Carracedo-Reboredo *et al.*
Journal of Cheminformatics (2024) 16:9
https://doi.org/10.1186/s13321-024-00802-7

Journal of Cheminformatics

RESEARCH

Open Access

MATEO: intermolecular α-amidoalkylation theoretical enantioselectivity optimization. Online tool for selection and design of chiral catalysts and products

Paula Carracedo-Reboredo[1,2], Eider Aranzamendi[1], Shan He[1,3], Sonia Arrasate[1], Cristian R. Munteanu[2], Carlos Fernandez-Lozano[2], Nuria Sotomayor[1*], Esther Lete[1*] and Humberto González-Díaz[1,4*]

Abstract

The enantioselective Brønsted acid-catalyzed α-amidoalkylation reaction is a useful procedure is for the production of new drugs and natural products. In this context, Chiral Phosphoric Acid (CPA) catalysts are versatile catalysts for this type of reactions. The selection and design of new CPA catalysts for different enantioselective reactions has a dual interest because new CPA catalysts (tools) and chiral drugs or materials (products) can be obtained. However, this process is difficult and time consuming if approached from an experimental trial and error perspective. In this work, an Heuristic Perturbation-Theory and Machine Learning (HPTML) algorithm was used to seek a predictive model for CPA catalysts performance in terms of enantioselectivity in α-amidoalkylation reactions with $R^2 = 0.96$ overall for training and validation series. It involved a Monte Carlo sampling of > 100,000 pairs of query and reference reactions. In addition, the computational and experimental investigation of a new set of intermolecular α-amidoalkylation reactions using BINOL-derived *N*-triflylphosphoramides as CPA catalysts is reported as a case of study. The model was implemented in a web server called MATEO: InterMolecular Amidoalkylation Theoretical Enantioselectivity Optimization, available online at: https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/mateo. This new user-friendly online computational tool would enable sustainable optimization of reaction conditions that could lead to the design of new CPA catalysts along with new organic synthesis products.

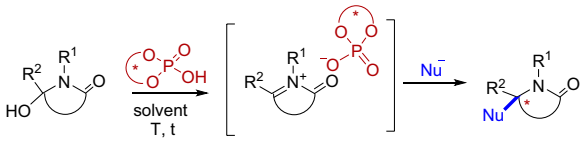
Keywords Chiral phosphoric acid catalysts, Cheminformatics, Machine learning, Amidoalkylation

*Correspondence:
Nuria Sotomayor
nuria.sotomayor@ehu.eus
Esther Lete
esther.lete@ehu.eus
Humberto González-Díaz
humberto.gonzalezdiaz@ehu.eus
[1] Department of Organic and Inorganic Chemistry, Faculty of Science and Technology, University of The Basque Country (UPV/EHU), P.O. Box 644, 48080 Bilbao, Spain
[2] Department of Computer Science and Information Technologies, Faculty of Computer Science, CITIC-Research Center of Information and Communication Technologies, University of A Coruña, Campus Elviña s/n, 15071 A Coruña, Spain
[3] IKERDATA S.L., ZITEK, University of Basque Country UPVEHU, Rectorate Building, 48940 Leioa, Spain
[4] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain

Introduction

Chiral Phosphoric Acid (CPA) and related catalysts are widely recognized and versatile tools in catalysis and organic synthesis useful for the synthesis of chiral drugs products [1–3]. The selection and design of new CPA catalysts for different enantioselective reactions has a dual interest because new CPA catalysts (tools) and chiral drugs or materials (products) can be obtained [4]. However, this process is difficult and time consuming if approached from an experimental trial and error perspective. Quantum Computational Chemistry tools may help to unravel the mechanism of reactions and help in the design of new CPA catalysts [5, 6]. Unfortunately, these techniques are less useful when it is necessary a fast scanning/optimization of new CPA catalysts for large libraries of reactions with diverse substrates, nucleophiles, products, and conditions (temperature, time, catalyst load, etc.). Cheminformatics methods relying upon Artificial Intelligence/Machine Learning (AI/ML) algorithms could help to speed up the discovery of new molecules [7–9] and in the design new chiral catalysts and products without engaging in a long term, empirical or quantum investigation [10–13]. Therefore, there is a need to develop fast-track computational tools able to predict the enantiomeric excess saving time and experimental resources. However, the application of AI/ML techniques to the study of enantioselective reactions is still uncommon due to the inherent complexity of the problem. In addition, most models are not implemented in public online web servers or they are not available for researchers or companies. In this context, it is remarkable Sigman's et al. platform for CPA catalysts and organophosphorous ligand design [14, 15]. In these works, the authors predict reactivity using structural information of the query reactants/products. However, useful experimental/operational conditions of already known reference reactions similar to the query reaction are not considered. Recently, our group has faced this problem by introducing the Perturbation-Theory and Machine Learning (PTML) approach that employs as inputs both vectors of structural variables $\mathbf{D}_{kqi}$ and vectors of multiple experimental conditions $\mathbf{c}_{qj}$. These PTML algorithms have been applied in medicinal chemistry, vaccine design, nanotechnology, and in catalysis as well [16–21]. In fact, we have previously reported a preliminary PTML model for the design of CPA catalysts for intermolecular α-amidoalkylation reactions [22]. However, the model was not implemented on a public online web server and is difficult to use by an experimentalist.

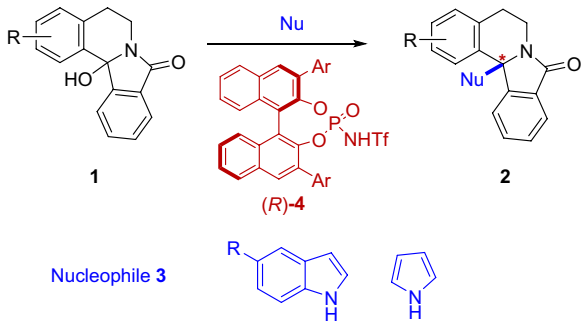
Consequently, in this work, we are going to focus on the development of a public web server for the selection and design of CPAs catalysts for enantioselective intermolecular α-amidoalkylation reactions (Scheme 1). In these reactions, the protonation of an α-hydroxylactam by the CPA would give a chiral conjugate base/*N*-acyliminium ion pair, which would be trapped by a nucleophile enantioselectively, generating a new tertiary or quaternary stereocenter [23, 24]. The α-amidoalkylation reaction of aromatic systems using *N*-acyliminium ions as electrophiles is a Friedel–Crafts-type reaction that has found widespread application in organic synthesis for the production of new drugs and natural products [25, 26]. For example, we have applied the procedure to the enantioselective synthesis of Nuevamine type alkaloids. Thus, indol and acyl moieties can be easily introduced in the alpha position of the nitrogen atom, using sterically demanding BINOL-derived CPA catalyst [27]. However, the enantioselectivity of these CPA catalyzed reactions is sensitive to many factors, from the nature of the nucleophile and the catalyst to the experimental conditions (solvent, temperature, etc.). In this context, many efforts have been made to understand the role of non-covalent interactions in organocatalyzed reactions and to rationalize and predict their stereochemical outcome using Quantum Chemical methods [28–30]. However, the chemical space accessible by organic synthesis is very wide, and all compatible combinations of substrate, nucleophile, catalyst, and solvent should have to be scanned.



Scheme 1 General scheme for CPA-catalyzed intermolecular α-amidoalkylation reactions

Therefore, the use of Cheminformatics models to explore the chemical space of these reactions becomes a very interesting option in order to reduce costs and time. Therefore, we decided to develop a new user-friendly online computational tool able to carry out screenings of this CPA-catalyzed intermolecular α-amidoalkylation reaction space for a large number of chiral catalysts, substrates, nucleophiles, solvents, chiral products, and reaction conditions. First, we carried out a re-evaluation of all the available data in our record to obtain a better estimate of the chemical space of these reactions. Next, we developed a new PTML model using Heuristics and Monte Carlo sampling calculations without relying on costly computational calculations. This PTML model was able to predict the enantioselectivity with $R^2 = 0.96$ after a comparative study 332 reactions, which can be paired in > 100,000 ways, as each reaction can be a query or reference reaction.

Later, we developed the web server called MATEO (interMolecular Amidoalkylation Theoretical



Scheme 2 Chiral BINOL-derived *N*-triflylphosphoramide-catalyzed intermolecular α-amidoalkylation reactions

Enantioselectivity Optimization), which is available at the online platform CPTMLTool (https://cptmltool.rnasa-imedir.com/). Finally, we have illustrated the practical use of the online tool with the experimental-theoretical study of a new set of CPA-catalyzed α-amidoalkylation reactions starting from bicyclic α-hydroxylactams **1** to construct the isoindoloisoquinoline framework **2** with a quaternary stereocenter. Electron-rich heteroaromatics (indole and pyrrole derivatives) **3** will be used as nucleophiles and chiral BINOL-derived *N*-triflylphosphoramides **4** as catalysts (Scheme 2). This new tool may help experimentalists in organic, medicinal, and materials chemistry to explore the chemical space of CPA-catalyzed α-amidoalkylation reactions and to optimize the reaction conditions for practical purposes.

Materials and methods

Dataset and parameter studied

In this paper, we have carried out the study of the enantiomeric excess $ee_R(\%)_{obs}$ parameter in intermolecular α-amidoalkylation reactions. The value $ee_R(\%)_{obs}$ allows to quantify the enantiomeric excess by applying an (*R*)-catalyst. This parameter is represented as $ee_R(\%)_{obs}$ = Sign (Prod)·Sign(Cat*R*)·$ee(\%)_{obs}$, where Sign(Prod) = 1 for (*R*)-product or Sign(Prod) = − 1 for (*S*)-product, taking into account an *R* or *S* notation of products experimentally obtained consistent with the Cahn-Ingold-Prelog (CIP) rules [31]. The function Sign(Cat) = 1 for all reactions carried out with an (*R*)-catalyst, irrespective of the product obtained. On the other hand, the sign was switched from + 1 to Sign(Cat) = − 1 for the reactions carried out with (*S*)-catalyst and the sign Sign(Prod) was changed to the opposed. This operation transform (*S*)-catalyst reactions into (*R*)-catalyst reactions with the same absolute value of enantiomeric excess but opposed sign of $ee_R(\%)_{obs}$. All reactions are expected to give the same result but with inverse configuration when you change the chirality of the Catalyst. Consequently, all reactions were studied as if they have been performed using an (*R*)-catalyst keeping the (*R*)-catalyst when originally used or switching the signs of Sign(Prod) and Sign(Cat) for (*S*)-catalyst reactions. In practice, this procedure will allow us to omit the use of chiral molecular descriptors for substrates, products, catalysts, etc., because all the chirality information will be included in the $ee_R(\%)$ terms for the query or reference reactions (see next sections). In fact, the method worked properly in this specific case because all the reactions give products with only one stereogenic center. Consequently, we have all the chirality information necessary included in both sides of the equation without necessity of using chiral molecular descriptors.

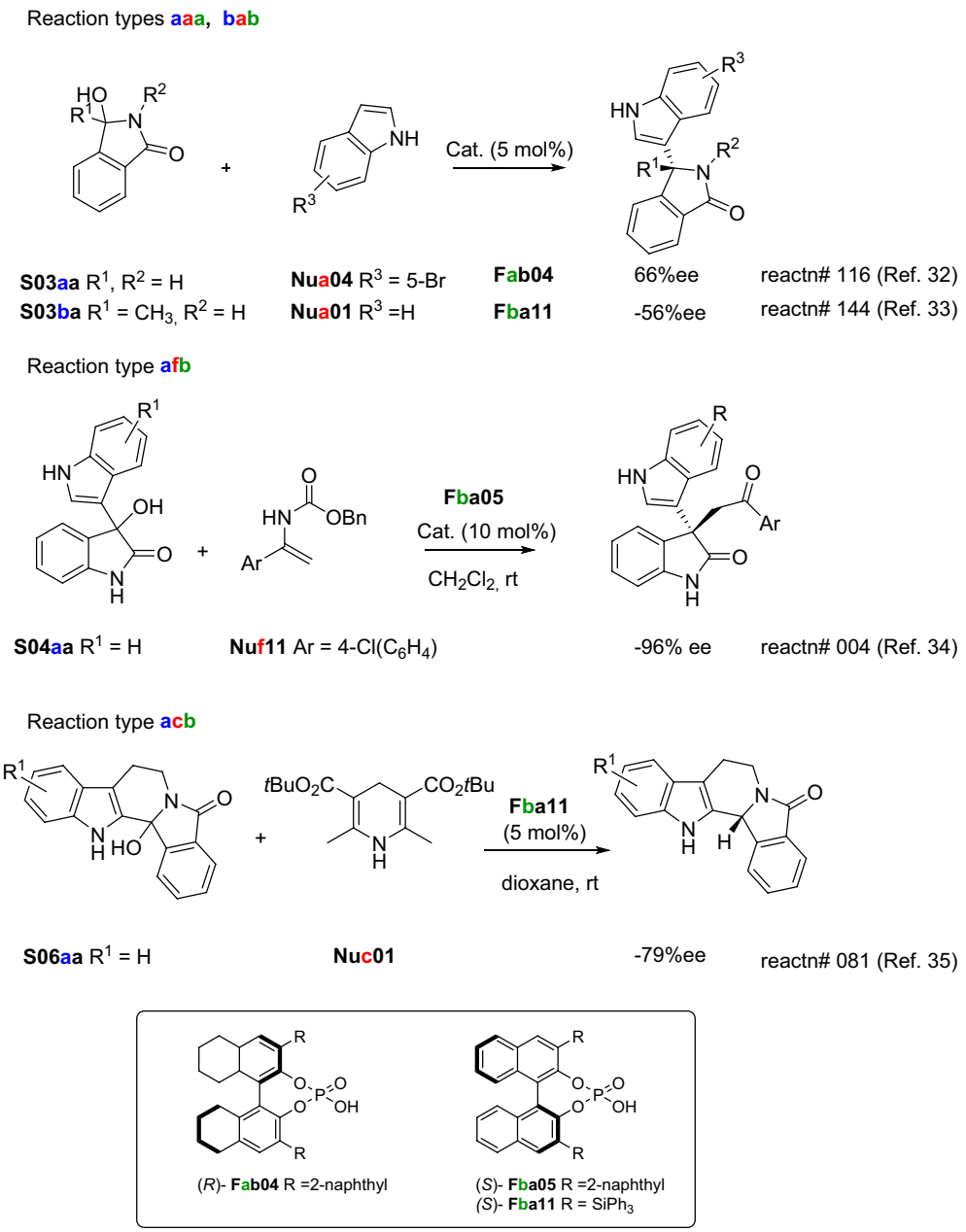
Reaction condition variables

Apart from defining the molecular descriptors, we also consider different reaction conditions variables $V_k(c_{qi})$ as input variables in order to quantify a k^{th} property (k = 1, 2, 3) related to a general reaction condition (c_q) and/or specific reactant. In this chemical reaction dataset, the variables taken into account for the i^{th} query reactions were: $V_1(c_{qi})$ = T(ºC) = Temperature, $V_2(c_{qi})$ = t(h) = reaction time and $V_3(c_{qi})$ = L(%) = catalyst loading. By analogy, the values of variables considered for each j^{th} reference reactions were: $V_1(c_{rj})$ = T(ºC) = Temperature, $V_2(c_{rj})$ = t(h) = reaction time, and $V_3(c_{rj})$ = L(%) = catalyst loading.

Dataset studied, compounds and reactions notation

A dataset of 332 CPA-catalyzed enantioselective intermolecular α-amidoalkylation reactions has been compiled, which comprised 324 reactions obtained from literature (see Additional file 3) and 8 new reactions studied in this work for the first time (see Table 8). These reactions have been grouped into 34 families according to the different structural patterns of the substrates, nucleophiles, and catalysts. There are different types of substrates **S** (mostly cyclic and bicyclic α-hydroxylactams, but also 3-hydroxyindolines) that are reacted with different types of nucleophiles **Nu** (indoles, pyrroles, Hantzsch esters, enols and enamides) using CPAs (phosphoric acids or the corresponding *N*-triflylphosphoramides and sulfonamides) as catalysts **Cat**.

