On the prediction of Mayr's nitrogen nucleophilicity parameter by a mixed conceptual density functional theory and atoms-in-molecules approach

Victoria Castor-Villegas, Vincent Tognetti,* Laurent Joubert*

Institut CARMEN UMR6064 (ex lab. COBRA), Université de Rouen, INSA Rouen, CNRS, 1 Rue Tesnière, Mont-Saint-Aignan, 76821 Cedex France.

Abstract

In this work, we propose a robust theoretical framework to predict experimental nucleophilicity values of nitrogen-containing molecules by combining quantum chemical descriptors derived from conceptual density functional theory and the quantum theory of atoms-in-molecules with machine learning algorithms. We assess the performances of several models —including support vector regression, Gaussian process regression, and gradient boosting decision trees— on Mayr's reactivity database, a widely recognised reference for experimental reactivity data based on kinetics measurements, acting on both global and atomic properties. Our results demonstrate strong predictive performance, highlighting the effectiveness of integrating quantum chemical descriptors with machine learning techniques for modelling nucleophilicity. Beyond numerical predictions, our approach also offers insight into the underlying chemical families, providing a richer characterisation of nucleophilic behaviour and enlightening the importance of neighbouring atoms to account for the reactivity of a given atomic site.

Keywords: nucleophilicity, conceptual density functional theory, quantum theory of atoms-in-molecules, machine learning, chemical reactivity.

Corresponding authors: vincent.tognetti@univ-rouen.fr, laurent.joubert@univ-rouen.fr

1 Introduction

The concept of nucleophilicity, first introduced by Ingold [1], constitutes a cornerstone for interpreting reaction mechanisms and guiding synthetic strategies in organic chemistry. It characterises the intrinsic propensity of a molecule to donate electrons in chemical reactions, directly influencing reaction kinetics. Despite its fundamental importance, accurately quantifying nucleophilicity remains challenging due to its intricate dependence on molecular structure, electronic effects, solvation, and reaction conditions.

Fundamentally, reactivity is a kinetic phenomenon defined by the tendency of a species to undergo a chemical reaction, with higher reactivity corresponding to a larger rate constant for a specified elementary reaction [2]. Quantitative reactivity scales, such as those developed by Mayr and co-workers, provide robust frameworks by correlating experimentally determined kinetic parameters to each individual reactants, thus facilitating comparative assessments of nucleophiles and electrophiles [3, 4]. More specifically, Mayr's double-scale approach, given by

$$\log(k_{20^\circ C}) = S_N(N_{Mayr} + E_{Mayr}), \tag{1}$$

quantitatively relates nucleophilicity N_{Mayr} and electrophilicity E_{Mayr} indices to reaction rates measured at 20°C by various spectroscopic experiments (while S_N is the nucleophile-specific sensitivity parameter, which will not be studied here). This strategy broadened the applicability of kinetic studies significantly, ensuring measurable reactions within practical rate limits (10^{-5} to 10^{8} $\mathrm{M}^{-1}\mathrm{s}^{-1}$), but it is actually far from being straightforward, requiring refined techniques and non-negligible experimental time. For such reasons it quickly captured the attention of computational chemists [4–6].

Indeed, quantum chemical methods offer two notable advantages: firstly, it may accelerate research through computational efficiency and ease of automation, surpassing experimental methods in terms of speed and resource economy; secondly, it may provide deeper mechanistic insights by uncovering electronic and structural details that are often inaccessible experimentally [7]. Theoretical predictions frequently make use of the Eyring equation, which relates calculated activation free energies in solution ($\Delta G^\dagger_{\mathrm{sol}}$) to experimentally determined rate constants k [8]:

$$k(T) = \kappa \frac{k_B T}{h c^\circ_{\mathrm{sol}}} \exp\left(-\frac{\Delta G^\dagger_{\mathrm{sol}}}{RT}\right), \tag{2}$$

where key physical constants and standard-state conditions are explicitly defined and where κ is the transmission coefficient. Nonetheless, achieving accurate predictions of $\log(k)$ using quantum chemical methods remains a formidable challenge, described metaphorically as the "jigsaw puzzle of quantum chemistry", primarily due to computational approximations inherent in electronic structure methods and solvation models, complicating the attainment of chemical accuracy (commonly defined within 1 kcal mol^{-1}) [7, 9, 10].

From the methodological point of view, substantial research has been dedicated to evaluating various exchange-correlation (XC) functionals for predicting kinetic parameters relevant to nucleophilicity. Hybrid functionals such as B3LYP, M06-2X, and ωB97X-D have frequently been employed due to their balanced treatment of electron exchange and correlation, significantly impacting the accuracy of computed reaction barriers and energetics [11–16]. These functionals generally offer improved performance over standard Generalised Gradient Approximation (GGA) functionals, especially for reaction energies and barrier heights, due to the inclusion of exact exchange. Double-hybrid functionals, which combine exact exchange and perturbation theory correlation (e.g., B2PLYP), represent a further step towards achieving higher accuracy by explicitly accounting for electron correlation effects more comprehensively, though at a higher computational cost [17].

Besides this electronic perspective, thermal and entropic contributions are other crucial aspects influencing kinetic predictions. Typically, theoretical calculations approximate these contributions at standard conditions of 298.15 K (25°C), even though Mayr's reactivity scale defines kinetic parameters at 20°C. This slight discrepancy introduces additional complexity and potential inaccuracies [12–16]. While many researchers adhere to this standard approximation due to computational convenience, Wang et al. notably employed the exact reference temperature of 20°C, aligning directly with Mayr's experimental framework, thereby improving the direct comparability between theoretical predictions and experimental data [11].

Solvation effects constitute another significant source of complexity (Mayr has regularly emphasized that N_{Mayr} and S_N are solvent-dependent), critically affecting the reaction environment and thus reaction kinetics. For instance, the N_{Mayr} value for the 1-methyl-imidazole has beend found equal to 11.9 in acetonitrile and to 9.9 in water solution. To address this point, theoretical studies have predominantly employed implicit solvation models, such as Polarizable Continuum Models (PCM) [18], due to their computational efficiency and reasonable accuracy for capturing bulk solvent effects [11, 13–17].

Nonetheless, recent studies highlight the limitations of purely implicit treatments for two main reasons: *i)* the accurate treatment of entropy usually computed from ideal gas formulas (see, for instance our recent work on the topic [19]), *ii)* in the presence of specific solvent-solute interactions (for instance hydrogen bonding) that significantly influence reaction pathways. Such limitations have prompted increasing interest in explicit solvation models or hybrid explicit-implicit approaches, which can better capture these specific interactions, albeit at increased computational expense [14–16].

Moreover, accurately accounting for conformational complexity remains essential yet challenging. Reactants and transition states often exhibit multiple accessible conformations, each associated with distinct free energies that collectively influence the reaction barrier and overall rate constants. Numerous recent studies have thoroughly explored the conformational space using robust molecular mechanics force fields such as MMFF94x and OPLS3, identifying the lowest-energy conformers to represent reaction intermediates and transition states in subsequent quantum chemical calculations [12, 14–16].

However, selecting only the global minimum conformation neglects energetically close conformations that could significantly contribute at equilibrium. A more rigorous approach involves population-weighted averaging of multiple conformations, considering Boltzmann distributions to yield more accurate free energy profiles and kinetic predictions [7]. Another method is to perform molecular dynamics (MD) simulations [20] that obviously considerably increase the computational time. MD is also the method of choice to estimate the transmission coefficient, but many trajectories are then necessary to evaluate the recrossing even probability with a low statistical uncertainty.

From all the previous points, it clearly appears that accurately computing reaction rates in solution in order to theoretically predict Mayr's parameters remains an elusive tasks owing to the inherent complexity that governs even so "simple" chemical reactions such as nucleophilic additions. One thus then wonders whether this should be achieved using another paradigm, namely Machine learning (ML), which is designed for dealing with complexity issues. ML techniques have actually emerged as powerful complements to conventional quantum chemical approaches, further enhancing predictive accuracy and providing novel interpretative insights into reactivity trends.

Nonlinear ML methods, including neural networks (NNs), Gaussian process Regression (GPR), and decision-tree-based algorithms like Gradient Boosting Decision Trees (GBDT), have demonstrated considerable potential when trained on comprehensive datasets comprising diverse chemical descriptors and experimentally validated kinetic parameters [21, 22]. While earlier approaches relying on Support Vector Machines (SVMs) yielded comparatively poor results due to limitations in capturing highly nonlinear and complex relationships, recent advances in feature selection and model optimisation have significantly improved the reliability and interpretability of ML-driven kinetic predictions [23]. The field is so active that many papers are regularly published on this topic [24–26], so that we do not consider here to make an extensive review on the use of ML to investigate chemical reactivity.

