot \Delta X_2(S)_{\text{Het}} \quad (5) \\ & + 22.405 \cdot \Delta X_2(IE)_{\text{Het}} + 7.697 \cdot \Delta X_2(P)_{\text{Het}} \\ & + 18.841 \cdot \Delta X_2(ArLi)_{\text{Het}} - 210.756 \cdot \Delta X_0(RLi)_{\text{rot}} \\ & - 6.846 \cdot \Delta RLi(eq) + 0.011 \cdot \Delta T_1 - 0.085 \cdot \Delta t_1(\text{min}) \\ & - 0.056 \cdot \Delta T_2 - 0.016 \cdot \Delta t_2(\text{min}) + 0.472 \\ n = & 10000 \quad R = 0.88 \quad R^2 = 0.78 \quad F = 2150.242 \quad p < 0.005 \end{aligned}$$

The first five variables express the contribution of the non-structural variables to the change in yield  $\Delta Yld(\%)$ . The other variables express the contribution of changes in the molecular structure of the different components of the reaction S, IE, P, ArLi, and RLi. The variables used to account for changes in the structure of the cyclized intermediates CI were not significant for the prediction of  $\Delta Yld(\%)$  according to this model. This could be because the information of the variable  $\Delta \chi_2(\text{CI})_{\text{Het}}$  is redundant with respect to the variable  $\Delta \chi_2(\text{P})_{\text{Het}}$ . Pareto's diagram for the variables (result not presented) shows that almost all the variables are important to the model. The variable  $T_1$  is the only one with a p-level higher than 0.05. We do not show the Pareto's diagram of the model because it can be easily constructed with the absolute values of the t-values presented in Table 3.<sup>56</sup> Notably, the present PTML model is able to predict correctly a very high number of perturbations both in training and in external validation series  $n = 10000$ . In fact, the model has values of  $R^2 = 0.78$  in training series. The correlation coefficient of the training series was  $R = 0.88$ .

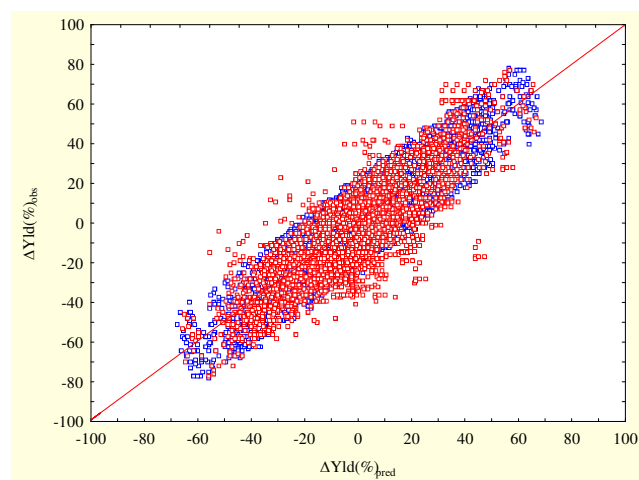
In addition, the correlation coefficient of the external validation series was  $R = 0.83$ . The values of  $\Delta Yld(\%)_{\text{obs}}$  vs.  $\Delta Yld(\%)_{\text{pred}}$  for 10000 pairs of reactions; training (blue) and validation (red) are depicted in Figure 1. It is important to point out that the residuals of this model have a normal distribution (Figure 2) and an average value = 0.56 (near to 0). Consequently, the model fulfills these two important parametric assumptions of the linear regression analysis.<sup>56</sup>



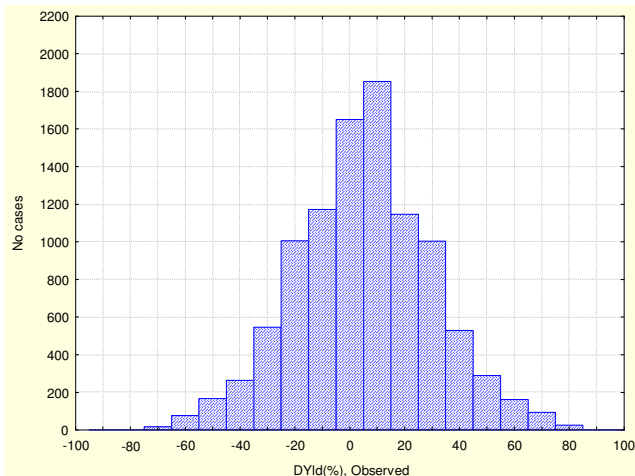
**Table 3.** Results of the PTML regression model

Coefficient	Input <sup>a</sup>	$\Delta Yld(\%)^b$	S.E. <sup>c</sup>	t <sup>c</sup>	p-level <sup>d</sup>
a <sub>1</sub>	$\Delta\chi_0(S)_{I/Br}$	11.970	0.42174	28.3819	0.000000
a <sub>2</sub>	$\Delta\chi_2(IE)_{Het}$	22.405	0.66544	33.6701	0.000000
a <sub>3</sub>	$\Delta\chi_2(P)_{Het}$	7.697	0.60475	12.7275	0.000000
a <sub>4</sub>	$\Delta\chi_2(S)_{Het}$	-29.302	0.63461	-46.1736	0.000000
a <sub>5</sub>	$\Delta\chi_2(ArLi)_{Het}$	18.841	0.51208	36.7928	0.000000
a <sub>6</sub>	$\Delta\chi_0(RLi)_{Tot}$	-210.756	13.31498	-15.8285	0.000000
b <sub>0</sub>	$\Delta T_1$	0.011	0.01729	0.6570	0.511176
b <sub>1</sub>	$\Delta t_1$	-0.085	0.00295	-28.6763	0.000000
b <sub>2</sub>	$\Delta T_2$	-0.056	0.00429	-12.9644	0.000000
b <sub>3</sub>	$\Delta t_2$	-0.016	0.00091	-17.1393	0.000000
b <sub>4</sub>	$\Delta RLi(eq)$	-6.846	0.23898	-28.6483	0.000000
e <sub>0</sub>	Independent term	0.472	0.14232	3.3146	0.000923

<sup>a</sup> Input variables of the model. <sup>b</sup> Coefficients of the variables in the model. <sup>c</sup> Standard error of the coefficients. <sup>d</sup> Student t-value. <sup>e</sup> p-level of error.



**Figure 1.** Observed vs. predicted  $\Delta Yld(\%)$  for 10000 pairs of reactions; training (blue) and external validation series (red).



**Figure 2.** Histogram of observed values of  $\Delta Yld(\%)_{obs}$

**3.3. PTML predictions with one reaction of reference.** As has been shown, the PTML model is able to predict the change in yield  $\Delta Yld(\%)_{pred}$  for a query reaction with respect to different reactions of reference (one by one). However, the prediction of  $\Delta Yld(\%)$  for different pairs of reactions is not the final objective of this model. The main interest of the synthetic chemist is the prediction of the yield of reaction  $Yld(\%)_{pred}$  of a new reaction using the yield of a known reaction as reaction of reference. In order to do these predictions (use of the model in practice), the model according to Equation 6 was used.

$$\begin{aligned}
 Yld(\%)_{pred} = & Yld(\%)_{ref} + 0.011 \cdot \Delta T_1 - 0.085 \cdot \Delta t_1(\text{min}) - 0.056 \cdot \Delta T_2 \\
 & - 0.016 \cdot \Delta t_2(\text{min}) - 6.846 \cdot \Delta RLi(eq) \\
 & + 11.970 \cdot \Delta X_0(S)_{I/Br} + 22.405 \cdot \Delta X_2(E)_{Het} \\
 & + 7.697 \cdot \Delta X_2(P)_{Het} - 29.302 \cdot \Delta X_2(S)_{Het} \\
 & + 18.841 \cdot \Delta X_2(AI)_{Het} - 210.756 \cdot \Delta X_0(RLi)_{Tot} + 0.472
 \end{aligned} \quad (6)$$

In this sense, case  $Yld(\%)_{pred} = Yld(\%)_{calc}$  is equal to the yield calculated with the model for one reaction of reference (see next section). It should be remembered that  $\Delta Yld(\%) = Yld(\%)_{pred} - Yld(\%)_{ref}$  to predict the  $Yld(\%)_{pred}$  using as input the value of yield  $Yld(\%)_{ref}$  for one reaction of reference and the values of the difference operators  $\Delta V(c_k, 'c_k)$  and  $\Delta V(m_q, 'm_q)$  for this pair of reactions (see details in Table 1). The reaction of reference in many problems is a reaction with a