All compounds have been labeled with a 5-element code Xyznn, X = S for Substrates, X = Nu for Nucleophiles, and X = Family of Catalysts; y = is the structural family (a, b, c,…), z = is the structural sub-family, if any (a, b, c, …), and nn = is the ID number of the compound in the dataset. When the structural sub-family is missing, the label y in the notation is omitted. Then, a code was created to classify each reaction in the dataset into different reactions types based on the structure of the molecules involved. Thus, the values of the family label y of the



Scheme 3. Selected examples of intermolecular α-amidoalkylation reactions included in the reference dataset, including molecule coding and reaction number (for Additional file 3)

Substrate, Nucleophile, and Catalyst were concatenated in this order to obtain the ID code of each reaction type. For example, the reaction of the Substrate **S03aa** with the Nucleophile **Nua04** and the Catalyst **Fab04** belongs to the reaction type with the ID code **aaa.** Scheme 3 shows selected examples of different reaction types included in the dataset using different types of cyclic hydroxylactams as substrates (**S03**, **S04**, **S06**) and different nucleophiles, such indoles (**Nua**) [32, 33] enamides (**Nuf**) [34] or Hantzsch esters as reducing agents (**Nuc**) [35], with CPAs catalysts (**F**). The full experimental detail of each of the 324 reference reactions (substrate, nucleophile, catalysts, catalyst loading product, solvent, temperature, time, yield, % ee) is included in the Supporting Information (Additional file 3), which also includes the SMILE code of the substrate, nucleophile and catalyst in each case. To have a general view of the chemical space in the dataset, general schemes for all reactions included in the

reference dataset are included in the Supporting Information (Additional file 1: Schemes S1 to S9). The structures and codification of substrates (**S**), nucleophiles (**Nu**), and catalysts (**cat.**) is included in the Supporting Information (Additional file 1).

Molecular descriptors calculation

First, the web tool MMDcalc was used to calculate the molecular descriptors $D_k(m_{sqi})_g$ and $D_k(m_{sri})_g$ of the molecules m_{sqi} and m_{sri} involved in the query and reference reactions [36]. The MMDcalc tool is an online web server available at the PTMLTool platform (https://cptmltool.rnasa-imedir.com/) for public use. This tool implements the Markov Chain Invariants for Networks Simulation and Design (MARCH-INSIDE) algorithm online. MARCH algorithm uses Markov Chains to calculate the average value of different atomic properties. These average values of atomic properties are calculated for predefined groups of atoms (g) inside the molecule and all their neighbors placed at topological distance (d). In the notation $D_k(m_{sqi})_g/D_k(m_{sri})_g$ the letter D = Descriptor, k = type of descriptor, s = sub-type of molecule, q = molecules involved in query reaction, r = molecules involved in reference reaction, i = ID number of the molecule, g = group of atoms inside the molecule. The general formula for the calculation is shown in Eq. 1 (see MARCH-INSIDE algorithm details in literature) [37].

$$D_k(m_{sqi})_g = \frac{1}{d_{max}} \sum_{d=1}^{d_{max}} \sum_{a \in g}^{\forall a \in g} D_d(m_{sqi})_a = \langle D_d(m_{sqi})_a \rangle \tag{1}$$

The k^{th} types (k = 1, 2, 3, 4, and 5) of molecular descriptors are: D_1 = Number of Valence Electrons (Zv), D_2 = van der Waals Volume (Vvdw), D_3 = Sanderson Electronegativity (χ), D_4 = Polarizability (α), and D_5 = Electron Affinity (EA). The sub-types (s) of query molecules m_{sqi}(s = 1, 2, 3, 4, and 5) are: m_{1qi} = Substrate$_{qi}$, m_{2qi} = Nucleofile$_{qi}$, m_{3qi} = Catalyst$_{qi}$, m_{4qi} = Solvent$_{qi}$, and m_{5qi} = Product$_{qi}$. The chemical functional groups or atom groups G_g (g = 1, 2, 3, 4, 5) are the following: G_1 = Saturated Carbon atoms (C_{sat}), G_2 = Unsaturated Carbon atoms (C_{uns}), G_3 = Heteroatoms (Het), G_4 = NonHalogen (X) Heteroatoms (HetNoX), and G_5 = Total (Tot). The groups of atoms indicate which atoms in the molecules were used as the basis for calculating the different local (g < 5) and/or total (g = 5) molecular descriptors.

ML linear model

In this section, $D_k(m_{sqi})_g$ values were introduced in order to look for a linear ML model. It is worth mentioning that each entry line of the dataset denotes only one query reaction (R_{qi}). The enantiomeric excess $ee_R(\%)_{qicalc}$ of the query reaction (R_{qi}) was predicted by applying both variables $V_k(c_{qi})$ as input depending on the experimental conditions and the molecular descriptors $D_k(m_{sqi})_g$ of the molecules taken into consideration in the reaction. With both sets of variables as inputs, we can seek a linear AI/ML additive model. A best practice, the following equality holds $ee_R(\%)_{calcqi} \approx ee_R(\%)_{qiobs,}$ when the additive linear hypothesis is correct. The general additive form of AI/ML model to be developed is the following.

$$ee_R(\%)_{calcqi} = \sum_{k=1}^{k_{max}} \sum_{s=1}^{s_{max}} a_{k,s} \cdot V_k(c_{qi}) + \sum_{k=1}^{k_{max}} \sum_{s=1}^{s_{max}} \sum_{g=1}^{g_{max}} b_{k,s,g} \cdot D_k(m_{sqi})_g + e_0 \tag{2}$$

PTML linear model

The PTML model is a well-known approach that can be used to predict the reactivity of a new case (reaction) through making comparisons with other known reactions. Our model can provide as output the $ee_R(\%)_{calcqi}$. On the other hand, the $ee_R(\%)_{calcqi}$ is calculated for a query reaction(R_{qi}) due to the observed enantiomeric excess $ee_R(\%)_{rjobs} = ee_R(\%)_{refj}$ of a reaction (R_{rj}) used as reaction of reference is already known. For this reason, the dataset applied to train/validate the PTML model, each entry line takes into consideration a pair of reactions, specifically a query reaction compared to a reference reaction (R_{qi} *vs.* R_{rj}). The PTML linear model enables to predict $ee_R(\%)_{calci}$ starting with the experimental value of $ee_R(\%)_{refj}$ of a reference reaction. Afterwards, the model includes the influences of different structural, operational or experimental conditions variations (perturbations) in the query in regard to the reference reaction. We use PT Operators (PTOs) in order to quantify these variations or perturbations. The parameter of PTOs are denoted as the form $\Delta D_k(m_{sqi}, m_{srj})_g$ for structural variations and $\Delta V_k(c_{qi}, c_{rj})$ for variations in the experimental reactions conditions. The formula of the PTML models used in this section are shown in Eqs. 3 and 4;

$$ee_R(\%)_{calcqi} = ee_R(\%)_{refj} + \sum_{k=1}^{k_{max}} \sum_{i=1}^{i_{max}} a_{k,i} \cdot \Delta V_k(c_{qi}, c_{rj}) + \sum_{k=1}^{k_{max}} \sum_{s=1}^{s_{max}} \sum_{g=1}^{g_{max}} b_k \cdot \Delta D_k(m_{sqi}, m_{srj})_g + e_0 \tag{3}$$

$$ee_R(\%)_{calcqi} = ee_R(\%)_{refj} + \sum_{k=1}^{k_{max}} \sum_{i=1}^{i_{max}} a_{k,i} \cdot \left[V_k\left(c_{qi}\right) - V_k\left(c_{rj}\right) \right] + \sum_{k=1}^{k_{max}} \sum_{i=1}^{i_{max}} \sum_{g=1}^{g_{max}} b_k \cdot \left[D_k(m_{sqi})_g - D_k(m_{srj})_g \right] + e_0 \quad (4)$$

In this work, the linear additive model used as a function of reference $ee_R(\%)_{robs}$ and two sets of PTOs represented by $\Delta V(c_{qi}, c_{rj})$ and $\Delta D(m_{sqi}, m_{srj})_g$ as input. The function of reference $ee_R(\%)_{robs}$ is equal to the observed values of enantiomeric excess *ee*(%), when the reference reaction used a (*R*)-catalyst with *R* configuration. We have developed two types of PTO in order to seek the PTML linear model. On the one hand, the first type of PTO is described as $\Delta V_k(c_{qi}, c_{rj}) = [V_k(c_{qi}) - V_k(c_{rj})]$. It takes into account the perturbations/deviations in the values of the k^{th} variables/conditions of reactions $V(c_{qi})$ of the q^{th} query reaction against the original values of the same variables $V_k(c_r)$ for the r^{th} reaction of reference. On the other hand, the second type of PTO is denoted as: $\Delta D_k(m_{sqi}, m_{srj}) = [D_k(m_{sqi}) - D_k(m_{srj})]_g$. It considers the perturbations/deviations in the values of the molecular descriptors of the query with respect to the reference molecules. Subsequently, the input variables for the reaction of the reference $V_k(c_{rj})$ are related to a k^{th} property (k = 1, 2, 3). The connection between the input variables and k^{th} property enables the connection in terms of general experimental conditions of reaction (c_{rj}) and/or specific reactants: $V_1(c_{rj}) = T(^oC) = Temperature$, $V_2(c_{rj}) = t(h) = reaction\ time$, and $V_3(c_{rj}) = L(\%) = catalyst\ loading$, for the reaction of reference (R_{rj}). The input variables denoted as $D_k(m_{ri})_g$ are the molecular descriptors of type k^{th} for the i^{th} molecules (m_{sri}) of type q^{th} involved in the reference reaction (R_{rj}). Analogously, the molecules m_{ri} taken part in the reaction of reference are $m_{r1j} = Substrate_{rj}$, $m_{r2j} = Nucleofile_j$, $m_{r3j} = Catalyst_{rj}$, and $m_{4rj} = Solvent_{rj}$. In addition, we use the k^{th} types of molecular descriptors as the same way as for the query reaction $D_1 = Number\ of\ Valence\ Electrons\ (Zv)$, $D_2 = Van\ der\ Waals\ Volume\ (Vvdw)$, $D_3 = Sanderson\ Electronegativity\ (\chi)$, $D_4 = Polarizability\ (\alpha)$, and $D_5 = Electron\ Affinity\ (EA)$. In Table 1, we illustrate the detailed information about of all the PTOs used as input variables in the PTML models.

AI/ML *vs.* PTML linear model development

So as to seek the AI/ML and PTML linear models, we apply Multivariate Linear Regression (MLR) and Linear Neural Network (LNN) algorithms by using the software STATISTICA [38]. In this sense, in the PTML regression models, the values of observed (experimental) enantiomeric excess $ee_R(\%)_{obsqi}$ against multiple values of reference $ee_R(\%)_{refj}$ have to be fitted. The regression model allows to generate artifacts in the standard distribution of the data [39]. The parameters $a_{k,s}$ $b_{k,s,g}$ and e_0 are the coefficients of the model to be fitted by AI/ML algorithms. The formula for the PTML linear regression models was fitted as presented in the Eq. 5;

$$\Delta ee_R(\%)_{qi} = \sum_{k=1}^{k_{max}} \sum_{s=1}^{s_{max}} a_{k,s} \cdot \Delta V_k(c_{qi}, c_{rj}) + \sum_{k=1}^{k_{max}} \sum_{s=1}^{s_{max}} \sum_{g=1}^{g_{max}} b_{k,s,g} \cdot \Delta D_k(m_{sqi}, m_{srj})_g + e_0 \quad (5)$$

Table 1 Definition of variables used as inputs of the PTML model

Experimental conditions (c_q)	Perturbation operators[b]	Type of operator
Reaction temperature (T)	$\Delta V(T) = \Delta T = T_q - T_r$	Temperature deviation
Reaction time (t)	$\Delta V(t) = \Delta t = t_q - t_r$	Time deviation
Catalyst loading [Load (%)]	$\Delta V(Load(\%)) = Load(\%)_q - Load(\%)_r$	Conc. difference
Molecules (m_q)[a]	Perturbation terms	Type of operator[a]
Substrate (Sub)	$\Delta D_k(Sub_{qi}, Sub_{rj})_g = [D_k(Sub_{qi})_g - D_k(Sub_{rj})]_g$	Structural variation
Product (Prod)	$\Delta D_k(Prod_{qi}, Prod_{rj})_g = [D_k(Prod_{qi})_g - D_k(Prod_{rj})]_g$	
Nucleophile (Nuc)	$\Delta D_k(Nuc_{qi}, Nuc_{rj})_g = [D_k(Nuc_{qi})_g - D_k(Nuc_{rj})]_g$	
Catalyst (Cat)	$\Delta D_k(Cat_{qi}, Cat_{rj})_g = [D_k(Cat_{qi})_g - D_k(Cat_{rj})]_g$	
Solvent (Solv)	$\Delta D_k(Solv_{qi}, Solv_{rj})_g = [D_k(Solv_{qi})_g - D_k(Solv_{rj})]_g$	

[a] Molecules (m) involved in the reaction with distinguishable roles: $m_{qsi} = Substrate\ (Sub_q)$, $Product\ (Prod_q)$, $Nucleophile\ (Nuc_q)$, $Catalyst\ (Cat_q)$, and $Solvent\ (Solv_q)$

[b] PTOs with formula $\Delta V(m_q, m_r)_g = [V(m_q)_g - V(m_r)]_g$. These PTOs measure the variation of the value of the molecular property/structural variable (V) in the query molecules m_q with respect to the value for molecule m_r with the same role in the reaction of reference. The values of $V_k(m_q)_g$ are average values of the properties $V_k = Sanderson\ Electronegativities\ (\chi)$, Polarizabilities, etc., for all the atoms in the group g and all their neighboring atoms placed at a topological distance $k \leq 5$. Consequently, these properties have been calculated for all the atoms in the molecule (Tot) or for subsets of atoms (group g). The groups of atoms studied are $g = unsaturated\ carbons\ (C_{uns})$, $saturated\ carbons\ (C_{sat})$, Heteroatoms (Het), Heteroatoms non-Halogen (HetNoX)

HPTML linear model

The PTML linear model built can predict diverse outputs for the same reaction taking into consideration the selected reference reactions. Therefore, in this section we introduced different Heuristics (H) in order to define the best reaction performance or set of reactions as reference. In this work, specifically we used two following heuristic. On the one hand, the first heuristic (H_1) can calculate the final predicted value as this form: $ee_R(\%)_{qrpred} = ee_R(\%)_{qrmin}$. This value is obtained using as reference the reaction with a minimum (Min) value of the PTOs in other words, the minimal deviation. Specifically, the heuristic (H_1) uses as reference, the reaction with a minimal difference/deviation (Δ) between the input variables $\Delta V(m_{qsi}, m_{rsj})$ and $\Delta V(c_{qi}, c_{rj})$ for all ($\forall$) pairs of reactions. On the other hand, the second heuristic (H_2) can calculate the value $ee_R(\%)_{qrpred} = ee_R(\%)_{qravg} = Avg(ee_R(\%)_{qrcalc})$. Particularly, the heuristic (H_2) uses as reference the values of variables $\Delta D(m_{qi}, m_{rj})$ (molecule structural variations) and $\Delta V(c_{qi}, c_{rj})$ (experimental conditions variations) for all ($\forall$) pairs of reactions. As the first step, we calculated the 331 different $ee_R(\%)_{qrcal}$ values, not including the query. Then, we obtained the final values as the average for all the references. These two heuristics can be described as illustrated in Eqs. 6 and 7.

$$H_1 : ee_R(\%)_{\text{qrpred}} = ee_R(\%)_{\text{qrmin}} \overset{\forall q,r}{\Rightarrow} Min\left\{PTOs\left[\Delta V\left(m_{qi}, m_{rj}\right), \Delta V\left(c_{qi}, c_{rj}\right)\right]\right\} \tag{6}$$

$$H_2 : ee_R(\%)_{\text{qrpred}} = ee_R(\%)_{\text{qravg}} \overset{\forall q,r}{\Rightarrow} Avg\left\{PTOs[(\Delta V\left(m_{qi}, m_{rj}\right), \Delta V\left(c_{qi}, c_{rj}\right)]\right\} \tag{7}$$

Monte carlo simulation

Most reactivity prediction models already reported take into consideration only the structure of the reactants but omit the values of temperature, catalyst loading, time of reaction, solvent polarity, etc. when predicting the enantiomeric excess of the reactions. In fact, many of the works focus only on yield at specific conditions of T, time, load, etc., and do not predict the enantiomeric excess. In addition, the values of enantiomeric excess, T, time, load, solvent polarity, etc. when measured experimentally contains a certain degree of error because most researchers do not measured them for triplicate or lead them uncontrolled like when using room temperature conditions. In this context, the Monte Carlo Simulation (MC) starts with the original values of the non-structural variables T, t, Load and using a random generator creates new values with small variations with respect to the original values. MC experiments are a wide-ranging class of computational algorithms that base on repeated random sampling to obtain numerical results. This method are among the most useful data sampling in Cheminformatics [40–42].