In the present paper, less ambitiously, we intend to show that the general method that we developed and applied (within linear and non-linear ML approaches) to predict electrophilicity using quantum descriptors [27–29] can be efficiently used to predict nucleophilicity across structurally and electronically diverse nucleophile classes. The peculiarity of our approach is that it is based on the consistent combination of two theories that shares the same basic ingredient, namely Conceptual Density Functional Theory (C-DFT) [30, 31] and the Quantum Theory of Atoms-in-Molecules (QTAIM) [32, 33], both based on the electron density function $\rho(\vec{r})$. A special care will also be devoted to the chemical interpretation of the features involved in the most promising models.

For instance, in a seminal paper, Orlandi and co-workers have critically assessed various molecular descriptors, emphasising that no single parameter —such as the energy of the Highest Occupied Molecular Orbital (HOMO)— can comprehensively capture nucleophilicity. They advocated for multidimensional regression models integrating electronic, steric, and solvation descriptors, wherein each effect is explicitly represented by tailored descriptors to achieve reliable predictive accuracy [34, 35]. Importantly, they also calculated the protonated products, as molecular models of the results of a nucleophilic attack, thereby probing how post-attack stabilisation influences the observed rate constants. Nonetheless, consistently quantifying steric effects across chemically distinct nucleophiles, such as olefins versus N-heterocyclic carbenes or amines, remains problematic due to intrinsic structural differences [34].

Moreover, correlations between nucleophilicity and thermodynamic parameters such as experimental pK_a values further highlighted the intricate interplay between kinetic and thermodynamic properties. These correlations are nuanced and significantly modulated by solvent effects and steric hindrance, complicating attempts to derive straightforward quantitative relationships [36–38]. Such complexities underscore the need for comprehensive, multifaceted computational approaches to reliably predict nucleophilicity indices.

To address the inherent complexities and uncertainties in nucleophilicity predictions, recent advancements have introduced uncertainty quantification (UQ) methodologies into Mayr's reactivity. Proppe et al. demonstrated that integrating UQ allows computational chemists to report theoretically calculated reaction rates ($\log k$) in a format directly analogous to experimental measurements, namely as expectation values accompanied by corresponding deviations or uncertainties [39].

Such an approach significantly enhances the transparency and reliability of theoretical predictions, explicitly acknowledging the limitations and approximations intrinsic to quantum chemical methods. Moreover, incorporating uncertainty estimates not only strengthened confidence in predicted kinetic parameters but also opened novel avenues for systematically benchmarking different computational approaches against experimental datasets, even in cases where experimental data is sparse or currently unavailable [40, 41].

The recent work by Liu and colleagues further represents a substantial advancement in the field by compiling the most extensive dataset to date for predicting Mayr's nucleophilicity (N_{Mayr}) and electrophilicity (E_{Mayr}) parameters using machine learning methods. Their dataset comprises 1115 nucleophilicity parameters alongside 285 electrophilicity parameters, encompassing chemically diverse compounds with varied nucleophilic reaction centres [23].

Recognising the critical role of solvent environments in influencing nucleophilicity, their comprehensive approach incorporated multiple sophisticated solvent descriptors. Specifically, they included solvent parameters such as Reichardt's, Laurence, Kamlet-Abboud-Taft, Catalán and Hansen parameters [42–48]. These diverse features collectively capture various physicochemical solvent characteristics, significantly improving the predictive capabilities and interpretability of ML models.

In their methodological approach, Liu et al. also systematically evaluated a broad range of molecular descriptors using Random Forest (RF) algorithms to identify the most informative descriptors for accurate predictions of nucleophilicity indices. However, their analysis also highlighted critical limitations associated with certain ML approaches, such as SVR, which provided relatively poor predictive performance compared to other methods. This emphasises the importance of carefully selecting appropriate ML techniques and underscores the necessity of rigorous descriptor selection processes to ensure optimal model accuracy and interpretability [23].

The benefits of an accurate and efficient model are even larger. Indeed, despite significant methodological advances, the experimental determination of nucleophilicity indices remains challenging for specific classes of highly reactive or structurally unstable nucleophiles —such as tertiary enamines— which highlights a crucial area where computational predictions can offer substantial benefits, particularly in scenarios where direct kinetic measurements are experimentally impractical or impossible [49–52].

In order to adress some of this issues, our paper will be thus divided as follows. In the next section, we provide an overview of the quantum chemical descriptors, the ML techniques we used, and the computational details. Then the database will be presented, before reporting the results and discussing them.

2 Theory

2.1 Quantum Chemical Descriptors

The C-DFT descriptors can be classified into three categories: *global* (one value for the whole molecule), *local* (functions $a(\vec{r})$ that depend only on one position in real-space, allowing for discussing regioselectivity), *non-local* (functions $b(\vec{r}, \vec{r}', ...)$ that depend on more than one 3D-space variables, also referred as kernels) ones. All of them can be also categorised as being either *basic* or *composite*, the first class corresponding to descriptors that are directly expressed as derivatives of the electronic energy E_e with respect to the two variables describing the system in the chosen ensemble representation [53]. Conversely, composite descriptors are obtained by combination of basic descriptors.

For instance, in the canonical ensemble, E_e is expressed as a function of the number of electrons N and as a functional of the external potential, $v(\vec{r})$, generated by the nuclei. The electronic chemical potential μ and molecular hardness η are basic global descriptors defined by

$$\mu = \left(\frac{\partial E_e[N, v(\vec{r})]}{\partial N} \right)_{v(\vec{r})}, \tag{3}$$

$$\eta = \left(\frac{\partial^2 E_e[N, v(\vec{r})]}{\partial N^2} \right)_{v(\vec{r})} \tag{4}$$

while Chattaraj's nucleophilicity index [54] which is a composite global descriptor derived from these quantities is defined as:

$$N_C = \frac{2\eta}{\mu^2}. \tag{5}$$

Still in the canonical ensemble, important local descriptors are the Fukui functions, which quantify the change in electron density upon addition or removal of an electron, thereby pinpointing the most nucleophilic and electrophilic sites in a molecule.

$$f^{\pm}(\vec{r}) = \frac{\partial}{\partial N}\left(\frac{\delta E_e}{\delta v(\vec{r})}\right)_{\substack{\partial N > 0, \\ \partial N < 0}} \tag{6}$$

Originally developed to remedy the fixed-orbital picture of FMO theory [55], these functions admit a simple approximation in terms of the HOMO and LUMO when orbital relaxation is neglected:

$$f^{+}(\vec{r}) \approx |\phi_{\mathrm{LUMO}}(\vec{r})|^2, f^{-}(\vec{r}) \approx |\phi_{\mathrm{HOMO}}(\vec{r})|^2. \tag{7}$$

Additionally, the non-local linear response kernel $\chi(\vec{r}, \vec{r}\,')$, extensively studied in C-DFT in the last years, describes how the electron density at position $\vec{r}$ responds to an external perturbation at other position $\vec{r}\,'$, revealing various chemical phenomena, including electron delocalization, inductive and mesomeric effects, and aromaticity reactivity[56]:

$$\chi(\vec{r}, \vec{r}\,') = \left(\frac{\delta^2 E_e[N, v(\vec{r})]}{\delta v(\vec{r})\delta v(\vec{r}\,')}\right)_N \tag{8}$$

Then, in order to condense the information embodied in local or non-local descriptors, QTAIM [32, 33] has been used. Condensation can be seen as information coarse-graining, going from an infinite amount of data to a finite one (here, related to the number of atoms). It also allows for translating the information embodied by such descriptors into the usual chemical language based on "sites". In a nutshell, QTAIM partitions, based on the field lines of the electron density gradient vector, the 3D-real space in non-overlapping domains that each corresponds (in the absence of non-nuclear attractors) to a topological atomic basin Ω_A. *Atomic* and *di-atomic values* features can thus be computed for any local $a(\vec{r})$ or non-local $b(\vec{r}, \vec{r'})$ functions by:

$$a(A) = \int_{\Omega_A} a(\vec{r}) d^3r \tag{9}$$

$$b(A, B) = \int_{\Omega_A}\int_{\Omega_B} b(\vec{r}, \vec{r'}) d^3r d^3r' \tag{10}$$

Whereas the canonical ensemble treats the particle number N as the privileged extensive variable the grand-canonical ensemble replaces N by its conjugate, the chemical potential μ, through a Legendre transformation. This operation generates the grand potential $\Omega(\mu, v(\vec{r}))$ and recasts all derived quantities. The transformation is conceptually identical to the shift from the internal energy $U(S, V)$ to the enthalpy $H(S, P)$. In this ensemble, the local softness functions, defined by

$$s^{\pm}(\vec{r}) = f^{\pm}(\vec{r})/\eta, \tag{11}$$

are basic descriptors while they are composite ones in the canonical representation.