1  
2  
3 low experimental value of yield [ $\text{Yld}(\%)_{\text{obs}}$ ] to be optimized. In this situation changes are  
4  
5 required on the values of experimental conditions,  $\Delta V(c_k, 'c_k)$  operators, in order to increase the  
6  
7 value of  $\text{Yld}(\%)_{\text{pred}}$  for the query reaction. The chemical structure of the reactants,  $\Delta V(m_q, 'm_q)$   
8  
9 operators, can also be changed. In these cases, it is recommendable to change only the value of  
10  
11 one experimental condition or only one reactant each time and keep all other values constant,  $c_k$   
12  
13 =  $'c_k$  and  $m_q = 'm_q$ . This strategy may help to keep the perturbations as small as possible and  
14  
15 increase the accuracy of the predictions.  
16  
17  
18

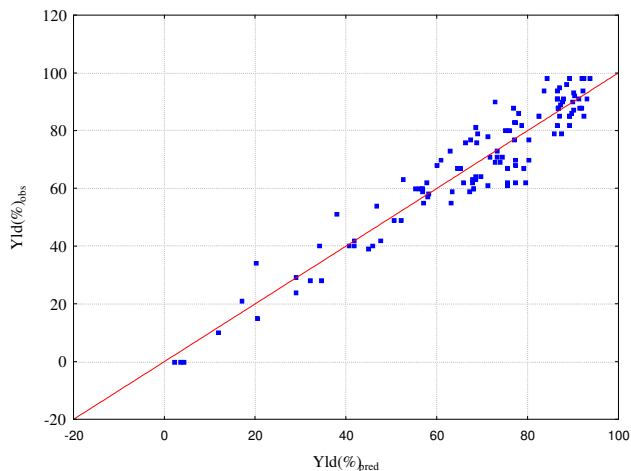
19 **3.4. PTML for predictions with multiple reactions of reference.** The model was built using  
20  
21 10000 different pairs of reactions selected at random from all possible pairs in our dataset of  $n =$   
22  
23 117 reactions. In this sense, for a single reaction  $\text{Yld}(\%)_{\text{pred}} = \text{Yld}(\%)_{\text{calc}}$  or yield calculated with  
24  
25 the model when we use only one reaction of reference. However, we can calculate multiple  
26  
27 values of yield  $\text{Yld}(\%)_{\text{calc}}$  for a single reaction if we use the  $n$  different reactions of reference in  
28  
29 the dataset. In many cases, the selection of the reaction of reference is clear (See previous  
30  
31 section). However, when the selection of the reaction of reference is not clear a method to work  
32  
33 with multiple reactions of reference is required. In this work, we have calculated the yield  
34  
35 predicted  $\text{Yld}(\%)_{\text{pred}} = \text{Avg}(\text{Yld}(\%)_{\text{calc}})$  as the average of all the values of yield calculated  
36  
37 [ $\text{Yld}(\%)_{\text{calc}}$ ] with the model using all reactions in our dataset as reference ( $n = 117$ ). The  
38  
39 statistical analysis is summarized in Table 4.  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 4.** Results obtained with multiple references

Model	B <sup>a</sup>	S.E.	t	p-level
a <sub>0</sub>	-10.43	2.40	-4.35	< 0.05
Yld(%) <sub>pred</sub>	1.13	0.03	33.94	< 0.05
Parameters	R	F	p-level	SEE
Values	0.95	1152.15	< 0.05	6.856

<sup>a</sup> B = Coefficients of the linear equation, S.E. = Standard Error, t = Student test parameter, p-level = level of error, Avg.Res. = Average of Residuals, R = Regression coefficient, F = Fisher ratio, SEE = Standard Error of Estimates.

Notably, the coefficient of regression obtained for predictions with multiple references is statistically significant with p-values < 0.05 and regression coefficients R = 0.95. This means that the model has  $R^2 = 0.9025$ , and explains >90% of variance. In addition, the predictions showed a value of  $r_m^2 = 0.64$ , higher than 0.5, which indicates that the model is acceptable. The index  $r_m^2$ , reported by Roy *et al.*,<sup>57</sup> penalizes other indices such as the classical  $q^2$ , which should be used with caution according to Golbraikh and Tropsha.<sup>58</sup> The average value of residuals for this method is Avg. Res. = - 1.0. A graphical representation of observed vs. predicted values of yield for 117 Parham reactions using multiple reactions of reference are shown in Figure 3.

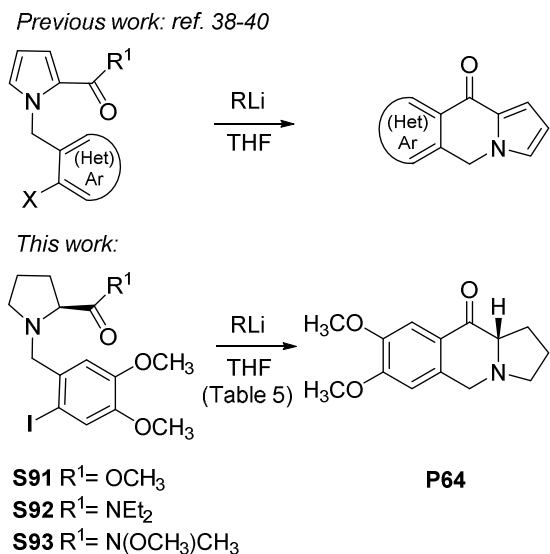


**Figure 3.** Observed vs. Predicted values of yield (%) using multiple reactions of reference

**3.5. Theoretical and experimental study of new reactions.** Once the PTML model was established, we decided to illustrate the practical use of the model with a real case. To this end, we carried out both the experimental and theoretical study of Parham reactions that had not been previously reported. The model would allow predicting the effect of changes (perturbations) in the input experimental conditions (changes in the substitution pattern of the substrate, the halogen, the internal electrophile and experimental conditions  $c_r$ ) for a high number of reactions with low cost of time. In fact, we carried out a 500000-points simulation to illustrate the use of the model.

*Parham cyclization of S91-S93.* We have previously shown that aryllithiums generated from metalation of *N,N*-diethyl-*[N-(o*-halobenzyl)]pyrrole-2-carboxamides undergo intramolecular cyclization to give pyrrolo[1,2-*b*]isoquinolines.<sup>38-40,59</sup> For this work, we decided to study the application of this metalation-cyclization sequence for the synthesis of tetrahydropyrrolo[1,2-*b*]isoquinolines. This is a structural framework present in natural products such as the lycorine class of *Amaryllidaceae* alkaloids and the phenanthroindolizidine alkaloids.<sup>60</sup> In this case, the corresponding *N-(o*-halobenzyl)]pyrrolidine derivatives **S91-S93** were selected as substrates (see

Supporting information for the preparation), in order to study the effect of using a pyrrolidine ring instead of a pyrrole on the reactivity. We selected a methyl ester (**S91**), *N,N*-diethylamide (**S92**) and Weinreb amide (**S93**) as internal electrophiles (Scheme 3). This procedure would allow the synthesis of enantiomerically pure tetrahydropyrroloquinoline **P64** starting from a compound from the chiral pool, such as L-proline.



### Scheme 3. Parham cyclization of **S91-S93**

When a methyl ester was used as internal electrophile (**S91**), it was necessary to use a bulky and non-nucleophilic reagent as MesLi as metalating agent in order to avoid the direct addition of the RLi to the carbonyl group (Table 5, entries 108-110. The numbering of the compounds and entries on Table 5 corresponds to the numbering on the reaction dataset in the Supporting Information). Under these conditions, the reaction was very fast (5 min) at low temperature, even using 1 equivalent of the metalating agent (Table 5, entry 109), but only moderate yields of **P64** were obtained. When diethylamide **S92** was used, the reaction could be carried out with *n*-BuLi, but a longer reaction time was required (entry 111). The best result was obtained with Weinreb amide **S93**, as could be expected (entries 112-117). The metalation reactions could be carried out

efficiently with *n*-BuLi and *t*-BuLi, and no significant difference in the yield of **P64** was observed with the use of TMEDA as additive (entry 113 vs. 114).