In this work, we used an MC algorithm to predict the enantiomeric excess of the reactions taking into consideration all these factors, which are of the major relevance to optimize the reaction in the laboratory. In order to demonstrate the robustness of the model we generated a new set of reactions with "perturbations" in the values of T, t, Load, etc. and retrained the models. The values of the values of T, t, Load, where changed randomly but inside the limits of min and max reported for this reactions. This allowed to test the robustness of the model in terms of ability of the model to continue working properly (giving good predictions) despite of changes/errors etc. in the reports of temperature, time, etc.

For this purpose, we generated a new set of reactions with "perturbations" in the values of T (°C), t(h), Load(%), etc. and retrained the models. The values of T (°C), t(h), Load(%) where changed randomly between the limits set in the minimum $V_k(c_{qi})_{min}$ and maximum $V_k(c_{qi})_{max}$ reported for this type of reactions. The synthetic data allow to test the robustness of the PTML model in terms of ability to continue giving good predictions despite of changes/errors, etc. In addition, the values of minimum $V_k(c_{qi})_{min}$, maximum $V_k(c_{qi})_{max}$, and step $V_k(c_{qj})_{step}$ for all the operational conditions were calculated (Table 2).

Table 2 Summary of basic statistics for reactions in the dataset

Stat.[a]	Dataset reaction conditions (c_{qi})[b]		
	T (°C)	T (h)	Load (%)
N_{reacc}	12	53	7
Avg	11.59	35.87	9.10
S.D	26.80	33.94	5.72
Min	− 78.00	1.00	2.00
Max	66.00	240.00	30.00
Range	144	239	28
Step	10	1	1
N_{expr}	14	239	28

[a] Stat. = Statistical parameters for the input parameters (operational conditions) of all the reactions present in our dataset: N_{reacc} = Number of reactions present in our dataset, Avg. = average value, S.D. = Standard deviation, Max. = *maximum* value, Min. = *minimum* value, Range = Max.—Min., Step = minimal change allowed in one experimental condition, N_{expr} = Number of experiments (reactions) changing one condition and keeping the others constant

[b] Operational conditions: T(°C) = temperature, t(h) = reaction time, Load(%) = catalyst loading

Afterwards, we used a MC model based on the following system of equations in order to create the new synthetic data.

Firstly, the Eqs. 8 and 9 were applied so as to generate new $V_k(c_{qi})_{new}$ values starting from the original minimum value $V_k(c_{qi})_{min}$ (Eq. 8). Later, with the Eq. (9), we obtained the new synthetic data value $V_k(c_{qi})_{synth}$ after introducing a boundary condition. This boundary condition is set up taking into consideration the conditions of α-amidoalkylation reactions. In other words, the boundary condition keeps the synthetic values $V_k(c_{qi})_{synth}$ within the range $[V_k(c_{qi})_{min}, V_k(c_{qi})_{max}]$. The synthetics values were created for the experimental condition variables $V_1(c_{qi}) = T(°C)$, $V_2(c_{qi}) = t(h)$, $V_3(c_{qi}) = L(\%)$. It means that the new synthetic data values are equal to $V(c_k)_{synth} = V(c_k)_{min} + rnd(0, N_{max}) \cdot V(c_k)_{step}$ iff (if and only if) they are lower than $V_k(c_{qi})_{max}$; otherwise, they are equal to $V_k(c_{qi})_{max}$. The function $Rnd(0, n_{max})$ is a generator of pseudo-random natural numbers (n = 0, 1, 2, … N_{max}) based on Mersenne-Twister MC algorithm (MT19937). The same system of equations was used to form new synthetic data for the input variables of the reference $V_k(c_{rj})$ equation.

$$V_k\left(c_{qi}\right)_{new} = \left(V_k\left(c_{qi}\right)_{min} + Rnd(0, n_{max}) \cdot V_k\left(c_{qi}\right)_{step}\right) \tag{8}$$

$$V_k\left(c_{qi}\right)_{synth} = if\left[V_k\left(c_{qi}\right)_{new} > V_k\left(c_{qi}\right)_{max}; V_k\left(c_{qi}\right)_{max}; V_k\left(c_{qi}\right)_{new}\right] \tag{9}$$

As mentioned above, we have only made small random changes to the values of the input variables t, T, and catalyst loading from the original ones. Consequently, in the new synthetic data cases generated by MC, we assumed that the deviations in the new values of input variables (perturbations) from the original ones are small enough to cause unetectable/non-measurable changes in the output values of ee_R(%). The supposition is based on practical empiric evidence, which seems to confirm that new reactions/repetitions carried out with small changes of a few degrees of Temperature, minutes of reaction time, or catalyst loading will not alter i the value of *ee*R(%) by a measurable amount. In fact, in Eq. (8) the new synthetic value is equal to the minimum value in all the dataset plus the value of the step multiplied by a random value getting values 0, 1, 2, n_{max}.

Experimental methods

We describe here the typical procedure for the enantioselective intermolecular α-amidoalkylation reaction leading to the synthesis of (+)-**2e** (See Table 8, entry 8). For full experimental details and characterization data for compounds **2a-d**, See Supporting Information file SI00.pdf).

*(+)-(R)-2,3-dimethoxy-12b-(1H-pyrrol-2-yl)-5,12b-dihydroisoindolo[1,2-a]isoquinolin-8(6H)-one(**2e**).* A solution of 12b-hydroxyisoindoloisoquinoline **1** (60 mg, 0.19 mmol), pyrrole **3e** (0.014 mL, 0.19 mmol) and *N*-triflylphosphoramide **4a** (28 mg, 0.038 mmol 20 mol%) in dry THF (5 mL) were stirred during 5 h at room temperature. The solvent was evaporated under reduced pressure, and the crude reaction mixture was purified by flash column chromatography (alumina, Hexane/EtOAc 3:7) to afford isoindolo[1,2-*a*]isoquinoline **2e** (68 mg, quant.); $[\alpha]_D^{20} = +40.3$ (c = 0.28; CH_2Cl_2). The enantiomeric excess was determined by HPLC to be 54% [Chiralcel OD, 15% Hexane/2-propanol, 1 mL/min, t_R (*S*) = 23.2 min (22.87%), t_R (*R*) = 29.4 min (77.13%)]. m.p. (Hexane/EtOAc): 254–256 °C; IR (Film): 3188 (NH) cm^{-1}, 1672 (CO) cm^{-1}; ^{1}H NMR (300 MHz, $CDCl_3$): δ 2.70–2.76 (m, 1H), 3.06 (ddd, *J* = 17.3, 11.1, 6.5 Hz, 1H), 3.23 (ddd, *J* = 12.6, 11.1, 4.8 Hz, 1H), 3.85 (s, 3H), 3.87 (s, 3H), 4.26 (ddd, *J* = 12.6, 6.5, 2.2 Hz, 1H), 5.86–5.88 (m, 1H), 6.08 (dd, *J* = 5.8, 2.7 Hz, 1H), 6.62 (s, 1H), 6.74 (td, *J* = 2.7, 1.5 Hz, 1H), 7.23 (s, 1H), 7.44 (t, *J* = 7.5 Hz, 1H),7.58 (t, *J* = 7.5 Hz, 1H), 7.70–7.72 (m, 2H), 8.70 (s, 1H)ppm; ^{13}C[^{1}H] NMR (75.5 MHz, $CDCl_3$):δ 28.7, 35.2, 55.9, 56.2, 65.7, 108.1, 110.5, 110.8, 111.7, 119.0, 123.7, 123.9, 127.1, 127.9, 128.8, 131.5, 132.1, 147.1, 148.6, 148.9, 167.2 ppm; MS (CI) m/z (%): 361 (100) $[MH]^+$, 360 (50) $[M]^+$, 294 (37), 293 (33); HRMS (CI): cacld. for $C_{22}H_{21}N_2O_3$ $[MH]^+$: 361.1552; found: 361.1556.

Results and discussion

CPA catalyzed α-amidoalkylation reactions chemical space

As stated above, the chemical space of α-amidoalkylation reactions is very wide. In this work, the dataset is based on 332 reactions which contains 55 different substrates (cyclic and bicyclic hydroxylactams), 53 nucleophiles (enamides, indoles, etc.), 39 chiral catalysts (phosphoric acids, phosphoramides, etc.), and 17 different solvents undertaken by multiple experimental conditions (see Supporting Information, file SI00.pdf for structures and reaction schemes; see Additional file 3 for full details of each reference reaction, including reaction conditions, yield, enantiomeric excess, and SMILE codes for reactants and catalysts in each case). The combination of all possible substrates, catalysts, and reactions conditions to be explored is potentially high to be covered by trial and error experiments. To better understanding the amount of all possible combination, we illustrate an example, if reactions are run independently by changing one reactant at a time, a total of $N_{comb} = N(Subs_{qi}) \cdot N(Nuc_{qi}) \cdot N(C$

$at_{qi})\cdot N(Solv_{qi}) = 55\cdot 53\cdot 39\cdot 17 = 1{,}932{,}645$ unique combinations of molecule subtypes should be run. This could be a new source of interesting products [changes in $N(Subs_{qi})$ or $N(Nuc_{qi})$] or a way to improve the reaction efficiency [changes in $N(Cat_{qi})$ or $N(Solv_{qi})$]. This estimation considers only the combinations of different molecular entities. Unfortunately, the vast majority of these reactions remain unexplored in terms of high cost in time and resources.

On the other hand, there are also important variations in the three main experimental condition variables $V_k(c_{qi})$ [T(°C), t(h), and L(%)]. Table 2 shows different statistics parameters of these variables for the reported reactions. The integer values for maximum (T_{max}, t_{max}, and L_{max}), minimum (T_{min}, t_{min}, and L_{min}), and step (T_{step}, t_{step}, and L_{step}) are included. This is important because the expression Range $[V_k(c_{qi})] = V_k(c_{qi})_{max} - V_k(c_{qi})_{min}$ gives us the range of this variable that can be covered in actual practice in the laboratory. Consequently, when this range is divided by the minimum value, we decided to change in practice Step $[V_k(c_{qi})]$, the number of experiments $N(c_{qi}) = Range[V_k(c_{qi}))/Step(V_k(c_{qi})]$ that we can run in order to explore this variable can be obtained. When reactions are run independently by changing one experimental condition at a time, a total of N_{exp} experiments must be run. This will be equal to $N_{exp} = N(c_1)\cdot N(c_2)\cdot N(c_3) = N(T)\cdot N(t)\cdot N(L) = [Range(T)/Step(T)]\cdot[Range(t)/Step(t)]\cdot[Range(L)/Step(L)] = [144/10]\cdot[(239/1]\cdot[(28/1] = 96{,}365$ optimization experiments for each unique combination of molecule sub-types giving as result an specific $Product_{qi}$ of the reactions R_{qi} (Table 2). The multiplication of both parts of the equation gives an estimate of the very large number of reactions accessible in this chemical space $N(R_{qi})_{max} = N_{comb}\cdot N_{exp} \approx 10^{11}$.The equations used to carry out the calculations of the number of reactions in this chemical space are shown below (Eq. 10) [39]:

$$N(R_{qi})_{max} = N(Sub_{qi})\cdot N(Nuc_{qi})\cdot N(Cat_{qi}) \cdot N(Solv_{qi})\cdot \frac{Range(T)}{Step(T)}\cdot \frac{Range(t)}{Step(t)} \cdot \frac{Range(L)}{Step(L)} \quad (10)$$

$$N(R_{qi})_{max} = \prod_{s=1}^{s=4}\left[N(m_{sqi})\right]\cdot \prod_{k=1}^{k=3}\left[\frac{V_k(c_{qi})_{max} - V_k(c_{qi})_{min}}{Step(V_k\left(c_{qi}\right)_{max})}\right]$$

$$N(R_{qi})_{max} = \prod_{s=1}^{s=4}\left[N(m_{sqi})\right]\cdot \prod_{k=1}^{k=3}\left[N(c_{qi})\right]$$

$$N(R_{qi})_{max} = N_{comb}\cdot N_{exp}$$

The previous calculation gives an idea on the dimension of chemical reaction space for enantioselective CPA-catalyzed intermolecular α-amidoalkylation reactions. It is inviable to study all possible combinations in the laboratory due to the time and cost in material and human resources. In the daily practice, chemists can use expert criteria and experimental design techniques to reduce the number of combinations to be tested, to decrease the range of the different experimental conditions variables, etc. This can support researchers to reduce meaningfully the number of reactions to perform in the practice. However, the use of the previous well-known experimental expert criteria, researchers will never test interesting products. Therefore, the main objective of this project was the development of a new user-friendly predictive regression model for these reactions. This predictive model may become a useful tool to reduce the time and cost of experimentation.