Finally, since our interest spans reactivity at defined temperatures, conformational sampling must be considered. For a property P, the following Boltzmann-weighted average $\overline{P}$ across the stable conformers k will be computed:

$$\overline{P} = \frac{\sum_k e^{-\frac{E_k^{SCF}}{RT}} P_k}{\sum_k e^{-\frac{E_k^{SCF}}{RT}}} \tag{12}$$

2.2 Machine Learning

When applied to chemical reactions, ML mainly serves two complementary purposes: *i*) classification to reveal latent structure in the molecular-descriptor space, and *ii*) regression to predict reactivity metrics such as the nucleophilicity parameter N_{Mayr} or the family. The following summary emphasises the algorithms actually employed in this study.

All descriptor matrices were centred and scaled to unit variance. Missing values, present in $< 2\,\%$ of the entries, were neglected (non-neutral species were excluded; see Section 3). Highly correlated features ($> 95\,\%$) were pruned to reduce redundancy, and the remaining variables were ranked via mutual information against the target [57]. The final feature set thus balances information content with model parsimony.

2.2.1 Classification Algorithms

Support Vector Machine (SVM) classifier. For supervised classification we use the C-SVC formulation of support vector machines [58], as implemented in `scikit-learn`'s `svm.SVC`, which wraps LIBSVM [59]. Given labelled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$ with $y_i \in \{-1, +1\}$, the soft-margin primal problem is:

$$\min_{\mathbf{w},\, b,\, \boldsymbol{\xi}} \ \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{such that} \quad y_i\big(\mathbf{w}^{\top}\phi(\mathbf{x}_i) + b\big) \geq 1 - \xi_i,\ \xi_i \geq 0. \tag{13}$$

Kernelisation yields the dual

$$\max_{\boldsymbol{\alpha}} \ \sum_{i=1}^{n} \alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{n} \alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad \text{such that} \quad 0 \leq \alpha_i \leq C,\ \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{14}$$

The decision function is

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b. \tag{15}$$

We tune C and kernel hyperparameters by cross-validation.
Note: Support Vector *Clustering* [60] is an unsupervised method and is not used here.

Choice of kernel. The kernel $K(\mathbf{x}, \mathbf{x}')$ determines the geometry of the feature space and therefore the shape of the decision boundary. We consider:

- **Linear kernel**: $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$ Equivalent to no transformation; effective when data are linearly separable.
- **Radial Basis Function (RBF) kernel**: $K(\mathbf{x}, \mathbf{x}') = \exp\bigl(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\bigr)$ Maps input to an infinite-dimensional feature space. Suitable for nonlinearly separable data; parameter $\gamma > 0$ controls the radius of influence of support vectors.

2.2.2 Regression Algorithms

Random Forest Regressor (RFR). A RF is an ensemble of T decision-tree regressors $\{h_t(\vec{x})\}_{t=1}^T$ grown on independent bootstrap samples of the training set [61]. At each split, a random subset of $m<p$ features is considered, decorrelating trees and reducing variance. The final prediction is the arithmetic mean

$$\hat{y}(\vec{x}) = \frac{1}{T} \sum_{t=1}^{T} h_t(\vec{x}). \tag{16}$$

Individual trees minimise the squared-error impurity $\mathcal{I}(S) = \sum_{(\vec{x}_i, y_i) \in S} (y_i - \bar{y}_S)^2$ within every leaf S, ensuring piecewise-constant approximations that capture nonlinear interactions without explicit functional assumptions.

Support Vector Regression (SVR). Employing the same kernel $K(\mathbf{x}, \mathbf{x}\,') = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$ used above for *classification*, ε-SVR seeks a function $f(\mathbf{x}) = \mathbf{w}^\top \varphi(\mathbf{x}) + b$ that deviates from targets by at most ε while penalising slack variables ξ_i, ξ_i^* outside the tube [62]:

$$\min_{\vec{w}, b, \xi_i, \xi_i^*} \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \quad \text{subject to} \quad \bigl| y_i - f(\vec{x}_i) \bigr| \le \varepsilon + \xi_i,\ \xi_i, \xi_i^* \ge 0. \tag{17}$$

The representer theorem yields $f(\vec{x}) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b$, where α_i, α_i^* are Lagrange multipliers obtained from the dual quadratic programme.

Feed-Forward Neural Network (FFNN). A fully connected network defines a hierarchical composition of affine maps and element-wise nonlinearities σ:

$$\hat{y}(\vec{x}) = \vec{W}^{(L)} \sigma\bigl(\vec{W}^{(L-1)} \sigma(\cdots \sigma(\vec{W}^{(1)} \vec{x} + \vec{b}^{(1)}) + \vec{b}^{(L-1)})\bigr) + \vec{b}^{(L)}, \tag{18}$$

where $\{\vec{W}^{(\ell)}, \vec{b}^{(\ell)}\}_{\ell=1}^L$ are learned parameters. Training minimises the mean-squared error $\mathcal{L} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}(\vec{x}_i))^2$ via stochastic gradient descent with back-propagation, optionally regularised by dropout or ℓ_2 weight decay to curb overfitting.

Gaussian Process Regression (GPR). We place a Gaussian process prior $f \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$ with $m \equiv 0$ [63]. For training inputs $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top$ and targets $\mathbf{y} \in \mathbb{R}^n$ observed with i.i.d. Gaussian noise variance σ^2, define the kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ by $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{k}_* = [\, k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_n, \mathbf{x}_*)\,]^\top$. Then the predictive distribution at $\mathbf{x}_*$ is Gaussian with

$$\mu_* = \mathbf{k}_*^\top \big(\mathbf{K} + \sigma^2 \mathbf{I}\big)^{-1} \mathbf{y}, \tag{19}$$

$$\sigma_*^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top \big(\mathbf{K} + \sigma^2 \mathbf{I}\big)^{-1} \mathbf{k}_*. \tag{20}$$

Kernel hyperparameters and σ^2 are estimated by maximising the log marginal likelihood. We report point predictions μ_* with credible intervals $\pm 2\sigma_*$.

All algorithms were implemented using `scikit-learn` [64] (`v` 1.5) with default settings unless stated otherwise. Source code and hyperparameter grids are provided in the Supporting Information.

2.3 Computational Details

All DFT calculations were performed with the ADF software [65, 66] using the PBE exchange-correlation functional [67] with the Grimme's dispersion correction with Becke-Johnson damping function (D3-BJDAMP) [68] and a triple-ζ basis set (TZP), with no relativistic effects added. We used the default settings for the numerical quality and integration grid, but we did not use the frozen core approximation.

The choice for a dispersion-corrected Generalized Gradient Approximation (GGA) functional was mainly governed by its significant trade-off between accuracy and computational time. Even if it is certainly outperformed by more modern approaches such as double hybrids, it is more suited for a large number of calculations (due to the large dataset considered here and to the inclusion of conformational effects). Besides, evaluating the C-DFT and QTAIM features at the best level of theory is not necessary since their computation is intended only to feed the ML models, which are able, in principle, to take into account and to indirectly correct the initial GGA deficiencies.

The automation of the computational workflow was achieved by a Python script, inspired by recipes already available in the Python Library for Automating Molecular Simulations (PLAMS) [69]. The pipeline is made of the following steps:

i) The RDKit [70] utility generates the 3D coordinates from the sole Simplified Molecular Input Line Entry System (SMILES) code and generates a set of molecular geometries to explore the conformational space;

ii) The obtained geometries are optimised at the Density-Functional based tight-binding (DFTB) level of theory, and the conformers are selected based on the lowest energy conformers, with a threshold of 5.0 kcal/mol in terms of relative SCF energies and 3.0 Å for the structural Root-mean-Square Deviation (RMSD);

iii) The selected conformers are then optimised at the final DFT level of theory, followed by a second conformer refinement, now with a threshold of 2.5

kcal/mol for energy and 1.5 Å for RMSD;

iv) The retained conformers are then used to calculate the QTAIM and C-DFT descriptors by means of single-point calculations;

v) All descriptors are then Boltzmann-averaged to afford the mean values according to eq. 12. For the sake of comparison, we also saved the numerical values for all descriptors corresponding to the most stable conformer.

Importantly, all descriptors used in this work were computed solely for the reactant structures, without introducing any modified or protonated or other activated forms to mimic the products of nucleophilic attack. This intrinsic analysis thus focuses exclusively on the electronic and topological information intrinsically present in the reactant molecule, as captured through QTAIM and C-DFT. By avoiding structural modifications or reaction intermediates, the models aim to predict reactivity from the chemical ground state, with the only additional layer being the consideration of conformers accessible at 20°C under solvation conditions.