**Table 5.** Parham cyclization of **S91-S93**

n	Subs.	RLi	RLi (equiv)	T (°C)	t (min)	Yld (%) <sup>a</sup>
108	<b>S91</b>	MesLi	(2)	-105	5	40
109	<b>S91</b>	MesLi	(1.5)	-105	5	51
110	<b>S91</b>	MesLi	(1)	-105	5	42
111	<b>S92</b>	<i>n</i> -BuLi	(2.2) <sup>b</sup>	-78	60	49
112	<b>S93</b>	<i>n</i> -BuLi	(2.2) <sup>b</sup>	-78	60	62
113	<b>S93</b>	<i>n</i> -BuLi	(2.2)	-78	60	67
114	<b>S93</b>	<i>t</i> -BuLi	(2.2) <sup>b</sup>	-78	60	61
115	<b>S93</b>	<i>t</i> -BuLi	(2.2) <sup>b</sup>	-78	30	62
116	<b>S93</b>	<i>t</i> -BuLi	(2.2) <sup>b</sup>	-78	15	70
117	<b>S93</b>	<i>t</i> -BuLi	(2.2) <sup>c</sup>	-78	15	62

<sup>a</sup> Yld (%)<sub>obs</sub> is the yield of isolated pure product **P64**. <sup>b</sup> TMEDA (2.3 equiv) was used as additive. <sup>c</sup> TMEDA (2.1 equiv) was used as additive

On the other hand, shorter reaction times led to higher yields (entry 115 vs. 116). Tetrahydropyrroloisoquinoline **P64** was obtained without racemization, as a single enantiomer of *S* configuration, starting from the enantiomerically pure pyrrolidines **S91-S93**. The absolute configuration (*S*) was confirmed by single-crystal X-ray analysis of **P64** (CCDC1560347 contains the supplementary crystallographic data for **P64**; see Supporting Information). These results seem to indicate that there is a significant difference in the reactivity between the pyrrolidines and pyrrole substrates, with the same internal electrophile. Thus, while the corresponding pyrroles required longer reaction times (3 h) or an increase of the temperature to room temperature after the LHE step to obtain the cyclized products,<sup>38</sup> cyclization of pyrrolidines took place at low temperature and in shorter reaction times (15 to 60 min) (see database in Supporting Information). It is clear that changes in the structure of the substrates require further

optimization of the reaction conditions. Consequently, we carried out the predictive study with our new PTML model. We selected the entry 116 (Table 5) as reaction of reference for a large-scale numerical simulation study.

*PTML prediction of new reactions using multiple references.* We applied the PTML model to carry out a prediction of the yield [ $\text{Yld}(\%)_{\text{pred}}$ ] for these new Parham reactions. The experimental conditions reported in Table 5 were used as reaction input. The results are shown in Table 6. Results for the 117 reactions in the dataset are included in Supporting Information.

**Table 6.** Predictive study of reactions 108 to 117.

Reactions		Yld(%) <sup>b</sup>			
n	Subs <sup>a</sup>	Obs.	Expt.	Pred.	Res.
108	<b>S91</b>	40	44.3	39.2	0.8
109	<b>S91</b>	51	44.3	42.6	8.4
110	<b>S91</b>	42	44.3	46.0	-4.0
111	<b>S92</b>	49	64.0	53.6	-4.6
112	<b>S93</b>	62	64.0	75.7	-13.7
113	<b>S93</b>	67	64.0	75.7	-8.7
114	<b>S93</b>	61	59.9	75.7	-14.7
115	<b>S93</b>	62	59.9	77.2	-15.2
116	<b>S93</b>	70	59.9	77.2	-7.2
117	<b>S93</b>	62	59.9	79.3	-17.3

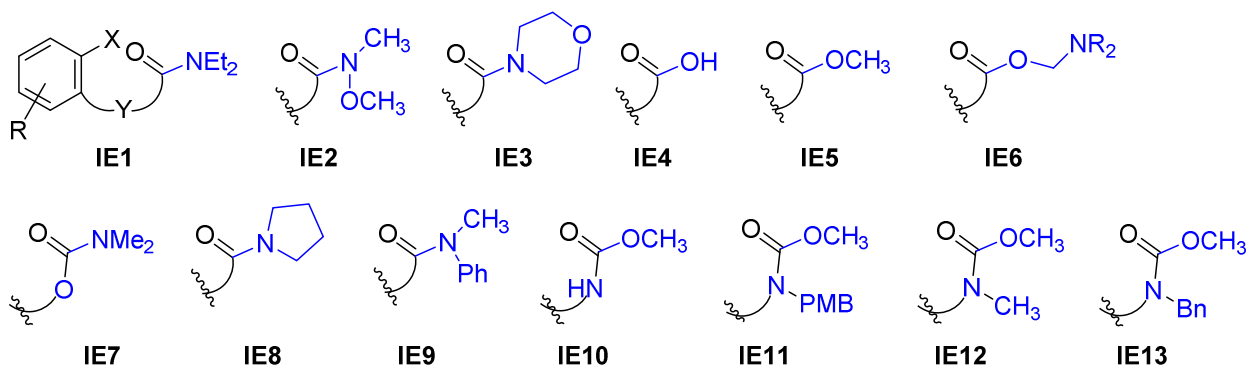
<sup>a</sup>Subs = Substrate. <sup>b</sup>Yield Obs. ( $\text{Yld}(\%)_{\text{obs}}$ ) is the yield of reaction observed experimentally (Table 5). Yield Expt. is the expected value of  $\text{Yld}(\%)$  calculated as  $\text{Yld}(\%)_{\text{expt}} = \text{Avg}(\text{Yld}(\%)_{\text{obs}})$  the average observed values of  $\text{Yld}(\%)_{\text{obs}}$  only for all the reactions carried out with the same RLi and solvent. Yield Pred.  $\text{Yld}(\%)_{\text{pred}} = \text{Avg}(\text{Yld}(\%)_{\text{calc}})$  is the value of yield calculated with the PTML model using the 117 reactions of reference. Yield Res.  $\text{Yld}(\%)_{\text{res}} = \text{Yld}(\%)_{\text{pred}} - \text{Yld}(\%)_{\text{obs}}$  is the residual value.

In general, all the reactions reported here have an observed experimental value of yield in the range 40-70%. The model predicts very well the observed values with values in a similar range 39-79 %. The values predicted here for reactions 108-117 were never used to train the model.



Table 6 shows the results obtained for the reactions studied experimentally (Table 5, entries 108-117).

*PTML simulation of Parham reaction space.* Next, we applied the PTML model to carry out a large-scale simulation. In this simulation, we measured the effect of perturbations in experimental conditions  $\Delta V(c_k)$  over the value of yield for new Parham reactions  $[Yld(\%)_{pred}]$ . To this end, we carried out the prediction of all the reactions in the dataset changing the values of the different kinetic (time) and thermodynamic (temperature) factors  $V(c_k)$ . In order to generate a large set of perturbations  $\Delta V(c_k) = V(c_k) - V('c_k)$ , new sets of experimental conditions  $V(c_k)$  have to be generated compared to the values of the reactions of reference  $V('c_k)$ . We generated 500000 different sets of experimental conditions  $V(c_k)$  for hypothetical reactions using a random interpolation procedure. After that, those values were used to calculate the values of 500000 sets of perturbations  $\Delta V(c_k) = V(c_k) - V('c_k)$  in the experimental conditions. Next, we substituted all these 500000 sets of values  $\Delta V(c_k)$  into the PTML model. As a result, we obtained the new values of yield  $[Yld(\%)_{pred}]$  for 500000 query reactions. Table 7 summarizes the results found in this numeric simulation experiment, grouped according to the different types of internal electrophiles (IE) collected in the data set, and shown in Figure 4, with gradient color, which is related to higher (green) or lower (red) Yield (%), in order to obtain the best visual result.



**Figure 4.** Classes of internal electrophiles included in the simulation (Table 7)

**Table 7.** Exploration of the 500000-points space for *t*-BuLi mediated reactions

LHE step	<i>t</i> -BuLi (equiv). temperature (T <sub>1</sub> ) <sup>a</sup> time (t <sub>1</sub> ) <sup>a</sup>	1	1	1	1	1	>1	>1	>1	>1	>1
CI step	Temperature (T <sub>2</sub> ) <sup>a</sup> time (t <sub>2</sub> ) <sup>a</sup>	cold	warm	cold	cold	warm	cold	warm	cold	cold	warm
IE <sup>b</sup>	<b>IE1</b>	63	60.4	60	53	50.6	57	51.8	52	46	45.4
	<b>IE2</b>	77	75.1	79	69	67.5	75	71.5	72	65	64
	<b>IE3</b>	76	75.3	77	69	67.4	74	72.5	73	66	64.7
	<b>IE4</b>	78	75.7	81	69	68.6	78	72.5	73	67	66.2
	<b>IE5</b>	57	54.1	60	48	46.8	54	50.9	51	45	42.8
	<b>IE6</b>	74	72.6	77	67	66.3	74	69.4	69	64	63.1
	<b>IE7</b>	84	82.8	86	76	75.4	82	79.3	79	73	71.9
	<b>IE8</b>	67	65.8	70	58	57	65	61.2	61	55	53.2
	<b>IE9</b>	73	71.6	75	63	61.9	70	67.9	68	60	58.2
	<b>IE10</b>	70	68.6	72	64	62.2	69	65.3	65	60	58.3
	<b>IE11</b>	44	41.7	47	37	35.1	43	39.8	40	33	31
	<b>IE12</b>	67	65.5	69	58	56.6	65	61.6	62	54	53.3
	<b>IE13</b>	61	59.9	64	54	53.1	61	56.9	57	51	49.3