ML linear model for α-amidoalkylation reactions

In the α-amidoalkylation reactions, there is no clear relationship between the chirality of the catalysts and the CIP notation of the product. In fact, in our literature dataset one can note the following ratio of Catalyst/Product chirality relationship, count, and ratio *(R)*/*(R)*140 reactions (43.2%), *(S)*/*(R)*102 reactions (31.5%), *(R)*/*(S)* 72 reactions (22.2%) and *(S)*/*(S)* 9 reactions (2.8%) of 324 reactions. There is only one reaction in the entire dataset with an *(S)*configuration catalyst and enantiomeric excess equal to zero. Therefore, it is very important to have a computational model to predict the absolute stereochemistry and the enantiomeric excess of the reaction product. This type of models could be used as a useful tool in order to address the design of new catalysts and/or selecting the optimal reaction conditions a priori. In this work, we decided to tackle this problem using AI/ML techniques. We trained this classic linear ML model using only the Original Data (OD) from reactions. The equation of this model is shown in Eq. 11;

Table 3 Results of the PTML regression model

Model	Input Vars[a]	Symbol	Coeff	$ee_R(\%)_{qr}$[b]	S.E.[c]	t[d]	p-level[e]
ML	$Load(\%)_{qr}$	$V_3(c_{qi})$	a_1	912.48	258.3259	3.53227	< 0.05
	$T(^oC)_{qr}$	$V_1(c_{qi})$	a_2	21.90	18.7647	1.16732	0.24
	$t(h)_{qr}$	$V_2(c_{qi})$	a_3	− 194.76	55.2274	− 3.52654	< 0.05
	$\alpha(Cat_{qi})_{Cuns}$	$D_4(m_{3qi})_2$	b_1	− 13.21	2.7499	− 4.80465	< 0.05
	$\alpha(Prod_{qi})_{HetNoX}$	$D_4(m_{5qi})_4$	b_2	− 45.02	21.8624	− 2.05905	< 0.05
	$EA(Prod_{qi})_{Csat}$	$D_5(m_{5qi})_1$	b_3	830.06	163.3526	5.08139	< 0.05
	$EA(Cat_{qi})_{HetNoX}$	$D_5(m_{3qi})_4$	b_4	− 0.34	0.0949	− 3.62574	< 0.05
	$\chi(Nuc_{qi})_{Het}$	$D_3(m_{2qi})_3$	b_5	0.22	0.0742	3.01949	< 0.05
	$\chi(Cat_{qi})_{HetNoX}$	$D_3(m_{3qi})_4$	b_6	− 2024.05	484.4448	− 4.17809	< 0.05
	$V(Sub_{qi})_{Tot}$	$D_2(m_{1qi})_5$	b_7	− 178.69	43.4355	− 4.11390	< 0.05
	$Zv(Cat_{qi})_{Cuns}$	$D_1(m_{3qi})_2$	b_8	− 1678.05	468.7747	− 3.57965	< 0.05
	$Zv(Solv_{qi})_{Cuns}$	$D_1(m_{4qi})_2$	b_9	− 34.41	11.6896	− 2.94399	< 0.05
	Intercept	–	e_0	− 0.70	0.5150	− 1.35948	0.18

[a] Input variables with coefficient b_k are the values of shift (Δ) in q-reac *vs*. r-reac for different properties: α = average atomic polarizability, EA = average atomic Electro Affinity, χ = average atomic Sanderson Electronegativity, Zv = average atomic number

[b] Coefficients of the variables in the model, the output variable is the Δ in enantiomeric excess *ee*(%)* of the q-reac with respect to the r-reac when both reactions have been carried out with(*R*)-catalyst

[c] Standard error of the coefficients

[d] Student t-value

[e] p-level of error

$$\begin{aligned} ee_R(\%)_{qpred} = & 912.48 \cdot Load(\%) + 21.90 \cdot T(^oC) - 194.76 \cdot t(h) \\ & - 13.21 \cdot \propto \left(Cat_{qi}\right)_{Cuns} - 45.02 \cdot \propto \left(Prod_q\right)_{HetNoX} \\ & + 830.06 \cdot \Delta EA\left(Prod_q\right)_{Csat} - 0.34 \cdot EA\left(Cat_{qi}\right)_{HetNoX} \\ & + 0.22 \cdot \chi\left(Nuc_{qi}\right)_{Het} - 2024.05 \cdot \chi\left(Cat_q\right)_{HetNoX} \\ & - 178.69 \cdot V\left(Sub_{qi}\right)_{Tot} - 1678.05 \cdot Zv\left(Cat_{qi}\right)_{Cuns} \\ & - 34.41 \cdot Zv\left(Solv_q\right)_{Cuns} - 0.70 \end{aligned} \tag{11}$$

$$n = 332(reactions) R^2 = 0.74 F = 59.2 p < 0.5$$

This ML model does not use reference reactions for comparison. The statistic parameters of the model are n = 332, Regression coefficient $R^2 = 0.74$, Fisher ratio F = 59.2, Standard Error of Estimates SEE = 37.1, p-level p < 0.05. More detailed information about coefficients and variables of the model as well as symbols and names of variables, Standard Error (SE), Students' t values, and p-level are given in Table 3. The model obtains 74.0% of variance (coefficient $R^2 = 0.74$), which is an acceptable prediction percentage for organic synthesis reactions (although extremely improbable). By the way, the SEE = 37.1 could be considered relatively high[39]. On the other hand, an essential short-coming of this classic ML linear model is that it does not provide us any evidence about the most similar reactions conveyed in the scientific literature. Consequently, this may limit our ability to deduce possible mechanisms and/or compare our results with others already known. Therefore, this ML model needs to be used along with another search strategy for similar molecules to obtain clues of similar reactions for a specific reaction under study. One option is to couple this model with similarity search strategies based on Tanimoto's similarity indices [43]. In fact, there are interesting works that report the coupling of Cheminformatics models with search strategies based on similarity [44–46]. A well-known example of online search tools is the Scifinder platform [47, 48].

PTML model for α-amidoalkylation reactions

As mentioned in the previous section, we have reported a PTML model for α-amidoalkylation reactions, although it is difficult to use in practice and not implemented on a publicly available online web server. Unfortunately, the input variables used in that model are not available as an open source code. For this reason, it could be advantageous to implement the model on a public online server. Consequently, we decided to develop a new linear PTML model using our own library to calculate the molecular descriptors. PTML reactivity models can study pair-wise reactions [39]. The model infers the reactivity of a query reaction (q) by comparing it to a previously known reference reaction (r). Some PTML models use different Heuristics (H) to match q and r reactions. These models can be called HPTML models. The Fig. 1 illustrates the general workflow that has been followed during this word to look for the new HPTML models. In step 1, the reference dataset and reaction pairs q *vs*. r were created. In step 2,

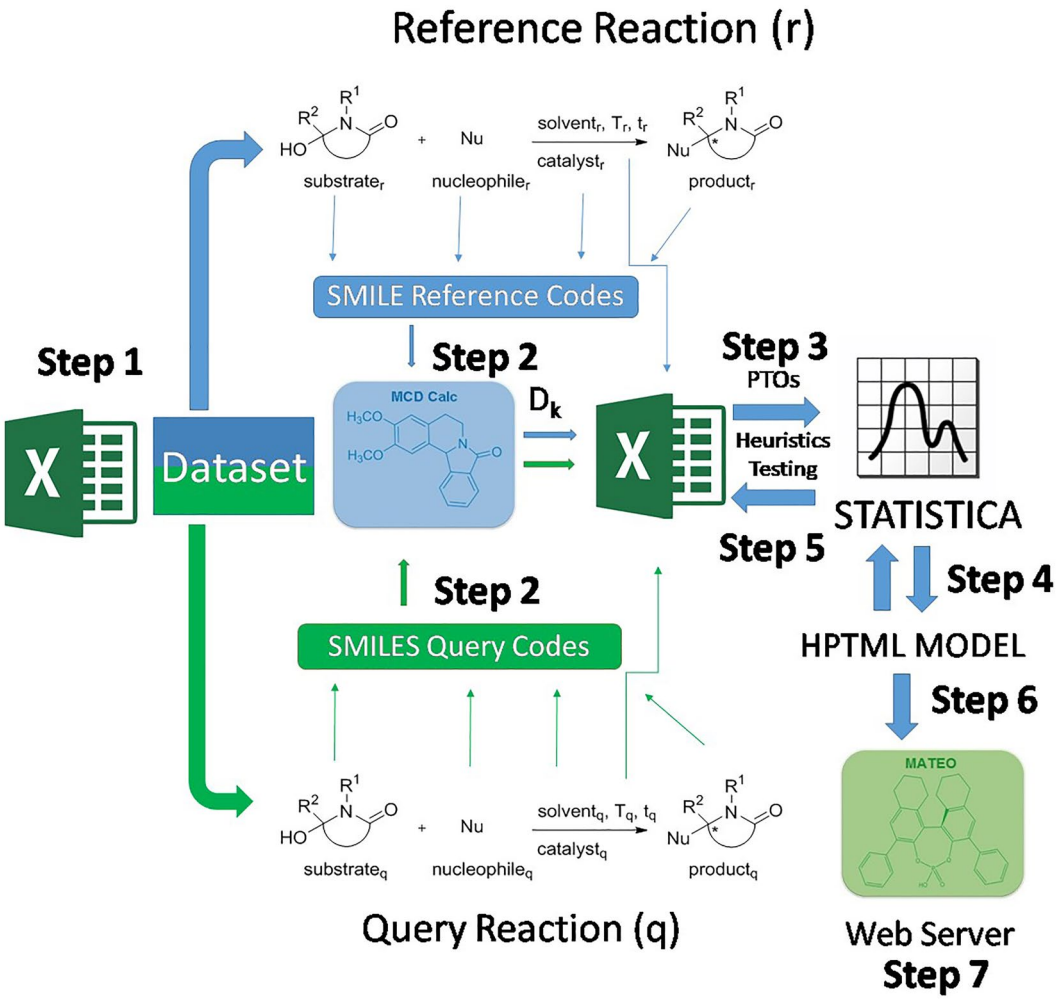


Fig. 1 HPTML models general workflow used in this work

the SMILE codes of the molecules (m_{qsi}, m_{rsj}) involved in both q and r reactions (substrates, nucleophiles, catalysts, solvents, products) were entered in the MCDCalc server [49] to calculate their molecular descriptors $D_k(m_{qsi})_g$ and $D_k(m_{rsj})_g$. In step 3, the PTOs for pairs of reactions were calculated. In step 4, the Multivariate Linear Regression (MLR) algorithm implemented in the STATISTICA [38] software was used to seek the PTML model. In step 5, heuristics H_1 and H_2 were tested interactively. In step 6, the best HPTML model was selected. Finally, in step 7, this model was implemented on a public web server (see the following sections). The best linear HPTML model found is shown in Eq. 12;

$$\begin{aligned}\Delta ee_R(\%)_{qr} = & -0.82 \cdot \Delta Load(\%) - 0.34 \cdot \Delta T(^oC) \\ & + 0.21 \cdot \Delta t(h) \\ & - 174.37 \cdot \Delta \propto \left(Cat_q, Cat_r\right)_{Cuns} \\ & - 1534.17 \cdot \Delta \propto \left(Prod_q, Prod_r\right)_{HetNoX} \\ & - 215.98 \cdot \Delta EA\left(Prod_q, Prod_r\right)_{Csat} \\ & - 1747.12 \cdot \Delta EA\left(Cat_q, Cat_r\right)_{HetNoX} \\ & - 42.49 \cdot \Delta \chi\left(Nuc_q, Nuc_r\right)_{Het} \\ & + 750.76 \cdot \Delta \chi\left(Cat_q, Cat_r\right)_{HetNoX} \\ & - 34.19 \cdot \Delta V\left(Sub_q, Sub_r\right)_{Tot} \\ & + 22.04 \cdot \Delta Zv\left(Cat_q, Cat_r\right)_{Cuns} \\ & - 12.46 \cdot \Delta Zv\left(Solv_q, Solv_r\right)_{Cuns} - 0.91 \end{aligned} \quad (12)$$

Table 4 Results of the PTML regression model

Model	Input Vars[a]	Symbol	Coeff.[b]	$\Delta ee_R(\%)_{qr}$[b]	S.E.[c]	t[d]	p-level[e]
PMTL	$\Delta Load(\%)_{qr}$	$\Delta V_3(c_{qi}, c_{rj})$	a_1	− 0.82	0.03243	− 25.3154	< 0.005
	$\Delta T(oC)_{qr}$	$\Delta V_1(c_{qi}, c_{rj})$	a_2	− 0.34	0.00594	− 57.8960	< 0.005
	$\Delta t(h)_{qr}$	$\Delta V_2(c_{qi}, c_{rj})$	a_3	0.21	0.00476	44.4627	< 0.005
	$\alpha(Cat_q, Cat_r)_{Cuns}$	$\Delta D_4(m_{3qi}, m_{3rj})_2$	b_1	− 174.37	2.67954	− 65.0741	< 0.005
	$\alpha(Prod_q, Prod_r)_{HetNoX}$	$\Delta D_4(m_{5qi}, m_{5rj})_4$	b_2	− 1534.17	26.38185	− 58.1525	< 0.005
	$EA(Prod_q, Prod_r)_{Csat}$	$\Delta D_5(m_{5qi}, m_{5rj})_1$	b_3	− 215.98	3.38484	− 63.8086	< 0.005
	$EA(Cat_q, Cat_r)_{HetNoX}$	$\Delta D_5(m_{3qi}, m_{3rj})_4$	b_4	− 1747.12	26.48292	− 65.9715	< 0.005
	$\chi(Nuc_q, Nuc_r)_{Het}$	$\Delta D_3(m_{2qi}, m_{2rj})_3$	b_5	− 42.49	1.33694	− 31.7788	< 0.005
	$\chi(Cat_q, Cat_r))_{HetNoX}$	$\Delta D_3(m_{3qi}, m_{3rj})_4$	b_6	750.76	8.98832	83.5259	< 0.005
	$V(Sub_q, Sub_r)_{Tot}$	$\Delta D_2(m_{1qi}, m_{1rj})_5$	b_7	− 34.19	0.70023	− 48.8225	< 0.005
	$Zv(Cat_q, Cat_r)_{Cuns}$	$\Delta D_1(m_{3qi}, m_{3rj})_2$	b_8	22.04	1.16398	18.9356	< 0.005
	$Zv(Solv_q, Solv_r)_{Cuns}$	$\Delta D_1(m_{4qi}, m_{4rj})_2$	b_9	− 12.46	0.16653	− 74.8101	< 0.005
	Intercept	–	e_0	− 0.91	0.18257	− 5.0065	< 0.005

[a] Input variables with coefficient b_k are the values of shift (Δ) in q-reac*vs*. r-reac for different properties: *α* average atomic polarizability, *EA* average atomic Electro Affinity, *χ* average atomic Sanderson Electronegativity, *Zv* average atomic number

[b] Coefficients of the variables in the model, the output variable is the Δ in enantiomeric excess *ee*(%)* of the q-reac with respect to the r-reac when both reactions have been carried out with(*R*)-catalyst

[c] Standard error of the coefficients

[d] Student t-value

[e] p-level of error

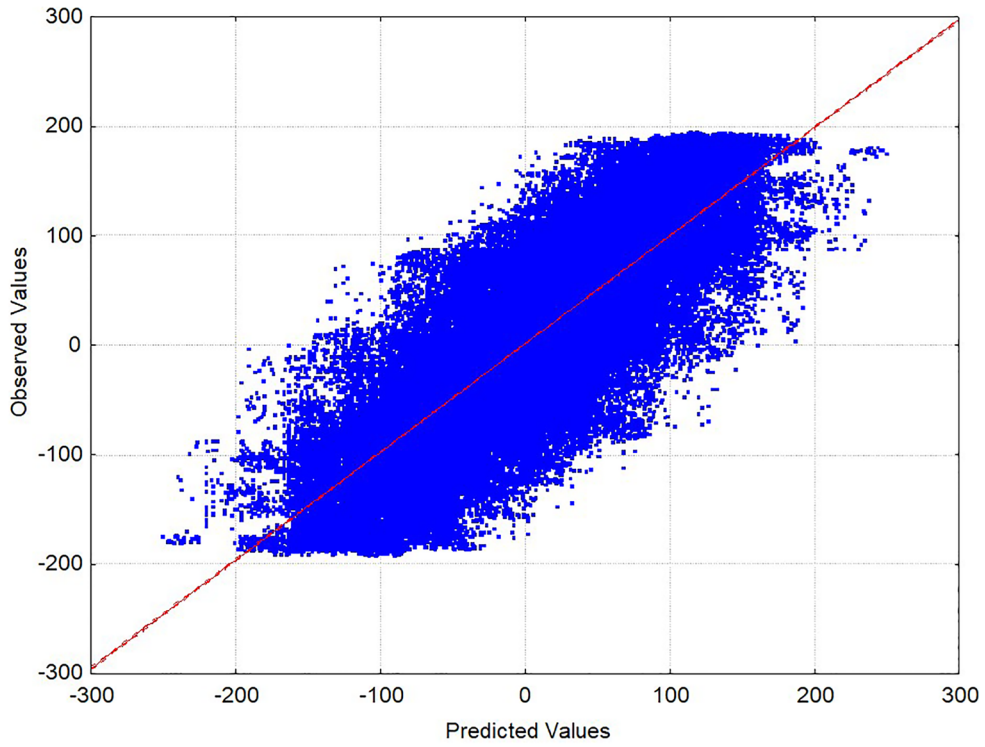


Fig. 2 Observed *vs*. Predicted ($\Delta ee_R(\%)_{qrobs}$ *vs*. $\Delta ee_R(\%)_{qrcalc}$) for equation Eq. 12 (R = 0.84 in training series). Only 10,000 reaction pairs of reactions (cases) are depicted due to software limitations

$$\begin{aligned} n &= 78732(react.pairs)R_{train} \\ &= 0.84F \\ &= 15238.7p < 0.5 \end{aligned}$$

The HPTML model was trained with a total of $n_{train}=78{,}732$ arbitrarily selected reaction pairs. The statistical parameters obtained for this model are the regression coefficient value of $R_{train}=0.84$ and Standard Error of Estimates SEE = 51.67 and a Fisher's ratio of F = 15,238.7 with a p-level < 0.05 in training series. This points out a important relationship between the observed relative values of $\Delta ee_R(\%)_{qrobs}$ and the predicted values $\Delta ee_R(\%)_{qrobs}$.