Solvent effects were described using the implicit Conductor-like Screening Model (COSMO) [71, 72], with default parameters applied in most cases. In a few instances involving solvent mixtures —such as MeOH/MeCN (45/55 or 91/9), or EtOH/MeCN (91/9)— the relevant parameters (solvent radii and relative permittivity) were explicitly specified, withvalues computed as weighted averages based on the composition of the pure solvents.

Descriptors used for the analysis were provided by the ADF software with the *Analysis Level* set to *Full*, which enabled the calculation of global and local descriptors based on C-DFT. The computed descriptors include the electronic chemical potential (μ), electronegativity (χ), hardness (η), softness ($S = 1/\eta$), electrophilicity index ($\omega = 1/N_C$), nucleofuge, electrofuge, electrodonating power, and electroaccepting power [73–75], all of them defined as global descriptors.

For the canonical atomic descriptors, we have the Fukui functions ($f^{\pm}$, f^0), and dual descriptor (DD) $f^{(2)}$ [76] in the canonical ensemble. In the grand canonical ensemble, $s^{\pm}$ and s^0. We also computed composite Fukui functions such as μ^+f^+, μ^-f^-, μf^+, composite DDs such as $\omega f^{(2)}$, $Sf^{(2)}$, $S^2f^{(2)}$, supplemented by condensed local electrophilicity and nucleophilicity. For non-local descriptors, we considered the condensed linear response $\chi(A, B)$. For the sake of simplicity, $\chi(A)$ denotes the value when A and B refer to the same atom. Additionally, we computed the recently revivified P and $Pf^{(2)}$ [53], not provided in the ADF output, where $P = \gamma^{-1}$.

Finally, atomic charges, denoted q, were evaluated using both the QTAIM and Hirshfeld partitioning schemes.

The ML models were implemented using `scikit-learn` in Python. The data were split into 80 % for training and 20 % for validation with the corresponding functions in `scikit-learn`, for all cases the splitting was done only one time with a random state for the split. Model optimization was performed using the minimization routine in `scipy`, with the ML model as the objective function and the L-BFGS-B method.

3 The Molecular Database

For the present work, we employed the full set of Mayr's reported nucleophilicity values for nitrogen-containing compounds for neutral molecules (indeed, modelling cations or anions in solution is not straightforward since it may require considering the counterion or the use of an explicit solvent layer). After this filtering process, a total of 244 systems containing nucleophilic nitrogen atoms were retained from the families (named following Mayr's typology): *i*) aliphatic amines, *ii*) aromatic amines, *iii*) amidines and imines, *iv*) amino acids, *v*) azoles and azoles anions, *vi*) guanidines, *vii*) hydrazines, hydroxylamines, etc., *viii*) imidazolines and related compounds, *ix*) isothioureas, *x*) pyridines, quinolines, etc. *xi*) other N-centered nucleophiles.

The dataset displays a broad chemical diversity, encompassing nearly all common organic functional groups. Notably, several nucleophiles contain multiple nitrogen atoms, as illustrated in Figure 1. In trimethylhydrazine, for instance, the two nitrogen atoms are chemically distinct, and both nucleophilic sites have been experimentally characterised, with separate N_{Mayr} values reported in the database. Conversely, in the guanidine derivative, only the non-cyclic nitrogen has a reported nucleophilicity value; the two cyclic nitrogen atoms, being equivalent in this context, are considered non-reactive. For azoles, a single N_{Mayr} value is given, which represents the reactivity of both nitrogen atoms involved in the nucleophilic attack.

This supports the interpretation of Mayr's scale as a site-specific rather than a purely molecular descriptor. Throughout this work, whenever a nitrogen atom within a multi-centre system lacks an associated N_{Mayr} value, it is treated as non-nucleophilic by default.

From the numerical point of view, the range for nucleophilicity values for the reactive systems spans a large range from 5 to 23, with a notably unimodal Gaussian-like distribution with a maximum frequency near 14 with a standard deviation of 3.2. Besides, we also considered the 1310 nucleophilic systems reported in Mayr's database to analyse if there is any particular trend related with the nuclei. As shown in Figure 2, there is a significant overlap between them, precluding any simple and direct correlation between the atomic number and the nucleophilicity.

There is thus no intrinsic atomic hierarchy: a nitrogen atom can be more or less nucleophilic than a carbon or a hydrogen, depending on its environment. Clearly, the other atoms like H, C, O shows nucleophilicity distributions that are much more spread and of a clear multimodal character (for instance one peak around 5 and another one around 18 for oxygen). A more detailed analysis of these trends is however outside the scope of this paper.

It should be noticed that we take all systems without considering the "star classification system" for any splitting of the data (at variance with some of our previous works). Let us recall that this classification is linked to the number of the reference electrophiles/nucleophiles used for the determination of the reactivity parameter. Hence, it constitutes a kind of evaluation of the reliability/accuracy of the numerical values, which compensates for the lack of

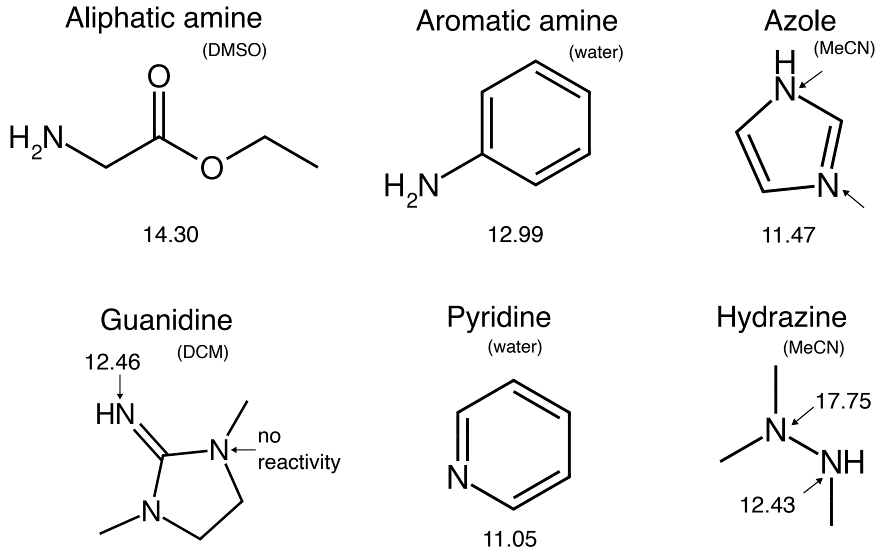


Figure 1: Representative examples of nitrogen-containing nucleophiles. For trimethylhydrazine, distinct N_{Mayr} values are reported for each nitrogen atom. In the guanidine example, only the non-cyclic nitrogen is considered reactive; the two equivalent cyclic nitrogens are treated as non-nucleophilic. In the azole example, a single N_{Mayr} value reflects the concerted involvement of both nitrogen atoms. Experimental solvents are indicated in parentheses.

measured numerical uncertainties. One has thus to keep in mind that one-star systems should be used only to give a good idea about the relative reactivities of strong nucleophiles towards weak electrophiles, and also that two-star systems could be subject to re-evaluation.

4 Results and Discussion

4.1 Discriminating between reactive and unreactive nitrogen atoms

As any chemical feature, reactivity can be discussed from a qualitative and a quantitative perspectives. In ML, the first one can be tackled from the classification point of view, in particular to distinguish between reactive and non-reactive nitrogen atoms in the full dataset. We thus investigated Support Vector Classification (SVC) models (linear and RBF kernels) based on Boltzmann-averaged QTAIM and C-DFT descriptors.