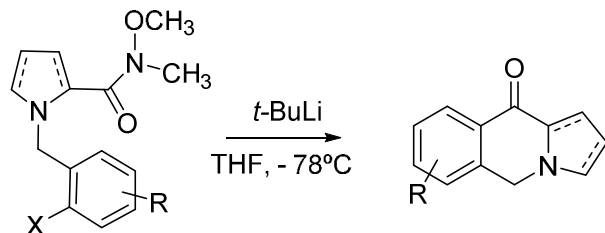
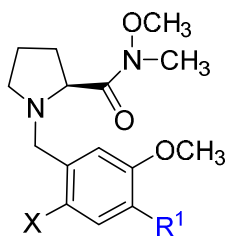
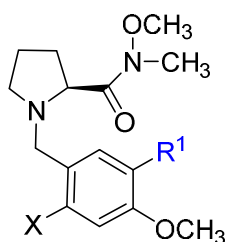
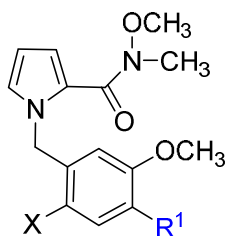
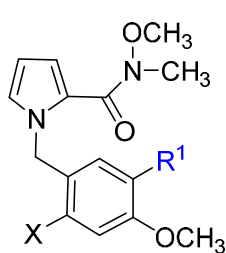
<sup>a</sup>Cold/warm or fast/slow cut-offs are the average values of reaction temperature and time for the two steps of reaction. These values are T<sub>1</sub> ≤ -82.2 °C (cold), t<sub>1</sub> ≤ 68.4 min (fast); T<sub>2</sub> ≤ -57.7 °C (cold), t<sub>2</sub> ≤ 101.8 min (fast). <sup>b</sup> IE: Internal Electrophile: see Figure 4. See Supporting information for the complete structures.

The simulation confirms the general trend expected for this type of reactions. Under all experimental conditions studied, the best results were obtained with carbamates as internal electrophiles, (**IE7**, Table 7, Figure 4), followed by amides. This could be explained assuming that, in these cases, metalation would be favored by a Complex Induced Proximity Effect (CIPE),<sup>61,62</sup> stabilizing the aryllithium intermediate. Among the amides, Weinreb or morpholinoamides (**IE2**, **IE3**) give consistently better results than simple amides (**IE1**, **IE8**, **IE9**), possibly due to the extra stabilization by chelation.<sup>37</sup> Regarding the reaction conditions, some useful trends can be observed. The amount (equivalents) of *t*-BuLi has not a relevant effect, although better results are generally predicted with the use of a stoichiometric amount,

1  
2  
3 compared to an excess. These predicted values also indicate that use of a short LHE time ( $t_1$   
4 short) is beneficial. On the other hand, better results are predicted when the time for the  
5 cyclization is extended ( $t_2$  long), rather than increasing the reaction temperature ( $T_2$  warm).  
6  
7

8  
9  
10 *PTML Hammett analysis of the effect of substituents on the aromatic ring and the electrophile.*

11  
12 The PMTL model can also be used to carry out a Hammett analysis, which is a well-known  
13 method to correlate the effect over reactivity of structural changes in reactants.<sup>63</sup> The method  
14 uses Hammett constants ( $\sigma$ )<sup>64</sup> or similar parameters to quantify the effect on reactivity of the  
15 introduction of chemical substituents in a given chemical system ( $\rho$ ). The method has been  
16 applied recently to important studies in computer-aided organic synthesis by Sigman *et al.*<sup>65-69</sup>  
17  
18 The Hammett equation used in this study is  $\text{Yld}(\%)_{\text{pred}} = \sigma \cdot \rho + a_0$ . In this equation,  $\text{Yld}(\%)_{\text{pred}}$   
19 are the values predicted with the PTML model. The reaction with the highest experimental value  
20 of yield in our experimental study (Table 5, entry116) was used as reference reaction in the  
21 PTML model. For this study, we selected the general reaction for the formation of  
22 pyrroloisoquinolines indicated in Scheme 4.  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**S1-8****S1** X = I  
**S5** X = Br**S2** X = I  
**S6** X = Br**S3** X = I  
**S7** X = Br**S4** X = I  
**S8** X = Br

$R^1$  = CH<sub>3</sub>, F, Cl, Br, I, OPh, OCOMe, O*i*Pr, NMe<sub>2</sub>,  
COOEt, CF<sub>3</sub>, Ph, NO<sub>2</sub>

**Scheme 4.** General reaction and structural modifications for the Hammett analysis (Table 8)

41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

In total, eight different series (**S1-8**) of compounds were studied, according to the general structures shown on Scheme 4. First, 13 different substituents ( $N_{\text{sbr}} = 13$ ), both electron donating and electron withdrawing, on positions 4 or 5 of the aromatic ring ( $R^1$ ) ( $N_{\text{brp}} = 2$ ) were selected using Weinreb amide as internal electrophile, with two different halogen atoms ( $X = \text{Br}, \text{I}$ ) ( $N_{\text{thd}} = 2$ ), and with a saturated (pyrrole) or unsaturated (pyrrolidine) ring ( $N_{\text{rie}} = 2$ ) (substrates type **S1-8**, Scheme 4, Table 8). This makes 104 different compounds for this series.

**Table 8.** Hammett analysis of structural perturbations

Structural Pattern <sup>a</sup>	Hammett Analysis (statistics) <sup>b</sup>						
	n	R	F	p	SEE	$\sigma$	$\rho$
<b>S1</b>	13	0.62	6.73	0.03	6.12	$\sigma_I$	-27.4
						$a_0$	85.6
<b>S2</b>	13	0.50	3.72	0.08	3.16	$\sigma_p$	-4.2
						$a_0$	75.5
<b>S3</b>	13	0.57	5.39	0.04	6.91	$\sigma_I$	-27.7
						$a_0$	83.3
<b>S4</b>	13	0.46	2.95	0.11	3.41	$\sigma_m$	-7.3
						$a_0$	76.2
<b>S5</b>	13	0.69	10.04	0.01	5.57	$\sigma_I$	-30.5
						$a_0$	87.3
<b>S6</b>	13	0.58	5.63	0.04	3.07	$\sigma_p$	-5.0
						$a_0$	76.2
<b>S7</b>	13	0.64	7.46	0.02	6.52	$\sigma_I$	-30.8
						$a_0$	85.0
<b>S8</b>	13	0.58	5.47	0.04	3.08	$\sigma_m$	-8.9
						$a_0$	77.3

<sup>a</sup>See Scheme 4 for series of compounds. See also Supporting Information. <sup>b</sup>n = number of cases (substituents), R = Regression coefficient, F = Fisher ratio, SEE = Standard Error of Estimates, p = p-level, significant p-values are < 0.05

To carry out this study, the SMILE codes of each compound were generated, and the molecular descriptors  $\Delta\chi_k$  were calculated and introduced into the PTML model to predict the new yield values [Yld(%)<sub>pred</sub>]. Last, a simple linear regression analysis of the Yld(%)<sub>pred</sub> vs. different constants of the substituents was carried out. Specifically, we used the Hammett parameters  $\sigma_m$  and  $\sigma_p$  to measure overall electron-donating or electron-withdrawing effects of substituents. We also used the constants  $\sigma_I$  and  $\sigma_R$  to measure inductive or resonance effects separately. These values were obtained from an excellent review of these methods reported by Hansch *et al.*<sup>64</sup> The values of the coefficients  $\rho$  for the  $\sigma$ -like constants of substituents are

depicted in Table 8 for the different series of compounds (see also Supporting Information). We investigated inductive effects for the substitution at the *meta* position with respect to the halogen atom (**S1**, **S3**, **S5**, **S7**), and resonance effects for the substitution at the *meta* position with respect to the halogen atom (**S2**, **S4**, **S6**, **S8**).

In general, the correlations are stronger in above 10% for compounds with substituent R<sup>1</sup> in *m* position (**S1**, **S3**, **S5**, **S7**) than for those *p*-substituted (**S2**, **S4**, **S6**, **S8**). The Yld(%)<sub>pred</sub> for compounds substituted in *m* position showed significant (p-level < 0.05) and stronger correlations R = 0.6 – 0.7 with the inductive effect constant  $\sigma_I$  and more negative values of  $\rho$  in the range  $\rho = -27$  to  $\rho = -31$ . Compounds substituted in *p* position showed weaker (R < 0.6) and/or non-significant correlations (p-level > 0.05). Table 9 shows the values of yield predicted with the PTML model [Yld(%)<sub>pred</sub>].