In addition, another subset of $n_{val}=28{,}836$ reaction pairs was used to validate the model. A regression coefficient $R_{val}=0.77$ and SEE = 60.225 were found for this validation series. The output of the model is $ee_R(\%)_{qrcalc}$. This variable represents the enantiomeric excess value calculated using a single reference reaction. The $ee_R(\%)_{calc}$ value quantifies the enantiomeric excess obtained using an (*R*)-catalyst. If $ee_R(\%)_{calc} > 0$, the product is predicted to have (*R*) notation; if $ee_R(\%)_{calc} < 0$, the product is predicted to have (*S*) notation; if $ee_R(\%)_{calc}=0$ racemic mixture. The overall p-level of the model is $p < 0.05$. All the variables introduced in the model are statistically significant (Table 4). The three first input variables quantify the effect of non-structural factors on the enantioselectivity parameter, $ee_R(\%)_{calc}$. The remaining input variables quantify the contribution of structural variations in the Substrate (Sub), Catalyst (Cat), Product (Prod), Nucleophile (Nuc), and Solvent (Solv).

PTML calculations with a single reference reaction

As we explained above, this PTML reactivity model studies pair-wise reactions. To avoid distortions in the distribution of the variables, PTML model uses the variable $\Delta ee_R(\%)_{qrobs}$ as objective function (see Eq. 5) [39]. This objective function is the function to fit and is equal to $\Delta ee_R(\%)_{qrobs}=ee_R(\%)_{qobs}-ee_R(\%)_{robs}$. As a result, the output of the new model is $\Delta ee_R(\%)_{qrcalc}=ee_R(\%)_{qcalc}- ee_R(\%)_{rcalc}$. For non-accurate models $\Delta ee_R(\%)_{qrcalc} \neq \Delta ee_R(\%)_{qrobs}$ (where ≠ indicates not ≈). Conversely, for a not-random accurate predictor, like this one, one can approximate $\Delta ee_R(\%)_{qrcalc} \approx \Delta ee_R(\%)_{qrobs}$. This presupposes that $ee_R(\%)_{qcalc} \approx ee_R(\%)_{qobs}$ and $ee_R(\%)_{rcalc} \approx ee_R(\%)_{robs}$. Therefore, for practical purposes, we use the model to predict the enantiomeric excess of new query reactions $ee_R(\%)_{qcalc}$, based on the observed enantiomeric excess of a reference reaction $ee_R(\%)_{qrobs}$. The approximation is only valid for not-random accurate predictors and takes into account that $ee_R(\%)_{rcalc} \approx ee_R(\%)_{robs}$ is always a known reference reaction, so it is necessary to rearrange the variables in Eq. 5 as shown in Eq. 13;

$$\begin{aligned} ee_R(\%)_{calcqi} = ee_R(\%)_{refj} &- 0.82 \cdot \Delta Load(\%) \\ &- 0.34 \cdot \Delta T(^oC) + 0.21 \cdot \Delta t(h) \\ &- 174.37 \cdot \Delta \propto \left(Cat_q, Cat_r\right)_{Cuns} \\ &- 1534.17 \cdot \Delta \propto \left(Prod_q, Prod_r\right)_{HetNoX} \\ &- 215.98 \cdot \Delta EA\left(Prod_q, Prod_r\right)_{Csat} \\ &- 1747.12 \cdot \Delta EA\left(Cat_q, Cat_r\right)_{HetNoX} \\ &- 42.49 \cdot \Delta \chi\left(Nuc_q, Nuc_r\right)_{Het} \\ &+ 750.76 \cdot \Delta \chi\left(Cat_q, Cat_r\right)_{HetNoX} \\ &- 34.19 \cdot \Delta V\left(Sub_q, Sub_r\right)_{Tot} \\ &+ 22.04 \cdot \Delta Zv\left(Cat_q, Cat_r\right)_{Cuns} \\ &- 12.46 \cdot \Delta Zv\left(Solv_q, Solv_r\right)_{Cuns} - 0.91 \end{aligned} \quad (13)$$

As a result of this approach, the model calculates different values of $ee_R(\%)_{calcqi}$ for the same reaction depending on the experimental value $ee_R(\%)_{refj}$ of the reaction used as reference in the pair [39]. Figure 2 illustrates the observed values of $\Delta ee_R(\%)_{qrobs}$ *vs.* the predicted (calculated) values of $\Delta ee_R(\%)_{calcqi}$ for 10,000 selected reaction pairs. We depict only 10000 pairs due to software plotting limitations (this the top number of points allowed by the software). A certain linear trend is observed (points with $\Delta ee_R(\%)_{qrcalc} \approx \Delta ee_R(\%)_{qrobs}$), however, despite being a predictor with adequate goodness of fit, there are many points with higher dispersion (points with $\Delta ee_R(\%)_{qrcalc} \neq \Delta ee_R(\%)_{qrobs}$).

In fact, PTML models may be included on a broader class of learning problems, such as delta ML, transfer ML, template selection ML, etc. [50–53]. In general, these models involve the use of a query item (item to be predicted) compared to a reference item (template, pair, known case, item from related domain, etc.). To calculate the output of a query item (quantum field, drug, protein, or reaction in this case), it is necessary to use an already known item or population of reference items as input. Query items can be in the same or a different data domain from the reference item. In this context, the low population (low number of available cases) of some of the studied data subset (data domains) is also a common problem. In our case, to calculate the value of $ee_R(\%)_{calcqi}$ for a query reaction (q), the observed $ee_R(\%)_{refj}$ values of an already known reference reaction (r) must be used as input. Here both the query and reference items come from the same data domain (both are the same type of reactions). The reaction of reference can be selected from our reaction dataset (same data domain) [54]. Consequently, for a new query reaction, there are n = 332 reactions in the dataset that can be used as the reference reaction, which pave the way for the question of which is/are the best candidate/candidates to be used as reference

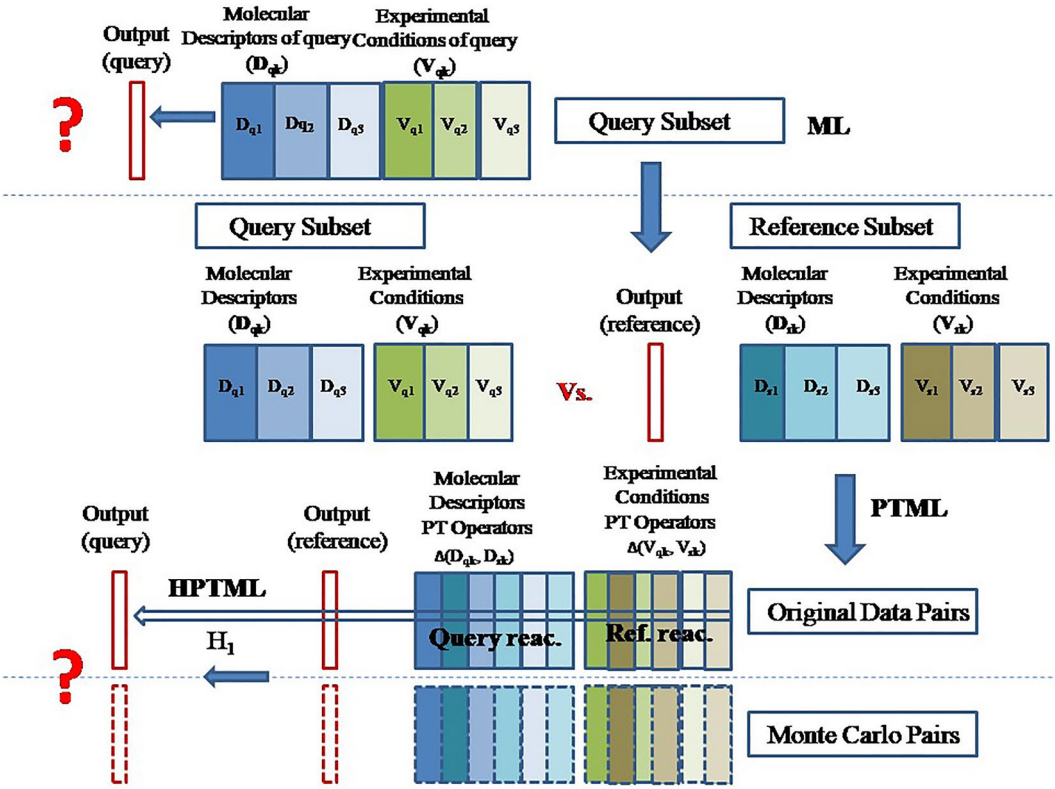


Fig. 3 HPTML data re-arrangement and MC data enrichment schematic illustration

Table 5 HPTML models obtained with different datasets *vs.* alternative heuristics

Algorithm[a]	Data	Heuristic	n_{reacc}[b]	n_{pairs}[c]	R^2	SEE	F	p-level
ML	OD	H_0	332	0	0.55	37.1	59.2	< 0.05
HPTML	OD	H_1	332	107626	0.81	29.5	1332.2	< 0.05
	OD	H_2	332	107626	0.64	39.3	603.0	< 0.05
HPTML	ODMC	H_1	332	109298	**0.96**	13.5	7560.6	< 0.05
+ MC	ODMC	H_2	332	109298	0.66	38.7	631.4	< 0.05

[a] *OD* Original Data, *MC* Monte Carlo, *ODMC* OD + MC enriched dataset

[b] n_{reacc} Number of reactions present in our dataset

[c] n_{pairs} Number of pairs of reactions present in our dataset

reaction in each case (see next section). Thus, 332 different values of $ee_R(\%)_{calcqi}$ can be calculated for the same query reaction based on the selected pairing reaction of reference. In this step, heuristic rules can be used to approximate the final predicted value $ee_R(\%)_{qpred}$ depending on the $ee_R(\%)_{calc}$ values of the model, as we have demonstrated previously to solve a similar problem [39].

HPTML model for prediction with multiple reactions of reference

As mentioned above, it is necessary to define the best reaction or set of reactions to use. Defining an appropriate reference reaction can also help reduce the dispersion and increase the value of the regression coefficient, because each query reaction will have a single predicted value. With this purpose, a Heuristic rule coupled to the PTML model can be used to select the best reference. Heuristic-based methods have been widely used in Cheminformatics to solve practical problems [55–57]. In our case, the combination of the PTML model with a Heuristic (H) rule defines the term HPTML = H + PTML algorithm. Two Heuristics (H_1 and H_2) were tested by calculating the $ee_R(\%)_{qrpred}$ values for the 332 reactions in our dataset, using the

Table 6 Selected subsets of reactions

Structural Patterns[a]			React Family	Absolute abundance [b]			*S*-catalyst products		*R*-catalyst products	
				OD	OD	MC	$<ee_S(\%)>_{qobs}$		$<ee_R(\%)>_{qobs}$	
Sub	Nuc	Cat	(Subset)	n_{reacc}	n_{pairs}	$n_{mcpairs}$	*S*	*R*	*R*	*S*
a	*a*	*a*	***aaa***	120	37570	605	− 43.2	76.1	43.2	− 76.1
a	*f*	*a*	***afa***	42	13943	211	− 71.0	76.0	71.0	− 76.0
h	*a*	*a*	***haa***	38	12616	190	ND	78.1	ND	− 78.1
c	*a*	*b*	***cab***	29	9628	145	− 17.0	69.8	17.0	− 69.8
a	*f*	*b*	***afb***	19	6307	95	− 53.0	87.4	53.0	− 87.4
a	*c*	*b*	***acb***	17	5644	85	ND	50.4	ND	− 50.4
c	*a*	*a*	***caa***	14	4648	70	− 15.5	27.7	15.5	− 27.7
e	*a*	*b*	***eab***	8	2656	40	ND	79.9	ND	− 79.9
a	*a*	*b*	***aab***	4	1328	20	− 21.0	ND	21.0	ND
d	*a*	*b*	***dab***	3	995	15	ND	85.7	ND	− 85.7

[a] *Sub* Substrate, *Nuc* Nucleophile, *Cat* Catalyst, patterns *a*, *b*, *c*, *aaa*, etc. are the different families of reactants/reactions, see the text, ND no data

[b] *OD* Original Data, *MC* Monte Carlo, *ODMC* OD + MC enriched dataset. n_{reacc} Number of reactions present in our dataset. n_{pairs} Number of pairs of reactions present in our dataset, $n_{mcpairs}$ Number of pairs of reactions present in our dataset in MC experiments

PTML trained with the OD set. These HPTML models based on Heuristics H_1 and H_2 were compared with a classic ML model. This classic ML model includes no PT terms and was built without using Heuristics (H_0). Figure 3 shows a schematic illustration of the ML, PTML, and HPTML data re-arrangement, as well as the MC data enrichment procedures used here.

Table 5 shows the statistical parameters for these studies (see only entries with Data = OD). Detailed information can be found in Additional file 2: Table S1of the Supporting Information file (Additional file 2). It should be noted that both HPTML models using Heuristics give good results with an OD regression coefficient in the range $R^2 = 0.64–0.81$ and $p < 0.05$. Specifically, the HPTML OD H_1 model has a higher regression coefficient ($R^2 = 0.81$ *vs.* 0.55) and a lower SEE ($R^2 = 29.5$ *vs.* 37.1) than the classic ML model. However, this SEE value is still relatively high. Interestingly, MC data enrichment improved both $R^2 = 0.96$ and SEE = 13.5 values of the HPTML OD H_1 model. In addition, the HPTML model automatically provides the most similar reference reaction from the reference dataset, including the reference of the article, which might give some clues about the possible reaction mechanism, etc. of the query reaction. In contrast, the classic ML model does not give information about the plausible reaction mechanism or similar reactions in the literature. Overall, these results justify the use of the HPTML algorithm instead of the classic ML algorithm.