More precisely, in order to identify the most relevant features for classification, the classification procedure exhaustively assessed all pairwise combinations across the full set of computed descriptors. In addition, three-dimensional clas-

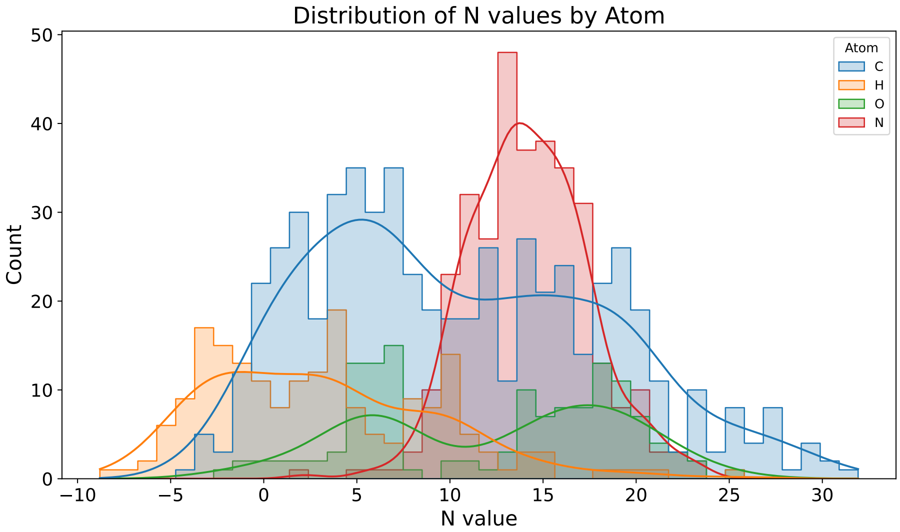


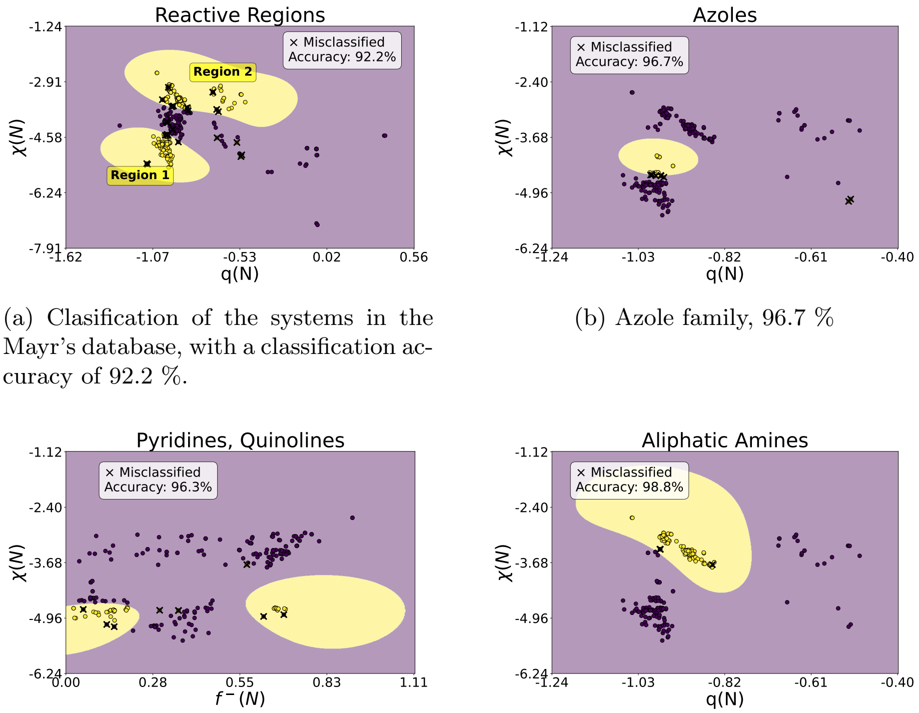
Figure 2: Histogram of the N_{Mayr} values for the systems in Mayr's database.

sification was explored by including a third descriptor, but this approach led to signs of overfitting, with no improvement in classification performance compared to the two-dimensional case.

The optimal partition —maximising classification accuracy— is shown in Figure 3a, achieving an overall accuracy of 92.2 % for the full dataset (i.e., no separation between training and test sets was applied in this exploratory classification). The background colours indicate the decision regions obtained from the classification: the purple region corresponds to the non-reactive class, while the yellow region corresponds to the reactive class. Data points are coloured according to their true class labels according to Mayr's experiments: purple for non-reactive sites and yellow for reactive sites. This allows visual comparison between the predictions and the actual values. For instance, a yellow point in the yellow zone corresponds to a site that is experimentally reactive and that is predicted also reactive by the model. Conversely, a purple point in the yellow zone corresponds to a site that is not experimentally reactive but that was erroneously predictive as reactive. Such misclassified sites are represented by a cross.

It is also valuable to use the same approach by restricting it to each of the three dominant chemical families since they are sufficiently populated to allow meaningful statistical evaluation, namely azoles (33 compounds), pyridines and quinolines (56 compounds), and aliphatic amines (104 compounds). The predictive performance further improved, reaching 99.2 % for azoles, 96.3 % for pyridines and quinolines, and 98.9 % for aliphatic amines. These results were obtained using the RBF kernel.

Among the most effective descriptor combinations, we found those involv-



(a) Clasification of the systems in the Mayr's database, with a classification accuracy of 92.2 %.

(b) Azole family, 96.7 %

(c) Pyridines and quinolines, 96.3 %

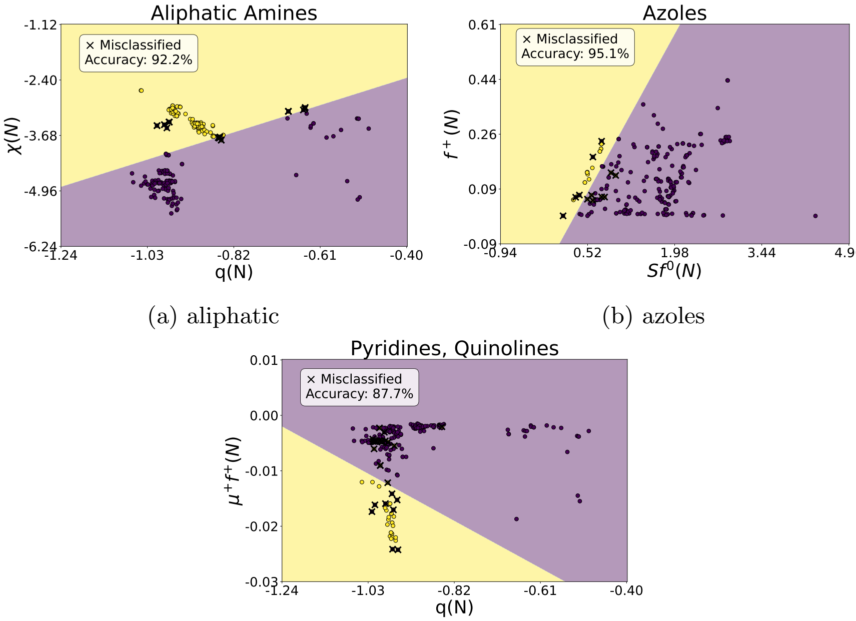
(d) Aliphatic amines, 98.9 %

Figure 3: (a) Classification of all systems, (b) Classification of the Azole family, (c) Classification of the Pyridines and Quinoline families, (d) Classification of the Aliphatic Amines family. Atomic units for all descriptors.

ing the electrostatic potential, condensed linear response descriptors, f^2, $\mu^- f^-$, and f^-. However, the final classification non-linear models reported in Figure 3 are based on the $q(N), \chi(N)$ combination (except for the pyridine family that involves the $q(N)$ and $f^-(N)$). Interestingly, the highest overall accuracy provides also a clear chemical interpretability since it mixes one feature describing charge control ($q(N)$, useful for hard species) and one related to orbital control ($\chi(N)$, well suited to soft species, since high values of χ corresponds to high polarisabilities), this non-linear analysis revealing two well-defined regions associated with reactive and non-reactive nitrogen atoms.

We also tested the linear kernel, but only achieved an overall accuracy of about 95 %, 87 %, 92 %, highlighting the significant non-linear character of the boundaries that separate the chemical families. A visual inspection of the decision regions (Figure 4) further illustrates that the misclassified points are artefacts of the linear assumption rather than genuine overlap between families.

The azole family deserve a particular comment. Indeed, the optimal classification performance was actually obtained when $\chi(N)$ was used in conjunction



(a) aliphatic

(b) azoles

(c) pyridines

Figure 4: Classification of reactive and non-reactive nitrogen sites using the SVC algorithm with the linear kernel.

with $f^+(N)$ that is a descriptor describing electrophilicity and not nucleophilicity, making this model a priori not chemically interpretable. However, one should also notice that the azoles form a compact cluster near zero for this descriptor. This is computationally not satisfying since these values may be more prone to numerical errors. Pleasingly, the already mentioned $q(N), \chi(N)$ combination yielded strong predictive results, and as it is more interpretable, this is the one that we keep for our figure. Indeed, it fails to correctly classify only two compounds: benzotriazole and 1-methylbenzotriazole, which are the only triazoles in the dataset, deviating significantly from the overall trend.

On the other hand, the pyridine and quinoline family remarkably exhibited a clear bifurcation into two distinct clusters, primarily differentiated by the $f^-(N)$ descriptor. One cluster was centred near zero, while the other grouped around a value of approximately 0.74. This separation suggests fundamentally different reactivity mechanisms within this family, despite structural similarity among its members.