**Table 9.** Results of Hammett analysis for **S5-8** series

R <sup>1</sup>	Yld(%) <sub>pred</sub>		Hammett constants					
	( <i>meta</i> )	( <i>para</i> )	$\sigma_p$	$\sigma_m$	$\sigma_I$	$\sigma_R$		
	<b>S5</b>	<b>S7</b>	<b>S6</b>	<b>S8</b>				
Me	79.6	74.3	73.2	73.0	-0.17	-0.07	0.01	-0.18
NMe <sub>2</sub>	81.8	80.3	78.6	78.4	-0.83	-0.16	0.15	-0.98
O <i>i</i> Pr	78.8	77.0	73.8	73.6	-0.45	0.1	0.34	-0.79
OPh	79.0	77.1	72.8	72.7	-0.03	0.25	0.37	-0.4
Ph	89.4	89.7	71.4	71.2	-0.01	0.06	0.12	-0.13
F	68.5	65.6	78.8	77.0	0.06	0.34	0.45	-0.39
I	77.8	72.5	77.6	76.0	0.18	0.35	0.42	-0.24
Cl	74.7	70.2	78.9	77.1	0.23	0.37	0.42	-0.19
Br	75.9	71.1	79.4	77.6	0.23	0.39	0.45	-0.22
CO <sub>2</sub> Et	70.4	69.3	77.7	76.1	0.45	0.37	0.34	0.11
CF <sub>3</sub>	86.7	87.0	75.7	76.0	0.54	0.43	0.38	0.16
OCOMe	71.9	70.8	78.4	78.7	0.31	0.39	0.42	-0.11

---

NO <sub>2</sub>	62.1	60.4	67.8	65.7	0.78	0.71	0.65	0.13
-----------------	------	------	------	------	------	------	------	------

---

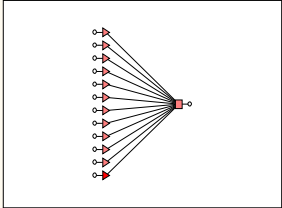
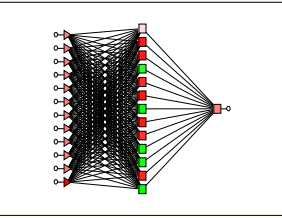
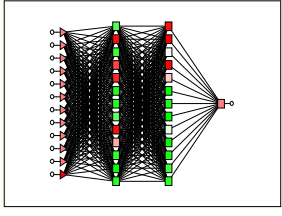
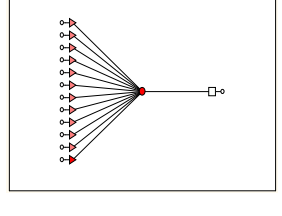
Table 9 also shows the respective values of the constants  $\sigma_m$ ,  $\sigma_p$ ,  $\sigma_I$  and  $\sigma_R$  for different substituents introduced in brominated systems (**S5-S8**). A detailed table of results for all the experiments is included in Supporting Information. This analysis, shows that slightly higher yields are predicted for the pyrrolidine systems compared to the pyrrole system (**S5** vs. **S7** and **S6** vs. **S8**). Regarding the substitution, although electron-withdrawing groups (*i.e.*: NO<sub>2</sub>, CO<sub>2</sub>Et) generally disfavor the cyclization, the overall effect of the substituents in the aromatic ring is quite moderate, and in some cases no correlation can be seen (*i.e.*: CF<sub>3</sub>). These results would be in agreement with computational studies carried out for related reactions, and suggest that different factors, and not only electronic effects, affect the course of the reaction.<sup>70</sup>

*PTML linear vs. non-linear models.* Finally, a comparative study of our linear model with non-linear models obtained using ANN algorithms was carried out. The ANN module of the software STATISTICA was used to process our dataset. In order to train the ANN models, we used the same variables previously selected by GLR stepwise methods. The use of the same variables allows us to compare the models in terms of performance, error, *etc.*, without introducing a bias error due to the variable selection strategy. Results are shown in Table 10.

In fact, the LNN model has the same variables and regression coefficient  $R = 0.88$  that our previous PTML linear regression model. In addition, we tested other non-linear ANN topologies like MLP and RBF. The MLP models trained may have one or two hidden layers of neurons. As can be seen, none of the ANN models tested outperforms the PTML linear models. For instance, the MLP models give similar results with  $R \approx 0.8$  in training and external validation series. In addition, the RBF topology has a notably lower  $R \approx 0.1$  (see Table 10). These results confirm

that the linear hypothesis used to seek the PTML model seems to be stronger for the present dataset.<sup>56</sup>

**Table 10.** PTML-ANN models

PTML-ANN <sup>a</sup>		Model parameters				
ANN Model profile	ANN Topology	Data set	R <sup>b</sup>	Error Mean	SD Error	SD Ratio
LNN 12:12-1:1		training	0.88	0.01	11.37	0.47
		test	0.83	2.60	14.40	0.56
MLP 12:12-13-1:1		training	0.88	-3.45	11.46	0.47
		test	0.83	-0.67	14.47	0.56
MLP 12:12-13-13-1:1		training	0.86	1.26	12.54	0.52
		test	0.79	6.01	15.82	0.62
RBF 12:12-1-1:1		training	0.08	0.00	24.22	1.00
		test	-0.05	4.84	25.96	1.01

<sup>a</sup> PTML-ANN model profiles indicates:  $N_{iv}:N_i-N_{h1}-N_{h2}-N_o:N_{ov}$ ,  $N_{iv}$  = Number of Input Variables,  $N_i$  = Number of Input neurons,  $N_{h1}$  = Number of neurons in first hidden layer,  $N_{h2}$  = Number of neurons in second hidden layer,  $N_o$  = Number of output neurons,  $N_{ov}$  = Number of output variables. <sup>b</sup> R = Regression coefficient for 10000 pairs of query vs. reference reactions.

#### 4. CONCLUSION



1  
2  
3 In this work, we have shown that PTML models are useful for predicting the reactivity in  
4 Parham reactions. In fact, combining PT operators and ML algorithms resulted useful to account  
5 for changes in experimental conditions and/or the structural variables of all the molecules  
6 involved in the query reaction as compared to a reaction of reference. The predictions made with  
7 the model are statistically significant in terms of correlation with respect to the values from the  
8 literature and new experimental values reported here by the first time. Non-linear PTML models  
9 based on ANN do not outperformed PMTL linear models. This result confirms the linearity of  
10 the model. On the other hand, Hammett analysis showed that the effect of the substitution on the  
11 aromatic ring on the reactivity (yield predicted) is only moderate. Experimental chemists could  
12 use the model described for the selection of optimal conditions of reaction (T, t, *etc.*) out of a  
13 chemical space of more than  $10^8$  possible combinations. The model could also be used for the  
14 selection of the most efficient structures, specially the IE, for the application of this type of  
15 reaction as a key step in the synthesis of a target compound, reducing the experimental  
16 screening. In this area, Density Functional Theory (DFT) is one of the most used methods to  
17 study chemical reactivity.<sup>71</sup> However, these calculations are difficult when complex reaction  
18 networks with intermediates interconnected by different Transition States (TS) are studied. In  
19 this sense, computational automated protocols such as Nudged Elastic Band (NEB),<sup>72</sup> Growing  
20 string methods and linear synchronous transit,<sup>73</sup> or Global Reaction Route Mapping (GRRM),<sup>74</sup>  
21 have been tested. Compared to those methods, the main advantage of PTML models would be  
22 the possibility of predicting reactivity without relying upon the analysis of TS. Thus, they are  
23 useful to carry out large simulations without the need of high computing capacities. In fact, we  
24 report here a simulation of the reactivity space of Parham reactions (500000-point or more). This  
25 result also opens a new gateway to apply PTML methodology to other types of organic reactions  
26 with the consequent save of time, human and material resources.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12 ASSOCIATED CONTENT  
13  
14

**Supporting Information.** Synthesis and characterization of *N*-(*o*-iodobenzyl)pyrrolidines **S91-S92**; dataset of reactions (substrate, product, reaction conditions and yield,); structure of substrates and products in dataset (SMILE and compound code); structure of intermediates in dataset (SMILE and compound code); observed *vs.* predicted values for 10000 pairs of reactions; prediction of Yld(%) with multiple reactions of reference; Hammett analysis (pdf file); X Ray Analysis of **P64** (CCDC 1560347) (cif file). his material is available free of charge via the Internet at <http://pubs.acs.org>.