Interestingly, the pair-wise strategy can rapidly increase the number of cases, as you go from datasets with n items (reactions) to n x n items (pairs of reactions). In this case, we go from $n_{reacc} = 332$ reactions to $n_{pairs} = 107{,}626$ pairs of reactions, which could be an advantage of PTML model, since increasing the number of items to train the ML model can improve learning. However, those items that are underrepresented in the original data are still underrepresented in the new data in relative terms. In addition, you take the risk of including mismatched pair, that is, you take the risk of trying to predict an

Table 7 HPTML Data set *vs.* heuristics correlation matrix

	HPTML Models[a]		$ee_R(\%)_{qcalc}$				
			OD	OMCD	MCD	OD	MCD
Heuristic	Data	$ee_R(\%)_{qobs}$		H_1		H_2	
H_1	OD	0.90	1.00				
	ODMC	0.98	0.92	1.00			
	MC	0.99	0.90	0.98	1.00		
H_2	ODMC	0.80	0.84	0.81	0.80	1.00	
	MC	0.81	0.84	0.81	0.81	1.00	1.00

[a] *OD* Original Data, *MC* Monte Carlo, *ODMC* OD + MC enriched dataset

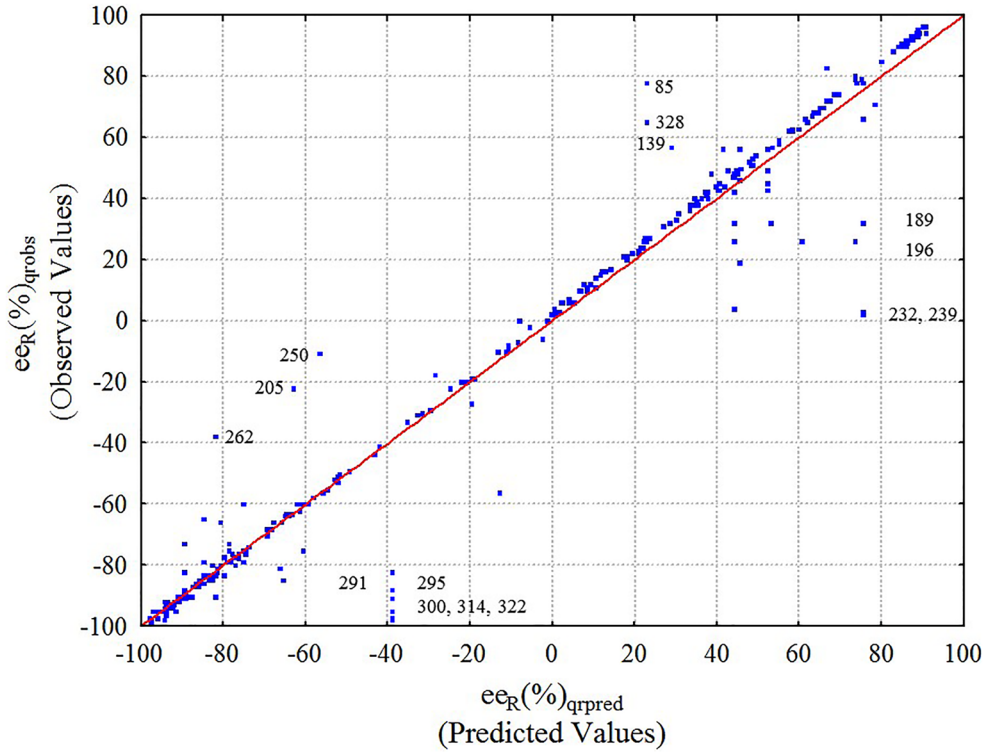
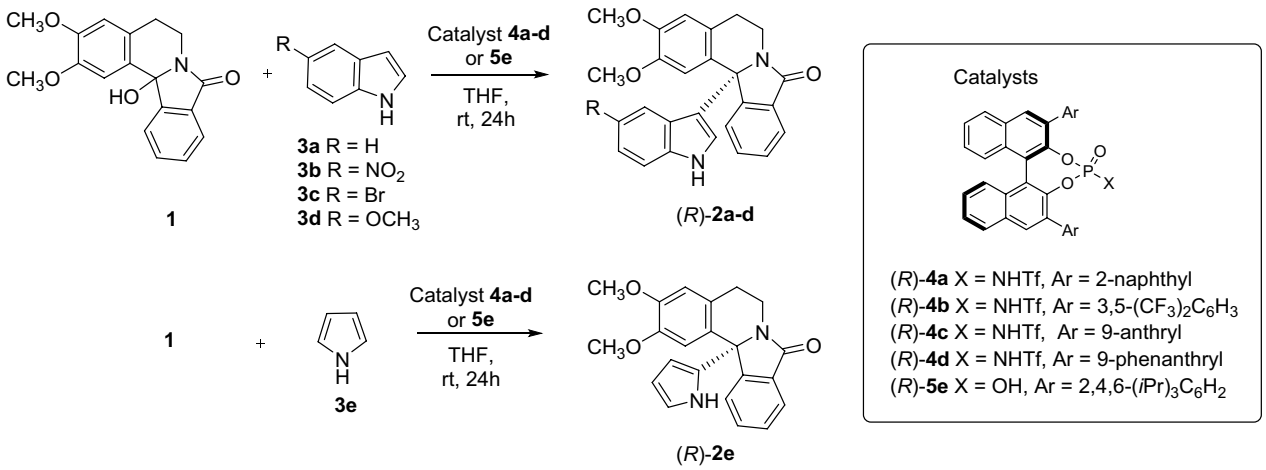


Fig. 4 HPTML ee_R(%) observed *vs.* predicted values with Eq. 12 ($R^2 = 0.98$) after applying both MC synthetic data and best Heuristic rule (ODMC + H_1). Overall data for training and validation series. The reaction number from the database (See Additional file 2) has been included for selected examples

underrepresented query item (reaction) using as reference an overrepresented item (reaction family) that is not similar to the reference. For example, reactions from the ***aaa*** family are generally the most represented with $n_{reacc} = 120$ cases (36.14% of cases) and $n_{pairs} = 37{,}570$ (34.91%) including many pairs with reactions from the same family. In contrast, reactions from the ***dab*** family are very poorly represented (low abundance) with only $n_{reacc} = 3$ cases (0.9% of cases) appearing in $n_{pairs} = 995$ pairs of reactions. Almost all of these pairs are formed with reactions from other families and the relative abundance remains low (0.9%).

Table 6 shows the absolute and relative abundance of different reaction families (subsets) in the original dataset and the number of pairs formed with them. It should be noted that the formation of pairs of mismatched reactions can lead to inaccurate predictions. For example, predicting a query reaction from the ***aab*** family may give an inaccurate result if we use a reaction from the ***haa*** family as reference, because ***aab*** reactions have an average enantiomeric excess $< ee_R(\%) >_{qobs} = 21.0$ while ***haa*** reactions have $< ee_R(\%) >_{qobs} = -78.1$. Both reaction families not only have a markedly different average enantiomeric excess, but also give products with reverse (*R*) or (*S*) CIP notation of absolute configuration [31]. The compound codes, SMILE codes, and chemical structures of the different families of substrates, nucleophiles, and catalysts are shown on the Supporting Information file SI00.pdf.

In this regard, synthetic data generation techniques can be used to palliate the presence of low populated data subsets. In any case, the total abundance of each enriched data subset should remain essentially constant to avoid creating data artifacts. MC sampling methods have widely used in chemistry for similar purposes [58]. To palliate this situation, we have used a Mersenne-Twister MC algorithm (MT19937) [59] for data enrichment by creating new synthetic data. Therefore, synthetic data cases of the input variables $V_k(c_{qi}) = T(°C)_{qi}$, $t(h)_{qi}$, or $L(\%)_{qi}$ of query reactions were generated using a MC algorithm (see system of equations in Materials and Methods section). The same MC algorithm (system of equations) was used to generate new synthetic data for the input variables of the equation of reference $V_k(c_{rj})$. Nevertheless, the molecular descriptors $D_k(m_{sqi})$ and $D_k(m_{srj})$ were never modified in the MC data enrichment simulation, because one can reasonably expect that small changes in the input reaction condition



Scheme 4. New enantioselective α-amidoalkylation reactions of indoles **3a-d** and pyrrole **3e** with Triflamide catalysts **4a-d** and their comparison with CPA catalyst **5e.** Synthesis of enantioenriched isoindoloisoquinolines **2a-e** (Table 8)

Table 8 Enantioselective intermolecular α-amidoalkylation reactions of *N*-triflamides *vs*. phosphoric acids as catalysts

Reactions	entry	Nuc_{qi}	Cat_{qi}	Prod	Yield (%)[a]	$ee_R(\%)_{obs}$[b]
New reactions (Catalysts **4a**-**4d**)	1	**3a**	**4a**	**2a**	90	93
	2	**3a**	**4b**	**2a**	70	0
	3	**3a**	**4c**	**2a**	70	26
	4	**3a**	**4d**	**2a**	70	65
	5	**3b**	**4a**	**2b**	94	11
	6	**3c**	**4a**	**2c**	quant	67
	7	**3d**	**4a**	**2d**	quant	52
	8	**3e**	**4a**[c]	**2e**	quant	54
Reported reactions (Catalyst **5e**)	9	**3a**	**5e**[d]	**2a**	70	74
	10	**3b**	**5e**[d]	**2b**	–	–
	11	**3c**	**5e**[d]	**2c**	74	58
	12	**3d**	**5e**[d]	**2d**	79	74
	13	**3e**	**5e**	**2e**	77	21
	14	**3e**	**5e**[e]	**2e**	18	12
	15	**3e**	**5e**[f]	**2e**	27	24

[a] Yield (%) of isolated pure compound, the symbols **2a** - **5e** are the reactants and products, see scheme 4

[b] Determined by chiral stationary-phase HPLC

[c] Reaction time: 5 h

[d] Reactions previously reported by our group using the (*R*)-phosphoric acid **5e** as catalyst

[e] Temperature: − 30 °C

[f] Catalyst loading 2.5 mol%, reaction time 48 h

variables [$V_{k(cqi)}$ = T(°C), t(h), or L(%)] do not to significantly change the output ee_R(%). However, the same cannot be guaranteed for changes in chemical structure. Thus, we obtained a slightly higher number of cases for very low abundant reactions. For example, we were able to add $n_{mcpairs}$ = 15, 20, or 40 new cases for the ***dab***, ***aab***, and ***eab*** families of reactions; but we kept their relative abundance essentially low in the range, 0.9–2.47%. Table 6 shows that both models trained with the ODMC dataset (OD enriched by MC) give essentially the same value of R = 0.8–0.9 and $p < 0.05$ obtained with OD alone. However, the error decreased from SEE = 29.5% to SEE = 13.5% using Heuristic H_1. Table 7 shows the correlation matrix for the outputs of all models that illustrates the high correlation obtained among them, R = 0.80–0.99. The results of $ee_R(\%)_{qrobs}$ observed *vs*. $ee_R(\%)_{qrpred}$

predicted with this HTPML model using ODMC dataset and H_1 heuristic are graphically depicted in Fig. 4, where each point corresponds to a reaction included in the dataset. It can be graphically observed that although an excellent correlation of the predicted and obtained *ee*(%) value is generally obtained, some values are far from the line of correlation. In selected cases, the corresponding reaction number from the database (See SI001.xls file) has been included. It is difficult to draw any conclusions from these cases, as the reactants used are structurally heterogeneous and the experimental conditions diverse as well. In any case, the model has already a very high $R^2 = 0.98$ value. We can conclude that using ODMC enriched data decreased the error of the model without decreasing the regression quality.

HPTML *vs.* Experimental study of new reactions

In this section, we report an additional test of the HPTML model comparing the computational predictions with the experimental study of new reactions. Thus, we performed both an experimental and a theoretical study of new intermolecular α-amidoalkylation reactions not previously reported in the literature. First, the α-amidoalkylation reactions carried out experimentally are described. Next, we report the use of the HPTML model to predict these reactions and compare the results with the experimental values.

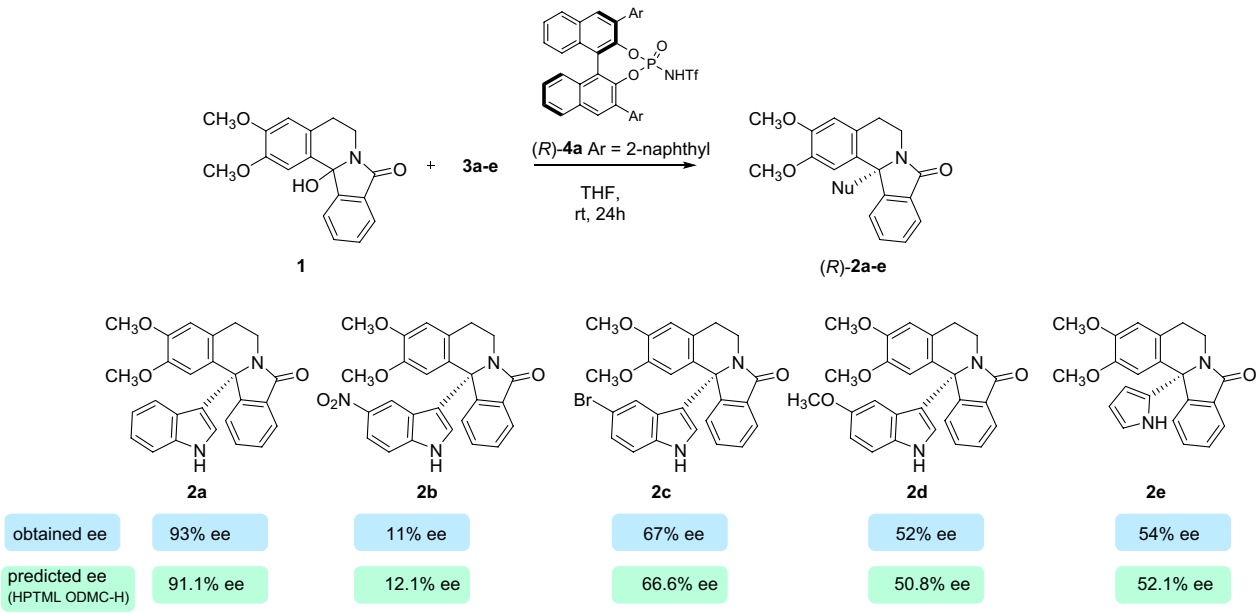
Experimental study of α-amidoalkylation reactions.

As stated above, the α-amidoalkylation reaction is a very attractive method for C–C bond formation in organic synthesis. In this context, we have previously reported [27] that the α-amidoalkylation reaction is an efficient procedure for the enantioselective synthesis of 12b-substituted isoindoloisoquinolines (Nuevamine-type alkaloids [60]) using BINOL-derived Brønsted acids as catalysts. It should be pointed out that these catalysts have been used in intermolecular α-amidoalkylation of indoles with cyclic *N*-acyliminium ions formed in situ from cyclic hydroxylactams to form tertiary or quaternary stereogenic centers, but this was the first example of bicyclic *N*-acyliminium intermediates in intermolecular α-amidoalkylation reactions of indoles [30].The best results were obtained using a sterically demanding CPA (20 mol% catalyst loading) under the following conditions: THF as solvent at room temperature for 24 h. However, in some cases, moderate enantioselectivity (enantiomeric excess) and/or yields were obtained. Therefore, we decided to test BINOL-derived *N*-triflylphosphoramides as catalysts to enhance the enantioselectivity of these reactions, because they are known to have an increased acidity when compared to the corresponding CPAs, so they can form tighter ion pairs leading to an improved reactivity [61, 62]. Thus, the *N*-triflylphosphoramides **4a**-**d** were synthesized [63, 64] and tested as catalysts in the reaction of

Table 9 HPTML study of new enantioselective intermolecular α-amidoalkylation reactions

Reaction inputs[a]		Reaction features[b]	New reactions (Table 8)							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Reactants		Nuc	**3a**	**3a**	**3a**	**3a**	**3b**	**3c**	**3d**	**3e**
		Cat	**4a**	**4b**	**4c**	**4d**	**4a**	**4a**	**4a**	**4a**
Input conditions		Load (%)	20	20	20	20	20	20	20	20
		T (^{0}C)	25	25	25	25	25	25	25	25
		T (h)	24	24	24	24	16	24	24	5
Heuristic	Data	ee_R (%)	Observed, Predicted, and Residual values							
–	OD	Observed	93	0	26	65	11	67	52	54
H_1	OD	Predicted	66.0	29.7	64.1	25.1	− 90.5	74.9	61.8	42.4
		Residual	27.0	− 29.7	− 38.1	39.9	101.5	− 7.9	− 9.8	11.6
	ODMC	Predicted	91.1	0.3	64.1	25.1	12.1	66.6	50.8	52.1
		Residual	1.9	− 0.3	− 38.1	39.9	− 1.1	0.4	1.2	1.9
H_2	OD	Predicted	− 36.2	22.0	− 50.8	− 50.9	− 121.3	− 36.1	− 49.2	− 57.9
		Residual	129.2	− 22.0	76.8	115.9	132.3	103.1	101.2	111.9
	ODMC	Predicted	− 34.3	21.6	− 49.6	− 49.2	− 119.3	− 34.5	− 47.7	− 56.2
		Residual	127.3	− 21.6	75.6	114.2	130.3	101.5	99.7	110.2