Lastly, it should be noted that, although the grand canonical ensemble provides the natural framework for defining C-DFT descriptors in terms of the chemical potential and is often argued to be the appropriate choice, particularly for open or electrochemical systems where particle exchange is relevant [77], some other studies have shown that canonical and grand canonical formulations yield nearly equivalent results in practice [78]. In our case, we find that a canonical description is sufficient to obtain reliable predictive accuracy for the systems under study.

While the use of QTAIM charges combined with the condensed linear response already provides a good classification accuracy of about 92 %, we observed that the description can be further improved by accounting for the contributions of neighbouring atoms. In this approach, the atomic charge and $\chi(A, B)$ are weighted by an exponential decay factor that takes the influence of surrounding atoms into account. For any atomic property P and atom A, it can be achieved by:

$$P_w(A) = P(A) + \sum_{B \neq A} e^{-\alpha_P R_{AB}} q(B), \tag{21}$$

where R_{AB} is the internuclear distance. For instance in the case $P = q$, a negatively charged atom surrounded by several positively charged neighbours will consequently "appear" less negatively charged for another charge approaching it. The previous equation is hence the simplest way to incorporate screening effects. exhibit a reduced effective reactivity. Similarly, for the linear response kernel (and more generally for any kernel):

$$\chi_w(A) = \chi(A) + \sum_{B \neq A} e^{-\alpha_\chi R_{AB}} \chi(A, B), \tag{22}$$

Considering now the $q_w(N), \chi_w(N)$ combination, we optimized the α_q and α_χ decay parameters (a very high value would imply that the neighbouring atoms have only negligible contributions) to maximise the classification accuracy using the SVC approach with the RBF kernel. We obtained a symmetric solution with $\alpha_q = \alpha_\chi = 1.0 \mathring{A}^{-1}$. This choice achieved an improved accuracy of 97.5 %. Moreover, the two distinct regions of predicted reactivity identified in the previous analysis were merged into a single unified region that captures the reactivity for all N atoms, as shown in Figure 5.

Noteworthy, it can be added that the influence of the neighbouring atoms for the linear-response kernel can be taken into account using the eigenvalues of the di-atomic QTAIM condensed kernel, following the methodology recently described by Grincourt et al. [79]. Indeed, these eigenvalues can be mapped to each atom by looking at the highest absolute coefficient in the linear expansion giving the corresponding eigenvectors. We are currently working on that point.

4.2 Predicting Mayr's nucleophilicity values

After showing that our approach was successful in determining if a nitrogen atom is reactive or not, we now attempt to predict Mayr's nucleophilicity. We first considered linear regression models based on descriptors such as atomic charges and HOMO energies, which met with limited success. Indeed, these models yielded a mean absolute error (MAE) of approximately 2.5 and exhibited considerable scatter, lacking any consistent trend or correlation. This result

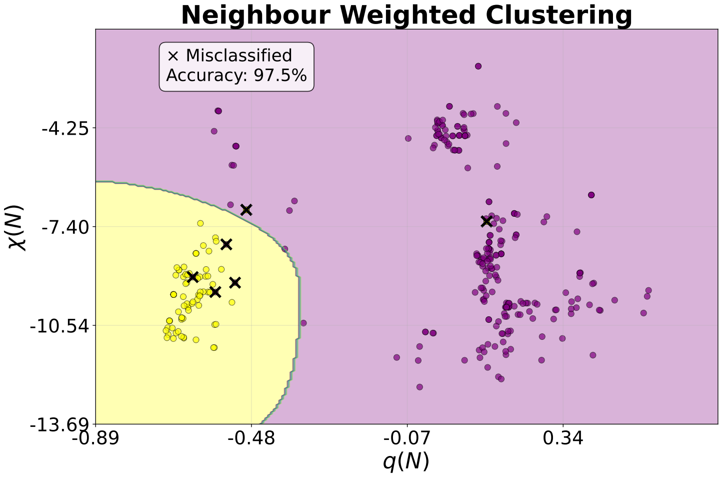


Figure 5: Classification of reactive and non-reactive regions using weighted atomic descriptors for the whole dataset. Atomic units.

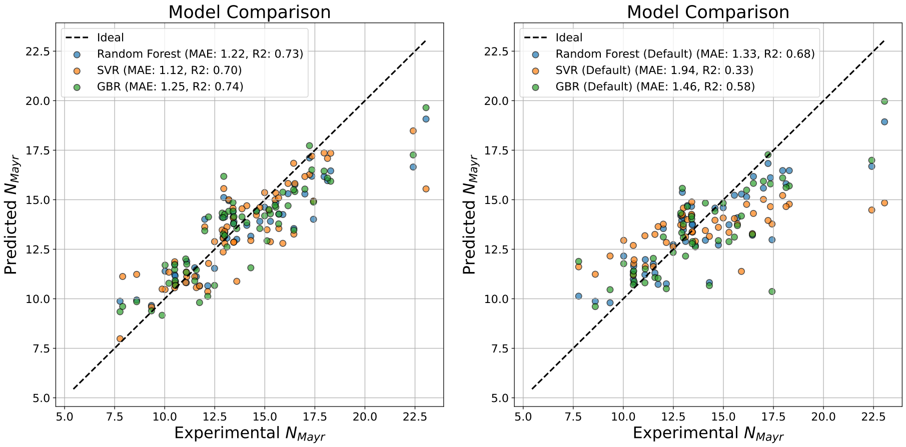
underscores the inadequacy of simple linear relationships in capturing the complexity of nucleophilic behaviour.

To address the limitations observed with traditional regression techniques, advanced ML approaches —including Random Forests, Support Vector Regression (SVR), Gradient Boosted Regression (GBR), and neural networks— were explored. These models significantly improved predictive performance, reducing the mean absolute error (MAE) to approximately 1.2, although some dispersion remained. Nonetheless, the ability of these ML methods to capture underlying chemical trends underscores their value, particularly when applied within chemically meaningful clusters identified earlier.

Figure 6 compares models trained with tuned hyperparameters to their counterparts trained with default settings. Both strategies produce similar error profiles and overall MAE, with hyperparameter optimisation yielding only modest gains (most visibly for SVR). This behaviour suggests that the descriptor space is already informative enough for robust prediction, even without extensive tuning.

A key advantage of the RF models is their feature-importance analysis (Figure 7a), which we used to design compact predictors. For the "small" models trained with default hyperparameters, we retained only the seven most important descriptors identified by RF. In contrast, the tuned models were trained on the full descriptor set to maximise any potential gain from optimisation. To maintain methodological continuity with the classification stage, the reduced set prioritised the $\chi(N)$ which had underpinned our classification analysis and was the descriptor to cut the set, delating also the Hirshfeld charge since it has basically the same weight as the QTAIM charge.

Complementary correlation analysis (Figure 7b) highlighted redundancies



(a) Different ML models

(b) Default values of ML libraries

Figure 6: ML

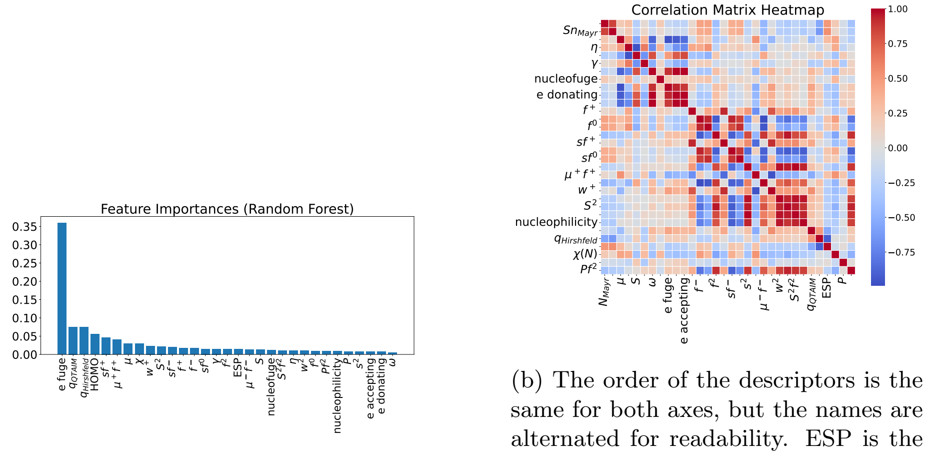
among several descriptors and clarified their relationships with N_{Mayr}. These insights guided the removal of low-contribution or strongly collinear features, streamlining the models without degrading accuracy and thereby reducing computational cost. An interestingly correlation was observed between the descriptors N_{Mayr} and S_N.