25  
26  
27  
28  
29 AUTHOR INFORMATION  
30  
3132 **Corresponding Authors**  
33

\*E-mail: [humberto.gonzalezdiaz@ehu.eus](mailto:humberto.gonzalezdiaz@ehu.eus)

\*E-mail: [nuria.sotomayor@ehu.eus](mailto:nuria.sotomayor@ehu.eus)

36  
37  
38  
39  
40  
41  
42  
43 ACKNOWLEDGMENT  
44

45  
46 Ministerio de Economía y Competitividad (FEDER CTQ2016-74881-P) and (CTQ2013-41229-  
47 P) and Gobierno Vasco (IT1045-16) are gratefully acknowledged for their financial support.  
48  
49 Technical and human support provided by Servicios Generales de Investigación SGIker  
50  
51 (UPV/EHU, MINECO, GV/EJ, ERDF and ESF) is also acknowledged. The authors also  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 acknowledge the kind preliminary revision and comments made by the editor Dr. Alexander  
4  
5 Tropsha.  
6  
7

## 8 ABBREVIATIONS

9  
10 ANN, Artificial Neural Networks; ArLi, Aryllithium Intermediate; CI, Cyclized Intermediate;  
11  
12 GLR, General Linear Regression; IE, Internal Electrophile; LHE, Lithium-Halogen Exchange;  
13  
14 ML, Machine Learning; PT, Perturbation Theory; PTML Perturbation-Theory Machine  
15  
16 Learning; Yld(%)<sub>obs</sub>, Yield of reaction observed (experimental); Yld(%)<sub>ref</sub>, Yield of a reaction of  
17  
18 reference (experimental); Yld(%)<sub>calc</sub>, Yield of a reaction calculated using one reference;  
19  
20 Yld(%)<sub>pred</sub>, Yield of a reaction predicted by the model with one or multiple references; Yld(%)<sub>res</sub>,  
21  
22 Yield of a reaction residual  
23  
24  
25  
26  
27

## 28 REFERENCES

- 29  
30 1. G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch,  
31 and A. Varnek. Expert System for Predicting Reaction Conditions: The Michael Reaction  
32 Case. *J. Chem. Inf. Model.* **2015**, 55, 239–250.  
33  
34  
35  
36 2. Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided  
37  
38 Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, 33,  
39  
40 469–476.  
41  
42 3. Sigman, M. S.; Miller, J. J. Examination of the Role of Taft-Type Steric Parameters in  
43  
44 Asymmetric Catalysis. *J. Org. Chem.* **2009**, 74, 7633-7643.  
45  
46  
47 4. Harper, K. C.; Sigman, M. S. Three-dimensional Correlation of Steric and Electronic  
48  
49 Free Energy Relationships Guides Asymmetric Propargylation. *Science* **2011**, 333, 1875-  
50  
51 1878.  
52  
53 5. Harper, K. C.; Sigman, M. S. Predicting and Optimizing Asymmetric Catalyst  
54  
55 Performance Using the Principles of Experimental Design and Steric Parameters. *Proc.*  
56  
57 *Natl. Acad. Sci. USA.* **2011**, 108, 2179-2183.  
58  
59  
60

- 1  
2  
3 6. Miller, J. J.; Sigman, M. S. Quantitatively Correlating the Effect of Ligand-Substituent  
4 Size in Asymmetric Catalysis Using Linear Free Energy Relationships. *Angew. Chem.*  
5 *Int. Ed.* **2008**, *47*, 771-774.  
6  
7
- 8  
9 7. Sigman, M. S.; Miller, S. Linear Free-Energy Relationship Analysis of a Catalytic  
10 Desymmetrization Reaction of a Diarylmethane-bis(phenol). *Org. Lett.* **2010**, *12*, 2794-  
11 2797.  
12  
13
- 14  
15 8. Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional Steric Parameters in the  
16 Analysis of Asymmetric Catalytic Reactions. *Nat. Chem.* **2012**, *4*, 366-374.  
17  
18
- 19  
20 9. Harper, K. C.; Vilardi, S. C.; Sigman, M. S. Prediction of Catalyst and Substrate  
21 Performance in the Enantioselective Propargylation of Aliphatic Ketones by a  
22 Multidimensional Model of Steric Effects. *J. Am. Chem. Soc.* **2013**, *135*, 2482-2485.  
23  
24
- 25  
26 10. Milo, A.; Bess, E. N.; Sigman, M. S. Interrogating Selectivity in Catalysis Using  
27 Molecular Vibrations. *Nature* **2014**, *507*, 210-214.  
28  
29
- 30  
31 11. Li, Y.; Li, H.; Pickard, F. C. T.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.;  
32 Sankaranarayanan, S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters  
33 from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492-4503.  
34  
35
- 36  
37 12. Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld,  
38 O. A.; Tkatchenko, A.; Muller, K. R. Assessment and Validation of Machine Learning  
39 Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**,  
40 *9*, 3404-3419.  
41  
42
- 43  
44 13. Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data  
45 to Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726-741.  
46  
47
- 48  
49 14. Müller, A. T.; Hiss, J. A.; Schneider, G. Recurrent Neural Network Model for  
50 Constructive Peptide Design. *J. Chem. Inf. Model.* **2018**, *58*, 472-479.  
51  
52
- 53  
54 15. Jaeger, S.; Fulle, S.; Turk, S. Mol2vec: Unsupervised Machine Learning Approach with  
55 Chemical Intuition. *J. Chem. Inf. Model.* **2018**, *58*, 27-35.  
56  
57  
58  
59

- 1  
2  
3 16. Turk, S.; Merget, B.; Rippmann, F.; Fulle, S. Coupling Matched Molecular Pairs with  
4 Machine Learning for Virtual Compound Optimization. *J. Chem. Inf. Model.* **2017**, *57*,  
5 3079-3085.  
6  
7  
8  
9 17. Xu, Y.; Pei, J.; Lai, L. Deep Learning Based Regression and Multiclass Models for Acute  
10 Oral Toxicity Prediction with Automatic Chemical Feature Extraction. *J. Chem. Inf.*  
11 *Model.* **2017**, *57*, 2672-2685.  
12  
13  
14  
15 18. Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle  
16 Electron Transfer Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 2413-2423.  
17  
18  
19  
20 19. Ramsundar, B.; Liu, B.; Wu, Z.; Verras, A.; Tudor, M.; Sheridan, R. P.; Pande, V. Is  
21 Multitask Deep Learning Practical for Pharma? *J. Chem. Inf. Model.* **2017**, *57*, 2068-  
22 2076.  
23  
24  
25  
26 20. Ericksen, S. S.; Wu, H.; Zhang, H.; Michael, L. A.; Newton, M. A.; Hoffmann, F.M.;  
27 Wildman, S. A. Machine Learning Consensus Scoring Improves Performance Across  
28 Targets in Structure-Based Virtual Screening. *J. Chem. Inf. Model.* **2017**, *57*, 1579-1590.  
29  
30  
31  
32 21. Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of  
33 Organic Chemistry Reactions. *ACS Cent. Sci.*, **2016**, *2*, 725–732.  
34  
35  
36  
37 22. Kayala, M. A.; Azencott, C. A., Chen, J. H., Baldi, P. Learning to Predict Chemical  
38 Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209-2222.  
39  
40  
41  
42 23. Coley, C. W.; Barzilay, R.; Jaakkola T. S.; Green, W. H.; Jensen, K. F. Prediction of  
43 Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434-443.  
44  
45  
46 24. Sadowski, P.; Fooshee, D.; Subrahmanya, N.; Baldi P. Synergies Between Quantum  
47 Mechanics and Machine Learning in Reaction Prediction. *J. Chem. Inf. Model.* **2016**, *56*,  
48 2125-2128.  
49  
50  
51  
52 25. Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel  
53 Fingerprint for Chemical Reactions and its Application to Large-Scale Reaction  
54 Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55*, 39-53.  
55  
56  
57  
58  
59  
60