[a] *OD* Original Data, *MC* Monte Carlo, *ODMC* OD + MC enriched dataset. *Nuc* = Nucleophile, *Cat* = Catalyst, *Load (%)* Catalyst loading (%), the symbols **3a** - **4d** are *Nuc* and *Cat*, see scheme 4



Scheme 5. Experimentally obtained ee values *vs.* predicted (HPTML ODMC-H) for the obtention of **2a-e** using catalyst **4a**

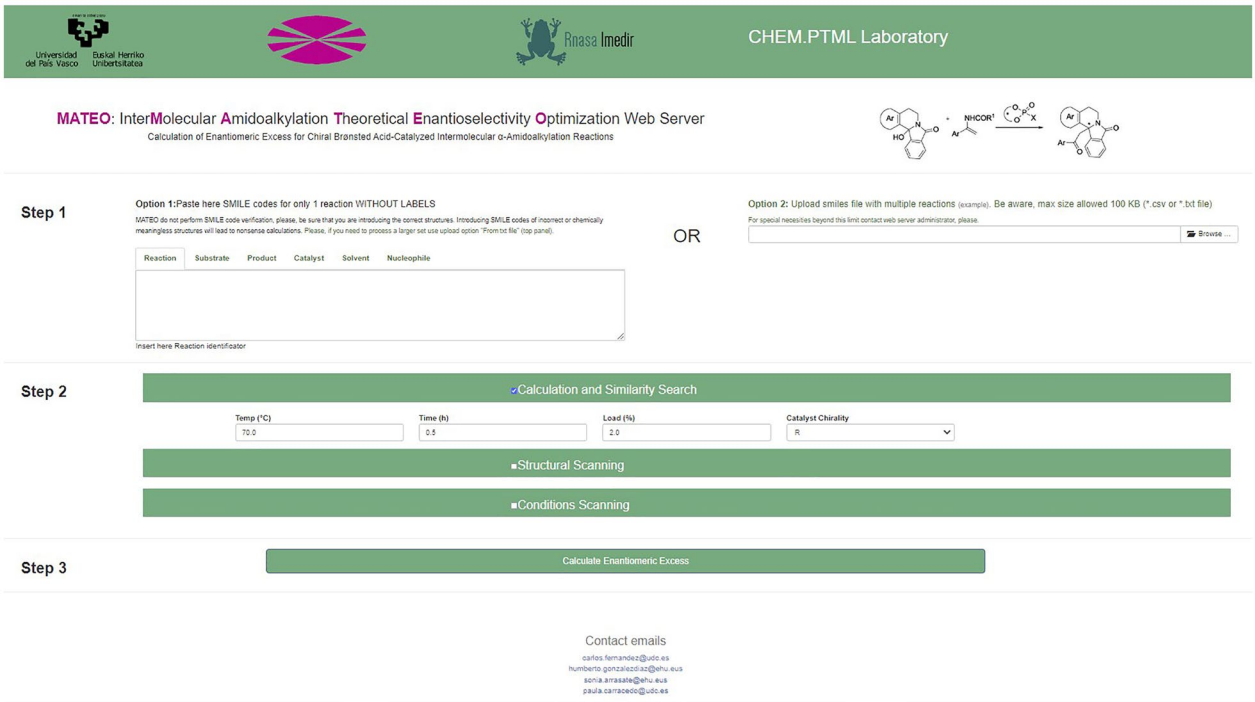


Fig. 5 MATEO web server user interface

12b-hydroxyisoindoloisoquinoline **1** with the indoles **3a-d** (Scheme 4). Table 8 summarizes these new results compared with those previously obtained with phosphoric acid **5e**, which has demonstrated to be the most efficient catalyst for indole [30].The best results were obtained with the catalyst **4a**, although good to excellent yields were achieved with all the phosphoramides. Successfully, we were able to improve our previous result obtaining

Table 10 MATEO Web server operational conditions

Stat.[a]	MATEO application operational conditions[b]		
	T (°C)	T (h)	Load (%)
Default	25.00	0.5	2.00
Min	− 78.00	0.5	2.00
Max	70.00	72.0	5.00
Step	20	1.0	1.00

[a] *Stat.* Statistical parameters for the input parameters (operational conditions) of all the reactions present in our dataset: *Max.* maximum value, *Min.* minimum value, *Step* minimal change allowed in one experimental condition

[b] Operational conditions: *T(°C)* temperature, *t(h)* reaction time, *Load(%)* catalyst loading

with the corresponding phosphoric acids, obtaining **2a** with excellent yield and enantioselectivity (90, 93% *ee*). In addition, the intermolecular α-amidoalkylation reaction was extended to 5-substituted indoles **3b-d**, obtaining excellent yields, even when a strong acceptor group (NO_2) was introduced (Table 8, entry 5). However, the use of the substituted indoles led to lower enantiomeric excesses (Table 8, entries 5–7). The reaction could also be applied to other electron-rich heteroaromatics as pyrrole **3e**, obtaining **2e** quantitatively, although with moderate *ee* (Table 8, entry 8). In this case, the reaction was cleaner and faster (reaction completed in 5 h) than when using phosphoric acid **5e** as catalyst (Table 8, entries 13–15).

HPTML prediction of new α-amidoalkylation reactions

Next, using the developed HPTML ODMC H_1 model, we predicted the values of ee_R(%) for the new enantioselective intermolecular α-amidoalkylation reactions. We first calculated the molecular $D_k(m_{qsi})_g$ descriptors of all the molecules ($Substrate_{qi}$, $Nucleophile_{qi}$, $Catalyst_{qi}$, $Solvent_{qi}$, and $Product_{qi}$) involved in the new query reactions (R_q) using the web server MCDCalc [38]. Then, the Heuristic $H_1$was used to find the best reference reaction for each new query reaction. Next, we substituted in the model equation the values of the molecular descriptors $D_k(m_{qsi})_g$ and $D_r(m_{rsj})_g$ of the molecules, as well as the values of the input experimental conditions variables $V_k(c_{qi})$ and $V_k(c_{rj})$, from both the query (R_q) and reference reaction (R_r), respectively. Table 9 shows the predicted ee_R(%) values for each reaction compared to the values

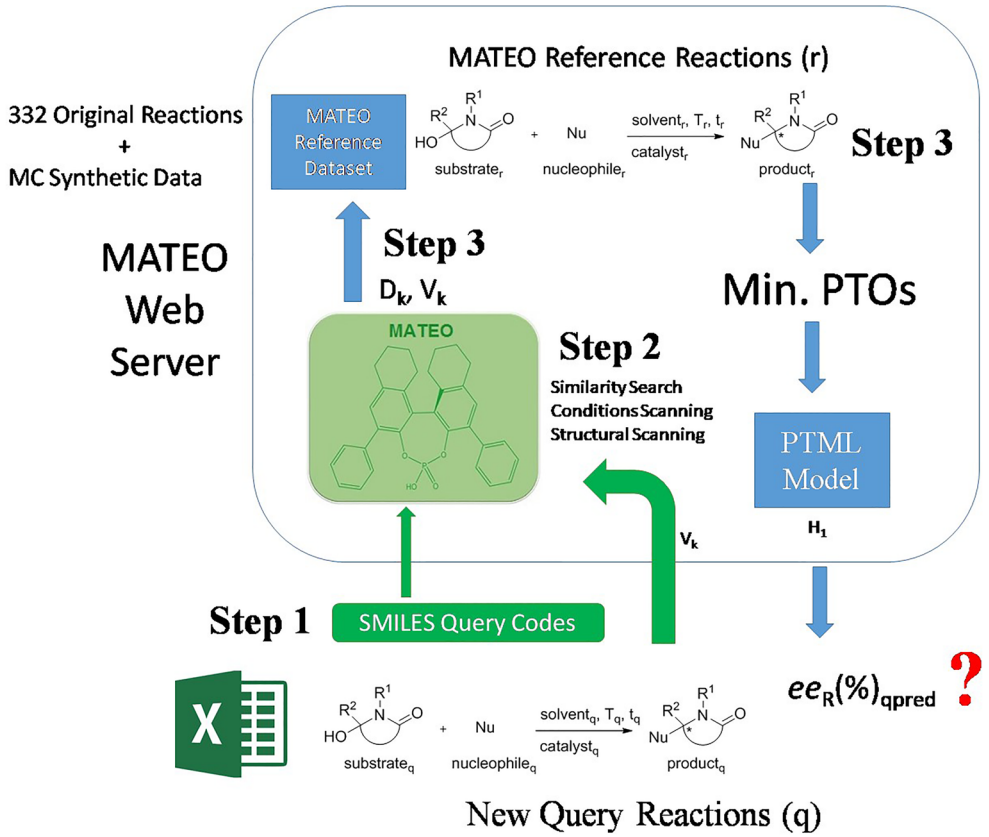


Fig. 6 MATEO server use workflow

predicted with the other Datasets (OD *vs.* ODMC) and Heuristics (H_1 and H_2).

The other HPTML models have notably larger residuals values, confirming our decision to discard them as good predictors for this type of reaction. In general, the best results are obtained with the HPTML ODMC H_1 model. For a total of 6 out of 8 reactions the model almost perfectly predicts the observed values of $ee_R(\%)_{qrobs}$ with residual values in the range $ee_R(\%)_{qrres} = -1.1$–1.9% (reactions 1, 2, 5–8) (Table 9). The experimental and predicted values for the obtention of **2a-e** using catalyst **4a** are represented in Scheme 5. For the other two reactions, the model correctly predicts the absolute stereochemistry of the final products, although with a relatively higher error. In addition to the results of training and validations series, these results validate the HPTML ODMC H_1 model as a useful predictor for enantioselective intermolecular α-amidoalkylation reactions. The Microsoft Excel software was used to run all these calculations. However, this HPTML calculation algorithm is slow because it is not automatic and need more than one software applications (MCDCalc, Excel) to run. Furthermore, the model is not available for use by other groups and requires some degree of expertise in Cheminformatics, so we decided to implement it on a public web server.

MATEO web server

The HPTML model was implemented on a new public web server called MATEO: interMolecular Amidoalkylation Theoretical Enantioselectivity Optimization. MATEO server is available for public use online (free of charge) through the link: https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/mateo. The graphical interface of the web server is shownin Fig. 5.Users worldwide can upload their own sets of query reactions to predict the values of $ee_R(\%)_{qrcalc}$ under different experimental conditions (solvent, time, temperature, catalyst loading), see Table 10.

Figure 6 graphically illustrates (from bottom to top) the steps required to use this web server. Step 1 is to upload the chemical structures of all the molecules involved in the reaction. The server is required to upload the structures in the Simplified Molecular Input Line Entry Specification (SMILES) code format [65]. SMILES has become a simplified and memory-optimal way of managing molecular structures widely used in Cheminformatics today [66, 67]. These codes can be pasted directly on the web interface or uploaded as a text file. The server allows uploading large collections of reactions with different combinations of substrate, nucleophile, and catalyst. This could be useful for exploring large libraries of molecules (products, substrates, and nucleophiles) and/or for the design of new catalysts. The server also allows uploading of the solvent structure, making it easy to explore a large variety of solvents. In Step 2, three general types of calculations can be selected: (1) Similarity Search, (2) Structural Scan, or (3) Conditions Scan. Option (1) allows us to predict the enantiomeric excess values, in addition to obtaining a report of the most similar reactions from the references in our dataset. Option (2) allows uploading the specific structures (substrate, nucleophile, catalyst, and/or solvent) and running a scan of these molecules under reaction conditions similar to those reported in the literature. Option (3) allows to keep the structure parameters constant (same molecules), while the software performs a scan of different combinations of input variables (temperature, time, catalyst loading). Table 10 shows the range (minimum, maximum) and step of the variables allowed by the server.

In this context, Goodman et al. have recently developed a rule-based web tool BINOPtimal for the online selection of CPA catalysts in a related reaction, the addition of nucleophiles to imines, by analyzing the reagent structures [68]. MATEO is web server allows the user to make quantitative predictions of enantiomeric excess parameter $ee_R(\%)$ at different reaction temperature, time, catalysts loading or solvent polarity, which are known factors that affect the enantioselectivity of α-amidoalkylation reactions. Therefore, MATEO web server will be useful to guide not only the catalyst selection but also the experimental conditions.

Conclusions

In conclusion, we have shown that classic linear ML models are not very accurate in predicting the enantioselectivity of α-amidoalkylation reactions using physicochemical properties calculated with a Markov chain approach as input. Besides, these linear ML models do not allow detecting the most similar reaction directly from the model. The PTML algorithm outperforms the classic linear ML model using the same dataset and molecular descriptors. Moreover, the HPTML algorithm based on PTML model + heuristic rule allows direct detection of the most similar reference reactions. In addition, MC synthetic data re-sampling/enrichment procedures reduce the procedural error. The final HPTML model responds very well in computational experiments with validation series. The HPTML model also reproduces very well the experimental values of a new series of reactions studied experimentally by the first time in this work. Finally, the implementation of the HPTML model on the MATEO online server makes the algorithm available for public use worldwide with a user-friendly interface.

Abbreviations

AI	Artificial intelligence
ANN	Artificial neural networks
CPA	Chiral phosphoric acid
GLR	General linear regression
ML	Machine learning
HPTML	Heuristic perturbation-theory and machine learning
LNN	Linear neural network
MARCH-INSIDE	Markov chain invariants for networks simulation and design
MATEO	InterMolecular amidoalkylation theoretical enantioselectivity optimization
MC	Monte carlo
ML	Machine learning
MLR	Multivariate linear regression
THF	Tetrahydrofuran
OD	Original data
PT	Perturbation theory
PTO	Perturbation theory operator
SE	Standard error
SEE	Standard error estimates
SMILE	Simplified molecular input line entry specification
$ee_R(\%)_{obs}$	Observed enantiomeric excess (experimental) using (*R*)-Catalyst
$ee_R(\%)_{ref}$	Enantiomeric excess of reference (experimental) using (*R*)-Catalyst
$ee_R(\%)_{calc}$	Enantiomeric excess using (*R*)-Catalyst calculated using one reference
$ee_R(\%)_{pred}$	Enantiomeric excess using (*R*)-Catalyst predicted by the model
$ee_R(\%)_{res}$	Residual enantiomeric excess using (*R*)-Catalyst

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00802-7.

Additional file1: The following files are available free of charge. General experimental methods; Synthetic procedures and structural determination for **2a-d**; Copies of HPLC chromatograms of racemic and enantioenriched **2a-d**; Copies of ^{1}H and ^{13}C NMR spectra

Additional file2: Dataset of reactions, molecular descriptors, SMILE codes, *etc*.

Additional file3: MATEO server reactions of reference

Acknowledgements
Technical and human support provided by General Research Services SGIker (UPV/EHU, MINECO, GV/EJ, ERDF and ESF) is also acknowledged.

Author contributions
SA, CRM, CFL, NS, EL, and HGD conceived the presented idea. PCR and CRM implemented the idea computationally, performed the computations and analysis. EA performed the organic synthesis experiments. SH carried out the data analysis and software validation. SA, CRM, CFL, NS, EL, and HGD supervised the findings of this work. All authors discussed the results and wrote the manuscript with input of all authors. All authors read and approved the final manuscript.