Finally, across this dataset, models trained with default hyperparameters were often competitive with their tuned counterparts. While careful optimisation can yield incremental improvements —and clearly outperforms any simple analytical equation— the marginal gains beyond defaults are limited. Taken together, the results support a practical workflow: use RF feature importances to define a compact, seven-descriptor model (preferentially including $\chi(N)$ for consistency with classification), and rely on default settings for rapid, robust predictions of N_{Mayr}.

4.3 Conformers

The difference between using only the most stable conformer and including all conformers accessible at 20°C in solvation proved to be negligible. Variations in the predicted results were comparable to the impact of adding or removing a molecule from the dataset, supporting the robustness of the method with respect to conformational diversity.

It is also important to note that, in the context of classification analysis, the use of all accessible conformers produced results nearly identical to those obtained using only the most stable conformer. This outcome reflects a physically meaningful interpretation. Our classification analysis reveals whether a reaction is thermodynamically favoured —that is, whether it occurs or not— but



(a) Importance of descriptors

(b) The order of the descriptors is the same for both axes, but the names are alternated for readability. ESP is the electrostatic potential at the nucleus.

Figure 7: Impotance of the descriptors and correlation between them.

does not capture kinetic aspects of reactivity, while the N_{Mayr} value is directly related to the reaction rate constant and therefore encodes kinetic information.

5 Conclusions

Although the QTAIM and Conceptual-DFT descriptors do not yet yield a fully satisfactory analytic expression for the nucleophilicity index N_{Mayr} of the nitrogen atoms, our ML model successfully reproduces the global trend of N_{Mayr} across the entire data set. The MAE obtained for N_{Mayr} is larger than that previously reported for the electrophilicity index E_{Mayr}, a difference that is consistent with the intrinsically broader energetic landscape associated with nucleophilic reactivity.

Nevertheless, the present analysis demonstrates that QTAIM charges and the condensed linear response already encode sufficient information to discriminate, between nitrogen atoms that engage in nucleophilic attack and those that remain inert. Even more importantly, this interpretation is reinforced when the influence of the neighbour environment is explicitly taken into account, underscoring the local, yet context-dependent, nature of nucleophilicity.

Acknowledgments

This work was partially supported by Normandie Université, the Région Normandie, the Centre National de la Recherche Scientifique, Université de Rouen Normandie, INSA Rouen Normandie, Université Caen Normandie, ENSICAEN, Labex SynOrg (ANR-11-LABX-0029), the graduate school for research XL-

Chem (ANR-18-EURE-0020 XL CHEM) and Innovation Chimie Carnot (I2C). The Centre Régional Informatique et d'Applications Numériques de Normandie (CRIANN) is warmly acknowledged for providing high-performance computational resources.

References

[1] C. K. Ingold. Significance of tautomerism and of the reactions of aromatic compounds in the electronic theory of organic reactions. *Journal of the Chemical Society (Resumed)*, page 1120, 1933.

[2] P. Muller. Glossary of terms used in physical organic chemistry (iupac recommendations 1994). *Pure and Applied Chemistry*, 66(5):1077–1184, January 1994.

[3] Herbert Mayr and Armin R. Ofial. Do general nucleophilicity scales exist? *Journal of Physical Organic Chemistry*, 21(7-8):584–595, May 2008.

[4] Herbert Mayr. Reactivity scales for quantifying polar organic reactivity: the benzhydrylium methodology. *Tetrahedron*, 71(32):5095–5111, August 2015.

[5] Patricia Pérez, Alejandro Toro-Labbé, Arie Aizman, and Renato Contreras. Comparison between experimental and theoretical scales of electrophilicity in benzhydryl cations. *The Journal of Organic Chemistry*, 67(14):4747–4752, May 2002.

[6] Claus Schindele, K. N. Houk, and Herbert Mayr. Relationships between carbocation stabilities and electrophilic reactivity parameters, e: Quantum mechanical studies of benzhydryl cation structures and stabilities. *Journal of the American Chemical Society*, 124(37):11208–11214, August 2002.

[7] Maike Vahl and Jonny Proppe. The computational road to reactivity scales. *Physical Chemistry Chemical Physics*, 25(4):2717–2728, 2023.

[8] Henry Eyring. The activated complex in chemical reactions. *The Journal of Chemical Physics*, 3(2):107–115, February 1935.

[9] Jeremy N. Harvey, Fahmi Himo, Feliu Maseras, and Lionel Perrin. Scope and challenge of computational methods for studying mechanism and reactivity in homogeneous catalysis. *ACS Catalysis*, 9(8):6803–6813, June 2019.

[10] Jonny Proppe, Tamara Husch, Gregor N. Simm, and Markus Reiher. Uncertainty quantification for quantum chemical models of complex reaction networks. *Faraday Discussions*, 195:497–520, 2016.

[11] Chen Wang, Yao Fu, Qing-Xiang Guo, and Lei Liu. First-principles prediction of nucleophilicity parameters for π nucleophiles: Implications for mechanistic origin of mayr's equation. *Chemistry - A European Journal*, 16(8):2586–2598, February 2010.

[12] Lian-Gang Zhuo, Wei Liao, and Zhi-Xiang Yu. A frontier molecular orbital theory approach to understanding the mayr equation and to quantifying nucleophilicity and electrophilicity by using homo and lumo energies. *Asian Journal of Organic Chemistry*, 1(4):336–345, November 2012.

[13] Harish Jangra, Quan Chen, Elina Fuks, Ivo Zenz, Peter Mayer, Armin R. Ofial, Hendrik Zipse, and Herbert Mayr. Nucleophilicity and electrophilicity parameters for predicting absolute rate constants of highly asynchronous 1, 3-dipolar cycloadditions of aryldiazomethanes. *Journal of the American Chemical Society*, 140(48):16758–16772, November 2018.

[14] Robert J. Mayer, Martin Breugst, Nathalie Hampel, Armin R. Ofial, and Herbert Mayr. Ambident reactivity of phenolate anions revisited: A quantitative approach to phenolate reactivities. *The Journal of Organic Chemistry*, 84(14):8837–8858, June 2019.

[15] Zhen Li, Robert J. Mayer, Armin R. Ofial, and Herbert Mayr. From carbodiimides to carbon dioxide: Quantification of the electrophilic reactivities of heteroallenes. *Journal of the American Chemical Society*, 142(18):8383–8402, April 2020.

[16] Jingjing Zhang, Quan Chen, Robert J. Mayer, Jin-Dong Yang, Armin R. Ofial, Jin-Pei Cheng, and Herbert Mayr. Predicting absolute rate constants for huisgen reactions of unsaturated iminium ions with diazoalkanes. *Angewandte Chemie International Edition*, 59(30):12527–12533, May 2020.

[17] Dominik S. Allgäuer, Harish Jangra, Haruyasu Asahara, Zhen Li, Quan Chen, Hendrik Zipse, Armin R. Ofial, and Herbert Mayr. Quantification and theoretical analysis of the electrophilicities of michael acceptors. *Journal of the American Chemical Society*, 139(38):13318–13329, September 2017.

[18] Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum mechanical continuum solvation models. *Chemical Reviews*, 105(8):2999–3094, July 2005.

[19] Victoria Castor-Villegas, Vincent Tognetti, and Laurent Joubert. On the prediction by density functional theory of entropies in solution within implicit solvation models. *Journal of Molecular Modeling*, 31(1), December 2024.

[20] Guillaume Hoffmann, Vincent Tognetti, and Laurent Joubert. On the influence of dynamical effects on reactivity descriptors. *Chemical Physics Letters*, 724:24–28, June 2019.

[21] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[22] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, November 2005.

[23] Yidi Liu, Qi Yang, Junjie Cheng, Long Zhang, Sanzhong Luo, and Jin-Pei Cheng. Prediction of nucleophilicity and electrophilicity based on a machine-learning approach. *ChemPhysChem*, 24(14), June 2023.

[24] Nicolai Ree, Jan M. Wollschläger, Andreas H. Göller, and Jan H. Jensen. Atom-based machine learning for estimating nucleophilicity and electrophilicity with applications to retrosynthesis and chemical stability. *Chemical Science*, 16(13):5676–5687, 2025.

[25] Nicolai Ree, Andreas H. Göller, and Jan H. Jensen. Automated quantum chemistry for estimating nucleophilicity and electrophilicity with applications to retrosynthesis and covalent inhibitors. *Digital Discovery*, 3(2):347–354, 2024.

[26] Sebastián A. Cuesta, Martín Moreno, Romina A. López, José R. Mora, José Luis Paz, and Edgar A. Márquez. Electropredictor: An application to predict mayr's electrophilicity e through implementation of an ensemble model based on machine learning algorithms. *Journal of Chemical Information and Modeling*, 63(2):507–521, January 2023.

[27] Guillaume Hoffmann, Vincent Tognetti, and Laurent Joubert. Can molecular and atomic descriptors predict the electrophilicity of michael acceptors? *Journal of Molecular Modeling*, 24(10), September 2018.