- 1  
2  
3 26. Kayala, M. A.; Baldi, P. Reaction Predictor: Prediction of Complex Chemical Reactions  
4 at the Mechanistic Level Using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526-  
5 2540.  
6  
7  
8  
9 27. Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S., Gajewska, E. P.;  
10 Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via  
11 Machine Learning: Are Current Descriptors Sufficient? *Sci. Reports.* **2017**, *7*, 3582.  
12  
13  
14  
15 28. Muller, C.; Marcou, G.; Horvath, D.; Aires-de-Sousa, J.; Varnek, A. Models for  
16 Identification of Erroneous Atom-to-Atom Mapping of Reactions Performed by  
17 Automated Algorithms. *J. Chem. Inf. Model.* **2012**, *52*, 3116-3122.  
18  
19  
20  
21 29. Podolyan, Y.; Walters, M. A.; Karypis, G. Assessing Synthetic Accessibility of Chemical  
22 Compounds Using Machine Learning Methods. *J. Chem. Inf. Model.* **2010**, *50*, 979-991.  
23  
24  
25  
26 30. Ma, X.; Li, Z.; Achenie, L. E.; Xin, H. Machine-Learning-Augmented Chemisorption  
27 Model for CO<sub>2</sub> Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528-  
28 3533.  
29  
30  
31  
32 31. Carrera, G. V.; Gupta, S.; Aires-de-Sousa J. Machine Learning of Chemical Reactivity  
33 from Databases of Organic Reactions. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 419-429.  
34  
35  
36  
37 32. González-Díaz, H.; Arrasate, S.; Gómez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-  
38 Porto, L.; Ruso, J. M. General Theory for Multiple Input-Output Perturbations in  
39 Complex Molecular Systems. 1. Linear QSPR Electronegativity models in Physical,  
40 Organic, and Medicinal Chemistry. *Curr. Top. Med. Chem.* **2013**, *13*, 1713-1741.  
41  
42  
43  
44 33. González-Díaz, H.; Arrasate, S.; Gómez-SanJuan, A.; Sotomayor, N.; Lete, E.; Speck-  
45 Planche, A.; Ruso, J. M.; Luan, F.; Cordeiro, M. N. D. S. Matrix Trace Operators: From  
46 Spectral Moments of Molecular Graphs and Complex Networks to Perturbations in  
47 Synthetic Reactions, Micelle Nanoparticles, and Drug ADME Processes. *Curr. Drug*  
48 *Metabol.* **2014**, *15*, 470-488.  
49  
50  
51  
52  
53  
54 34. C. Blázquez-Barbadillo, E. Aranzamendi, E. Coya, E. Lete, N. Sotomayor, and H.  
55 González-Díaz. Perturbation Theory Model of Reactivity and Enantioselectivity of  
56  
57  
58  
59  
60

- 1  
2  
3 Palladium-catalyzed Heck-Heck cascade reactions. *RSC Advances* **2016**, *6*, 38602–  
4 38610.  
5  
6  
7  
8 35. Aranzamendi E, Arrasate S, Sotomayor N, González-Díaz H, Lete E. Chiral Brønsted  
9 Acid-Catalyzed Enantioselective  $\alpha$ -Amidoalkylation Reactions: A Joint Experimental and  
10 Predictive Study. *ChemistryOpen*. **2016**, *5*, 540-549.  
11  
12  
13  
14 36. Parham, W. E.; Bradsher, C. K. Aromatic Organolithium Reagents Bearing Electrophilic  
15 Groups. Preparation by Halogen-Lithium Exchange. *Acc. Chem. Res.* **1982**, *15*, 300–305.  
16  
17  
18 37. Sotomayor, N.; Lete, E. Aryl and Heteroaryllithium Compounds by Metal-Halogen  
19 Exchange. Synthesis of Carbocyclic and Heterocyclic Systems. *Curr. Org. Chem.* **2003**,  
20 *7*, 275-300.  
21  
22  
23  
24 38. Ruiz, J.; Ardeo, A.; Ignacio, R.; Sotomayor, N.; Lete, E. An Efficient Entry to  
25 Pyrrolo[1,2-*b*]isoquinolines and Related Systems through Parham Cyclization.  
26 *Tetrahedron* **2005**, *61*, 3311-3324.  
27  
28  
29  
30 39. Ruiz, J.; Lete, E.; Sotomayor, N. Intramolecular Cyclization of functionalized  
31 Heteroaryllithiums. Synthesis of Novel Indolizinone-based Compounds. *Tetrahedron*  
32 **2006**, *62*, 6182-6189.  
33  
34  
35  
36 40. Ardeo, A.; Lete, E.; Sotomayor, N. Metalation-Cyclisation Sequence on *N*-(*o*-  
37 Halobenzyl)pyrroles. Synthesis of Pyrrolo[1,2-*b*]isoquinolones. *Tetrahedron Lett.* **2000**,  
38 *41*, 5211-5214.  
39  
40  
41  
42 41. Sibi, M. P.; Shankaran, K.; Alo, B. I.; Hahn, W. R.; Snieckus, V. Overriding Normal  
43 Friedel-Crafts Regiochemistry in Cycliacylation. Regiospecific Carbodesilylation and  
44 Parham Cyclization routes to 7-Methoxy-1-indanols *Tetrahedron Lett.* **1987**, *28*, 2933-  
45 2936.  
46  
47  
48  
49  
50 42. Aidhen, I. S.; Ahuja, J. R. A Novel Synthesis of Benzocyclobutenones. *Tetrahedron Lett.*  
51 **1992**, *337*, 5431-5432.  
52  
53  
54  
55 43. Lear, Y.; Durst, T. Synthesis of Regiospecifically Substituted 2-  
56 Hydroxybenzocyclobutenones. *Can. J. Chem.* **1997**, *75*, 817-824.  
57  
58  
59  
60

- 1  
2  
3 44. Gould, S. J.; Melville, C. R.; Cone, M. C.; Chen, J.; Carney, J. R. Kinamycin  
4 Biosynthesis. Synthesis, Isolation, and Incorporation of Stealthin C, an  
5 Aminobenzo[*b*]fluorene. *J. Org. Chem.* **1997**, *62*, 320-324.  
6  
7  
8  
9 45. Paleo, M. R.; Castedo, L.; Domínguez, D. A New Synthesis of 4-Aryl-2-benzazepine-  
10 1,5-diones. *J. Org. Chem.* **1993**, *58*, 2763-2767.  
11  
12  
13 46. Paleo, M. R.; Lamas, C.; Castedo, L.; Domínguez, D. A New synthesis of Phthalides by  
14 Internal Trapping in ortho-Lithiated Carbamates Derived from Benzylic alcohols. *J. Org.*  
15 *Chem.* **1992**, *57*, 2029-2033.  
16  
17  
18  
19 47. Lamas, C.; Castedo, L.; Domínguez, D. Conversion of Dibenzoxepinones to  
20 Aristocularine Alkaloids. *Tetrahedron Lett.* **1990**, *31*, 6247-6248.  
21  
22  
23 48. Bracher, F. A Regioselective Synthesis of Azafluorenone Alkaloids. *Synlett* **1991**, 95-96.  
24  
25  
26 49. Poirier, M.; Chen, F.; Bernard, C.; Wong, Y.-S.; Wu, G. G. An Anion-Induced Regio-  
27 and Chemoselective Acylation and Its Application to the Synthesis of an Anticancer  
28 Agent. *Org. Lett.* **2001**, *3*, 3795-3798.  
29  
30  
31  
32 50. Gore, M. P.; Gould, S. J.; Weller, D. D. Total Synthesis of Phenanthroviridin Aglycon:  
33 the First Naturally-Occurring Benzo[*b*]phenanthridine. *J. Org. Chem.* **1991**, *56*, 2289-  
34 2292.  
35  
36  
37  
38 51. Moreau, A.; Couture, A.; Deniau, E.; Grandclaoudon, P.; Lebrun, S. First Total Synthesis  
39 of Cichorine and Zinnimidine. *Org. Biomol. Chem.* **2005**, *3*, 2305-2309.  
40  
41  
42  
43 52. Lamblin, M.; Couture, A.; Deniau, E.; Grandclaoudon, P. Alternative and Complementary  
44 Approaches to the Asymmetric Synthesis of C3 Substituted NH Free or *N*-Substituted  
45 Isoindolin-1-ones. *Tetrahedron Asymmetry* **2008**, *19*, 111-123.  
46  
47  
48  
49 53. Wang, Z.; Li, Z.; Wang, K.; Wang, Q. Efficient and Chirally Specific Synthesis of  
50 Phenanthro-Indolizidine Alkaloids by Parham-Type Cycloacylation. *Eur. J. Org. Chem.*  
51 **2010**, 292-299.  
52  
53  
54  
55  
56  
57  
58  
59  
60