Funding
The authors acknowledge financial support from Grant PID2019-104148 GB-I00 and PID2022-137365NB-I00 funded by MCIN/AEI/10.13039/501100011033 and Grant IT1558-22 funded by Basque Government/Eusko Jaurlaritza, 2022–2025.CITIC is funded by the Xunta de Galicia through the collaboration agreement between the Department of Culture, Education, Vocational Training and Universities and the Galician universities to strengthen the research centers of the Galician University System (CIGUS).

Availability of data and materials
MATEO web server was implemented for public use by experimental organic chemists, see link: https://cptmltool.rnasa-imedir.com/CPTMLTools-Web/mateo.The code of the software was uploaded to a GitHub repository and is available free for use by cheminformatics researchers with MIT license. The links are the following. For the MATEO server code the link is: https://github.com/glezdiazh/MATEO. For libraries used to calculate the molecular descriptors the link is: https://github.com/muntisa/RMarkovTI.All data files (SI00, SI01, and SI02) have been uploaded to a public data repository and are available for use free of charge under universal commons creative license (CC0). The links are, SI00.pdf file link: https://doi.org/https://doi.org/10.6084/m9.figshare.21981740.v2, Additional file 2: https://doi.org/https://doi.org/10.6084/m9.figshare.21971690.v2, and Additional file 3: https://doi.org/https://doi.org/10.6084/m9.figshare.21971696.v2.

Declarations

Ethics approval and consent to participate
Bioethics approval is not applicable (not laboratory animals or personal data is used). All authors consent to participate in the paper.

Competing interests
The authors declare that they have no competing interests.

Received: 1 March 2023 Accepted: 11 January 2024
Published online: 23 January 2024

References

1. Parmar D, Sugiono E, Raja S, Rueping M (2014) Complete field guide to asymmetric BINOL-phosphate derived Brønsted acid and metal catalysis: history and classification by mode of activation; Brønsted acidity, hydrogen bonding, ion pairing, and metal phosphates. Chem Rev 114:9047–9153
2. Parmar D, Sugiono E, Raja S, Rueping M (2017) Addition and correction to complete field guide to asymmetric BINOL-phosphate derived Brønsted acid and metal catalysis: History and classification by mode of activation; Brønsted acidity, hydrogen bonding, ion pairing, and metal phosphates. Chem Rev 117:10608–10620
3. Akiyama T (2012) Asymmetric C-C bond formation using chiral phosphoric acid. In: Christman N, Bräse S (eds) Asymmetric Synthesis II: More Methods and Applications. Wiley, Weinheim, pp 261–266
4. Wu X, Gong LZ (2014) Chiral phosphoric acid-catalyzed asymmetric multicomponent reactions. In: Zhu J, Wang Q, Wamg MX (eds) Multicomponent reactions in organic synthesis. Wiley, Weinheim, pp 439–470
5. Zhu L, Mohamed H, Yuan H, Zhang J (2019) The control effects of different scaffolds in chiral phosphoric acids: a case study of enantioselective asymmetric arylation. Catal Sci Technol 9:6482–6491
6. ElKerdawy A, Güssregen S, Matter H, Hennemann M, Clark T (2014) Quantum-mechanics-based molecular interaction fields for 3D-QSAR. J Cheminform 6:1–2
7. Spjuth O (2018) Novel applications of machine learning in cheminformatics. J Cheminform 10:1–2
8. Drakakis G, Koutsoukas A, Brewerton SC, Evans DD, Bender A (2013) Using machine learning techniques for rationalising phenotypic readouts from a rat sleeping model. J Cheminform 5:1–1
9. Ye Z, Ouyang D (2021) Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. J Cheminform 13:1–13
10. Ruscher M, Herzog A, Timoshenko J, Jeon HS, Frandsen W, Kuhl S, Roldan Cuenya B (2022) Tracking heterogeneous structural motifs and the redox behaviour of copper-zinc nanocatalysts for the electrocatalytic CO(2) reduction using operando time resolved spectroscopy and machine learning. Catal Sci Technol 12:3028–3043
11. Takahashi K, Ohyama J, Nishimura S, Fujima J, Takahashi L, Uno T, Taniike T (2023) Catalysts informatics: paradigm shift towards data-driven catalyst design. Chem Commun 59:2222–2238
12. Sarma BB, Maurer F, Doronkin DE, Grunwaldt JD (2023) Design of single-atom catalysts and tracking their fate using operando and advanced X-ray spectroscopic tools. Chem Rev 123:379–444

13. Freeze JG, Kelly HR, Batista VS (2019) Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. Chem Rev 119:6595–6612
14. Tsai CC, Sandford C, Wu T, Chen B, Sigman MS, Toste FD (2020) Enantioselective intramolecular allylic substitution via synergistic palladium/chiral phosphoric acid catalysis: insight into stereoinduction through statistical modeling. Angew Chem Int Ed Engl 59:14647–14655
15. Gensch T, Dos Passos GG, Friederich P, Peters E, Gaudin T, Pollice R, Jorner K, Nigam A, Lindner-D'Addario M, Sigman MS, Aspuru-Guzik A (2022) A comprehensive discovery platform for organophosphorus ligands for catalysis. J Am Chem Soc 144:1205–1217
16. Dieguez-Santana K, Gonzalez-Diaz H (2021) Towards machine learning discovery of dual antibacterial drug-nanoparticle systems. Nanoscale 13:17854–17870
17. Barbolla I, Hernandez-Suarez L, Quevedo-Tumailli V, Nocedo-Mena D, Arrasate S, Dea-Ayuela MA, Gonzalez-Diaz H, Sotomayor N, Lete E (2021) Palladium-mediated synthesis and biological evaluation of C-10b substituted dihydropyrrolo[1,2-b]isoquinolines as antileishmanial agents. Eur J Med Chem 220:113458
18. Ortega-Tenezaca B, Gonzalez-Diaz H (2021) IFPTML mapping of nanoparticle antibacterial activity vs. pathogen metabolic networks. Nanoscale 13:1318–1330
19. Sampaio-Dias IE, Rodriguez-Borges JE, Yanez-Perez V, Arrasate S, Llorente J, Brea JM, Bediaga H, Vina D, Loza MI, Caamano O, Garcia-Mera X, Gonzalez-Diaz H (2021) Synthesis, pharmacological, and biological evaluation of 2-furoyl-based MIF-1 peptidomimetics and the development of a general-purpose model for allosteric modulators (ALLOPTML). ACS Chem Neurosci 12:203–215
20. Santana R, Zuluaga R, Ganan P, Arrasate S, Onieva E, Gonzalez-Diaz H (2020) Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. Nanoscale 12:13471–13483
21. Santana R, Zuluaga R, Ganan P, Arrasate S, Onieva Caracuel E, Gonzalez-Diaz H (2020) PTML model of ChEMBL compounds assays for vitamin derivatives. ACS Comb Sci 22:129–141
22. Aranzamendi E, Arrasate S, Sotomayor N, Gonzalez-Diaz H, Lete E (2016) Chiral bronsted acid-catalyzed enantioselective alpha-amidoalkylation reactions: a joint experimental and predictive study. ChemistryOpen 5:540–549
23. Yazici A, Pyne SG (2009) Intermolecular addition reactions of N-acyliminium ions (Part II). Synthesis 2009:513–541
24. Rahman A, Lin X (2018) Development and application of chiral spirocyclic phosphoric acids in asymmetric catalysis. Org Biomol Chem 16:4753–4777
25. Han B, He X-H, Liu Y-Q, He G, Peng C, Li J-L (2021) Asymmetric organocatalysis: an enabling technology for medicinal chemistry. Chem Soc Rev 50:1522–1586
26. Merad J, Lalli C, Bernadat G, Maury J, Masson G (2018) Enantioselective Brønsted acid catalysis as a tool for the synthesis of natural products and pharmaceuticals. Chem-Eur J 24:3925–3943
27. Aranzamendi E, Sotomayor N, Lete E (2012) Brønsted acid catalyzed enantioselective α-amidoalkylation in the synthesis of isoindoloisoquinolines. J Org Chem 77:2986–2991
28. Wheeler SE, Seguin TJ, Guan Y, Doney AC (2016) Noncovalent interactions in organocatalysis and the prospect of computational catalyst design. Accounts Chem Res 49:1061–1069
29. Peng Q, Duarte F, Paton RS (2016) Computing organic stereoselectivity–from concepts to quantitative calculations and predictions. Chem Soc Rev 45:6093–6107
30. Maji R, Mallojjala SC, Wheeler SE (2018) Chiral phosphoric acid catalysis: from numbers to insights. Chem Soc Rev 47:1142–1158
31. Helmchen G (2016) The 50th anniversary of the cahn–ingold–prelog specification of molecular chirality. Angew Chem Int Ed 55:6798–6799
32. Yu X, Lu A, Wang Y, Wu G, Song H, Zhou Z, Tang C (2011) Chiral phosphoric acid catalyzed asymmetric friedel-crafts alkylation of indole with 3-hydroxyisoindolin-1-one: enantioselective synthesis of 3-indolyl-substituted isoindolin-1-ones. Eur J Org Chem 2011:892–897
33. Yu X, Wang Y, Wu G, Song H, Zhou Z, Tang C (2011) Organocatalyzed enantioselective synthesis of quaternary carbon-containing isoindolin-1-ones. Eur J Org Chem 2011:3060–3066
34. Guo C, Song J, Huang JZ, Chen PH, Luo SW, Gong LZ (2012) Core-structure-oriented asymmetric organocatalytic substitution of 3-hydroxyoxindoles: application in the enantioselective totalsynthesis of (+)-folicanthine. Angew Chem Int Ed 51:1046–1050
35. Yin Q, Wang S-G, You S-L (2013) Asymmetric synthesis of tetrahydro-β-carbolines *via* chiral phosphoric acid catalyzed transfer hydrogenation reaction org. Lett 15:2688–2691
36. Carracedo-Reboredo P, Corona R, Martinez-Nunes M, Fernandez-Lozano C, Tsiliki G, Sarimveis H, Aranzamendi E, Arrasate S, Sotomayor N, Lete E (2020) MCDCalc: markov chain molecular descriptors calculator for medicinal chemistry. Curr Top Med Chem 20:305–317
37. Gonzalez-Diaz H, Duardo-Sanchez A, Ubeira FM, Prado-Prado F, Perez-Montoto LG, Concu R, Podda G, Shen B (2010) Review of MARCH-INSIDE & complex networks prediction of drugs: ADMET, anti-parasite activity, metabolizing enzymes and cardiotoxicity proteome biomarkers. Curr Drug Metab 11:379–406
38. Hill T, Lewicki P, Lewicki P (2006) Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft Inc., Tulsa
39. Simon-Vidal L, Garcia-Calvo O, Oteo U, Arrasate S, Lete E, Sotomayor N, Gonzalez-Diaz H (2018) Perturbation-theory and machine learning (PTML) model for high-throughput screening of parham reactions: experimental and theoretical studies. J Chem Inf Model 58:1384–1396
40. Liu H, Deng J, Luo Z, Lin Y, Merz KM Jr, Zheng Z (2020) Receptor-ligand binding free energies from a consecutive histograms monte carlo sampling method. J Chem Theory Comput 16:6645–6655
41. Cabeza de Vaca I, Qian Y, Vilseck JZ, Tirado-Rives J, Jorgensen WL (2018) Enhanced monte carlo methods for modeling proteins including computation of absolute free energies of binding. J Chem Theory Comput 14:3279–3288
42. Cole DJ, Tirado-Rives J, Jorgensen WL (2014) Enhanced monte carlo sampling through replica exchange with solute tempering. J Chem Theory Comput 10:565–571
43. Bajusz D, Rácz A, Héberger K (2015) Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7:1–13
44. Škuta C, Cortés-Ciriano I, Dehaen W, Kříž P, van Westen GJ, Tetko IV, Bender A, Svozil D (2020) QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. J Cheminform 12:1–16
45. Cortes-Ciriano I, Firth NC, Bender A, Watson O (2018) Discovering highly potent molecules from an initial set of inactives using iterative screening. J Chem Inf Model 58:2000–2014
46. Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW (2009) How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model 49:108–119
47. Wagner AB (2006) SciFinder scholar 2006: an empirical analysis of research topic query processing. J Chem Inf Model 46:767–774
48. Ridley DD (2000) Strategies for chemical reaction searching in SciFinder. J Chem Inf Comp Sci 40:1077–1084
49. Carracedo-Reboredo P, Corona R, Martinez-Nunes M, Fernandez-Lozano C, Tsiliki G, Sarimveis H, Aranzamendi E, Arrasate S, Sotomayor N, Lete E, Munteanu CR, Gonzalez-Diaz H (2020) MCDCalc: markov chain molecular descriptors calculator for medicinal chemistry. Curr Top Med Chem 20:305–317
50. Pesciullesi G, Schwaller P, Laino T, Reymond J-L (2020) Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. Nat Commun 11:4874
51. Smith JS, Nebgen BT, Zubatyuk R, Lubbers N, Devereux C, Barros K, Tretiak S, Isayev O, Roitberg AE (2019) Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. Nat Commun 10:2903
52. Grambow CA, Li Y-P, Green WH (2019) Accurate thermochemistry with small data sets: a bond additivity correction and transfer learning approach. J Phys Chem A 123:5826–5835
53. Sun G, Sautet P (2019) Toward fast and reliable potential energy surfaces for metallic Pt clusters by hierarchical delta neural networks. J Chem Theory Comput 15:5614–5627
54. Feuz KD, Cook DJ (2015) Transfer learning across feature-rich heterogeneous feature spaces via feature-space remapping (FSR). ACM T Intel Syst Tec 6:1–27

55. Grazioli G, Roy S, Butts CT (2019) Predicting reaction products and automating reactive trajectory characterization in molecular simulations with support vector machines. J Chem Inf Model 59:2753–2764
56. Charpentier A, Mignon D, Barbe S, Cortes J, Schiex T, Simonson T, Allouche D (2018) Variable neighborhood search with cost function networks to solve large computational protein design problems. J Chem Inf Model 59:127–136
57. Abramyan TM, An Y, Kireev D (2019) Off-pocket activity cliffs: a puzzling facet of molecular recognition. J Chem Inf Model 60:152–161
58. Endo K, Yuhara D, Yasuoka K (2022) Efficient monte carlo sampling for molecular systems using continuous normalizing flow. J Chem Inf Model 18:1395–1405
59. Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM T Model Comput S 8:3–30
60. Moreau A, Couture A, Deniau E, Grandclaudon P (2005) Construction of the six-and five-membered Aza-heterocyclic units of the isoindoloisoquinolone nucleus by parham-type cyclization sequences-total synthesis of nuevamine. Eur J Org Chem 2005:3437–3443
61. Akiyama T (2007) Stronger brønsted acids. Chem Rev 107:5744–5758
62. Akiyama T, Mori K (2015) Stronger brønsted acids: recent progress. Chem Rev 115:9277–9306
63. Caballero-García G, Goodman JM (2021) N-Triflylphosphoramides: highly acidic catalysts for asymmetric transformations. Org Biomol Chem 19:9565–9618
64. Nakashima D, Yamamoto H (2006) Design of chiral N-triflyl phosphoramide as a strong chiral brønsted acid and its application to asymmetric diels– alder reaction. J Am Chem Soc 128:9626–9627
65. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
66. Pogány P, Arad N, Genway S, Pickett SD (2018) De novo molecule design by translating from reduced graphs to SMILES. J Chem Inf Model 59:1136–1146
67. Toropov AA, Toropova AP, Benfenati E, Leszczynska D, Leszczynski J (2010) SMILES-based optimal descriptors: QSAR analysis of fullerene-based HIV-1 PR inhibitors by means of balance of correlations. J Comput Chem 31:381–392
68. Reid JP, Ermanis K, Goodman JM (2019) BINOPtimal: a web tool for optimal chiral phosphoric acid catalyst selection. Chem Commun 55:1778–1781

Publisher's Note