- 1  
2  
3 54. Wang, C.; Wu, Z.; Wang, J.; Liu, J.; Yao, H.; Lin, A.; Xu, J. An Efficient Synthesis of 4-  
4 Isochromanones via Parham-type Cyclization with Weinreb Amide. *Tetrahedron* **2015**,  
5 *71*, 8172-8180.  
6  
7  
8  
9 55. Viña, D.; Uriarte, E.; Orallo, F.; González-Díaz H. Alignment-Free Prediction of a Drug-  
10 Target Complex Network Based on Parameters of Drug Connectivity and Protein  
11 Sequence of Receptors. *Mol. Pharm.* **2009**, *6*, 825-835.  
12  
13  
14  
15 56. Hill, T.; Lewicki, P. *Statistics: Methods and Applications, A Comprehensive Reference*  
16 *for Science, Industry, and Data Mining*, 1st Ed. Tulsa, OK, USA, StatSoft. Inc., 2006,  
17 830 pp.  
18  
19  
20  
21 57. Pratim Roy, P.; Paul, S.; Mitra, I.; Roy, K. On Two Novel Parameters for Validation of  
22 Predictive Qsar Models. *Molecules* **2009**, *14*, 1660-1701.  
23  
24  
25  
26 58. Golbraikh, A.; Tropsha, A. Beware of  $q^2$ ! *J. Mol. Graph. Model.* **2002**, *20*, 269-276.  
27  
28  
29 59. Ardeo, A.; Collado, M. I.; Osante, I.; Ruiz, J.; Sotomayor, N.; Lete, E. Recent Advances  
30 in the Parham Cyclization for the Synthesis of Heterocyclic Systems. *Targets in*  
31 *Heterocyclic Systems* **2001**, *5*, 393-418.  
32  
33  
34  
35 60. He, M.; Qu, C.; Gao, O.; Hu, X.; Hong, X. Biological and Pharmacological Activities of  
36 *Amaryllidaceae* alkaloids. *RSC Adv.* **2015**, *5*, 16562-16574.  
37  
38  
39 61. Wishler, M. C.; MacNiel, S.; Snieckus, V.; Beak, P. Beyond Thermodynamic Acidity: A  
40 Perspective on the Complex-Induced Proximity Effect (CIPE) in Deprotonation  
41 Reactions. *Angew. Chem. Int. Ed.* **2004**, *43*, 2206-2225.  
42  
43  
44  
45 62. Beak, P.; Meyers, A. I. Stereo- and Regiocontrol by Complex Induced Proximity Effects:  
46 Reactions of Organolithium Compounds. *Acc. Chem. Res.* **1986**, *19*, 356-363.  
47  
48  
49  
50 63. Hammett, L. P. *Physical Organic Chemistry*, 2nd edn., McGraw-Hill: New York, 1970.  
51  
52  
53 64. Hansch, C.; Leo, A.; Taft, R. W. Survey of Hammett Substituent Constants and  
54 Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165-195.  
55  
56  
57  
58  
59  
60

- 1  
2  
3 65. Santiago, C. B.; Milo, A.; Sigman M. S. Developing a Modern Approach to Account for  
4 Steric Effects in Hammett-type Correlations. *J Am. Chem. Soc.* **2016**, *138*, 13424-13430.  
5  
6  
7  
8 66. Bess, E. N.; De Luca, R. J.; Tindall, D. J.; Oderinde, M. S.; Roizen, J. L.; Du Bois, J.;  
9 Sigman, M. S. Analyzing Site selectivity in Rh<sub>2</sub>(esp)<sub>2</sub>-Catalyzed Intermolecular C-H  
10 Amination Reactions. *J Am. Chem Soc.* **2014**, *136*, 5783-5789.  
11  
12  
13  
14 67. Michel, B.W.; Steffens, L. D., Sigman, M. S. On the Mechanism of the Palladium-  
15 Catalyzed *tert*-Butylhydroperoxide-mediated Wacker-type Oxidation of Alkenes Using  
16 Quinoline-2-oxazoline Ligands. *J Am. Chem. Soc.* **2011**, *133*, 8317-8325.  
17  
18  
19  
20 68. Mueller, J. A.; Goller, C. P.; Sigman, M. S. Elucidating the Significance of β-Hydride  
21 Elimination and the Dynamic Role of Acid/Base Chemistry in a Palladium-Catalyzed  
22 Aerobic Oxidation of Alcohols. *J Am. Chem. Soc.* **2004**, *126*, 9724-9734.  
23  
24  
25  
26 69. Mueller, J. A.; Sigman M. S. Mechanistic Investigations of the Palladium-Catalyzed  
27 Aerobic Oxidative Kinetic Resolution of Secondary Alcohols Using (-)-Sparteine. *J Am.*  
28 *Chem. Soc.* **2003**, *125*, 7005-7013.  
29  
30  
31  
32 70. Mattalia, J.-M.; Nava, P. A Computational Study of the Intramolecular Carbolithiation of  
33 Aryllithiums: Solvent and Substituent Effects. *Eur. J. Org. Chem.* **2016**, 394-401.  
34  
35  
36  
37 71. Burke, K.; Werschnik, J.; Gross, E. K. U. Time-Dependent Density Functional Theory:  
38 Past, Present, and Future. *J. Chem. Phys.* **2005**, *123*, 062206.  
39  
40  
41  
42 72. Koistinen, O. P.; Dagbjartsdóttir, F. B.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H. Nudged  
43 Elastic Band Calculations Accelerated with Gaussian Process Regression. *J. Chem. Phys.*  
44 **2017**, *147*, 152720.  
45  
46  
47  
48 73. Behn, A.; Zimmerman, P. M.; Bell, A. T.; Head-Gordon, M. Incorporating Linear  
49 Synchronous Transit Interpolation into the Growing String Method: Algorithm and  
50 Applications. *J. Chem. Theory Comput.* **2011**, *7*, 4019-4025.  
51  
52  
53  
54 74. Ohno, K.; Maeda, S. Global Reaction Route Mapping on Potential Energy Surfaces of  
55 Formaldehyde, Formic Acid, and their Metal-Substituted Analogues. *J. Phys. Chem. A.*  
56 **2006**, *110*, 8933-8941.  
57  
58  
59  
60

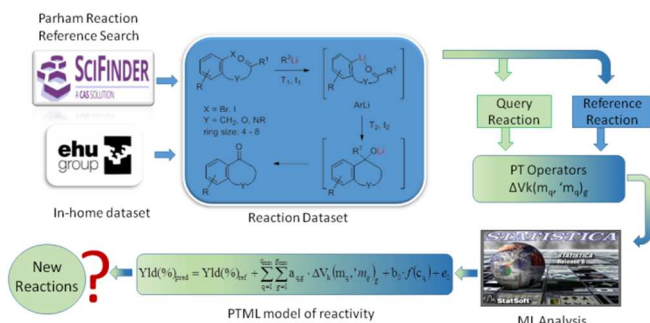
1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

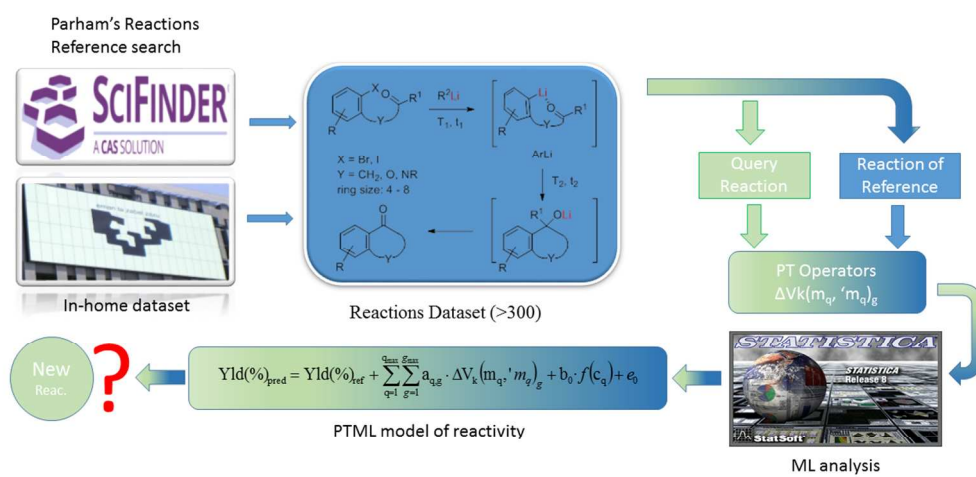
1  
2  
3  
4  
5  
6 **For Table of Contents use only**  
7  
8  
9  
10  
11

12 PTML: Perturbation-Theory and Machine Learning Model for High-Throughput Screening of  
13  
14 Parham Reactions. Experimental and Theoretical Studies  
15

16  
17 *Lorena Simón-Vidal,<sup>a</sup> Oihane García-Calvo,<sup>a</sup> Uxue Oteo,<sup>a</sup> Sonia Arrasate,<sup>a</sup> Esther Lete,<sup>a</sup> Nuria*  
18  
19 *Sotomayor,<sup>a,\*</sup> and Humberto González-Díaz<sup>a,b\*</sup>*  
20  
21  
22  
23  
24  
25  
26

27 **Graphical Abstract**  
28  
29





TOC

338x190mm (96 x 96 DPI)