[28] Aël Cador, Vincent Tognetti, Laurent Joubert, and Paul L. A. Popelier. Aza-michael addition in explicit solvent: A relative energy gradient–interacting quantum atoms study. *ChemPhysChem*, 24(24), November 2023.

[29] Guillaume Hoffmann, Muhammet Balcilar, Vincent Tognetti, Pierre Héroux, Benoît Gaüzère, Sébastien Adam, and Laurent Joubert. Predicting experimental electrophilicities from quantum and topological descriptors: A machine learning approach. *Journal of Computational Chemistry*, 41(24):2124–2136, July 2020.

[30] H. Chermette. Chemical reactivity indexes in density functional theory. *Journal of Computational Chemistry*, 20(1):129–154, January 1999.

[31] P. Geerlings, F. De Proft, and W. Langenaeker. Conceptual density functional theory. *Chemical Reviews*, 103(5):1793–1874, April 2003.

[32] Paul Popelier, F. Aicken, and S. O'Brien. *Atoms in molecules*, volume 1. 01 2000.

[33] R. F. W. Bader. *Atoms in Molecules: A Quantum Theory (International Series of Monographs on Chemistry)*. Oxford University Press, USA, 1994.

[34] Manuel Orlandi, Margarita Escudero-Casao, and Giulia Licini. Nucleophilicity prediction via multivariate linear regression analysis. *The Journal of Organic Chemistry*, 86(4):3555–3564, February 2021.

[35] Matthew S. Sigman, Kaid C. Harper, Elizabeth N. Bess, and Anat Milo. The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Accounts of Chemical Research*, 49(6):1292–1301, May 2016.

[36] Stefan T. A. Berger, Armin R. Ofial, and Herbert Mayr. Inverse solvent effects in carbocation carbanion combination reactions: The unique behavior of trifluoromethylsulfonyl stabilized carbanions. *Journal of the American Chemical Society*, 129(31):9753–9761, July 2007.

[37] Roland Appel, Robert Loos, and Herbert Mayr. Nucleophilicity parameters for phosphoryl-stabilized carbanions and phosphorus ylides: Implications for wittig and related olefination reactions. *Journal of the American Chemical Society*, 131(2):704–714, December 2008.

[38] Roland Lucius and Herbert Mayr. Constant selectivity relationships of addition reactions of carbanions. *Angewandte Chemie International Edition*, 39(11):1995–1997, June 2000.

[39] Johannes Proppe. Uncertainty quantification of reactivity scales. `https://gitlab.com/jproppe/mayruq`, 2025. Last accessed 17 July 2025.

[40] Ricardo A. Mata and Martin A. Suhm. Benchmarking quantum chemical methods: Are we heading in the right direction? *Angewandte Chemie International Edition*, 56(37):11011–11018, April 2017.

[41] Gregor N. Simm, Jonny Proppe, and Markus Reiher. Error assessment of computational models in chemistry. *CHIMIA*, 71(4):202, April 2017.

[42] Christian Reichardt. Pyridinium-n-phenolate betaine dyes as empirical indicators of solvent polarity: Some new findings. *Pure and Applied Chemistry*, 80(7):1415–1432, 2008.

[43] Christian Laurence, Julien Legros, Agisilaos Chantzis, Aurélien Planchat, and Denis Jacquemin. A database of dispersion-induction di, electrostatic es, and hydrogen bonding $\alpha 1$ and $\beta 1$ solvent parameters and some applications to the multiparameter correlation analysis of solvent effects. *The Journal of Physical Chemistry B*, 119(7):3174–3184, January 2015.

[44] Mortimer J. Kamlet, Jose Luis M. Abboud, Michael H. Abraham, and R. W. Taft. Linear solvation energy relationships. 23. a comprehensive collection of the solvatochromic parameters, .pi.*, .alpha., and .beta., and some methods for simplifying the generalized solvatochromic equation. *The Journal of Organic Chemistry*, 48(17):2877–2887, August 1983.

[45] Javier Catalán, Vicenta López, Pilar Pérez, Rosa Martin-Villamil, and José-Gonzalo Rodríguez. Progress towards a generalized solvent polarity scale: The solvatochromism of 2-(dimethylamino)-7-nitrofluorene and its homomorph 2-fluoro-7-nitrofluorene. *Liebigs Annalen*, 1995(2):241–252, February 1995.

[46] Javier Catalán, Cristina Díaz, Vicenta López, Pilar Pérez, José-Luis G. De Paz, and José Gonzalo Rodríguez. A generalized solvent basicity scale: The solvatochromism of 5-nitroindoline and its homomorph 1-methyl-5-nitroindoline. *Liebigs Annalen*, 1996(11):1785–1794, November 1996.

[47] Javier Catalán and Cristina Díaz. A generalized solvent acidity scale: The solvatochromism of o-tert-butylstilbazolium betaine dye and its homomorph o, o'-di-tert-butylstilbazolium betaine dye. *Liebigs Annalen*, 1997(9):1941–1949, September 1997.

[48] Charles M. Hansen. *Hansen Solubility Parameters: A User's Handbook, Second Edition*. CRC Press, June 2007.

[49] Tanja Kanzian, Sami Lakhdar, and Herbert Mayr. Kinetic evidence for the formation of oxazolidinones in the stereogenic step of proline-catalyzed reactions. *Angewandte Chemie International Edition*, 49(49):9526–9529, November 2010.

[50] Sami Lakhdar, Biplab Maji, and Herbert Mayr. Imidazolidinone-derived enamines: Nucleophiles with low reactivity. *Angewandte Chemie International Edition*, 51(23):5739–5742, April 2012.

[51] Hannes Erdmann, Feng An, Peter Mayer, Armin R. Ofial, Sami Lakhdar, and Herbert Mayr. Structures and reactivities of 2-trityl- and 2-(triphenylsilyl)pyrrolidine-derived enamines: Evidence for negative hyperconjugation with the trityl group. *Journal of the American Chemical Society*, 136(40):14263–14269, September 2014.

[52] Daria S. Timofeeva, Robert J. Mayer, Peter Mayer, Armin R. Ofial, and Herbert Mayr. Which factors control the nucleophilic reactivities of enamines? *Chemistry - A European Journal*, 24(22):5901–5910, March 2018.

[53] Guillaume Hoffmann, Frédéric Guégan, Vanessa Labet, Laurent Joubert, Henry Chermette, Christophe Morell, and Vincent Tognetti. Expanding horizons in conceptual density functional theory: Novel ensembles and descriptors to decipher reactivity patterns. *Journal of Computational Chemistry*, 45(20):1716–1726, April 2024.

[54] P. K. Chattaraj and B. Maiti. Reactivity dynamics in atom-field interactions: A quantum fluid density functional study. *The Journal of Physical Chemistry A*, 105(1):169–183, December 2000.

[55] Robert G. Parr and Weitao Yang. Density functional approach to the frontier-electron theory of chemical reactivity. *Journal of the American Chemical Society*, 106(14):4049–4050, July 1984.

[56] Paul Geerlings, Stijn Fias, Zino Boisdenghien, and Frank De Proft. Conceptual dft: chemistry from the linear response function. *Chemical Society Reviews*, 43(14):4989, 2014.

[57] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[58] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

[59] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, April 2011.

[60] Asa Ben-Hur, David Horn, Hava T. Siegelmann, and Vladimir Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, 2001.

[61] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[62] Bernhard Schölkopf and Alexander J. Smola. Learning with kernels. In *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[63] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[64] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[65] G. te Velde, F. M. Bickelhaupt, E. J. Baerends, C. Fonseca Guerra, S. J. A. van Gisbergen, J. G. Snijders, and T. Ziegler. Chemistry with ADF. *J. Comput. Chem.*, 22(9):931–967, 2001.

[66] Evert Jan Baerends, Nestor F. Aguirre, Nick D. Austin, Jochen Autschbach, F. Matthias Bickelhaupt, Rosa Bulo, Chiara Cappelli, Adri C. T. van Duin, Franco Egidi, Célia Fonseca Guerra, Arno Förster, Mirko Franchini, Theodorus P. M. Goumans, Thomas Heine, Matti Hellström, Christoph R. Jacob, Lasse Jensen, Mykhaylo Krykunov, Erik van Lenthe, Artur Michalak, Mariusz M. Mitoraj, Johannes Neugebauer, Valentin Paul Nicu, Pier Philipsen, Harry Ramanantoanina, Robert Rüger,

Georg Schreckenbach, Mauro Stener, Marcel Swart, Jos M. Thijssen, Tomáš Trnka, Lucas Visscher, Alexei Yakovlev, and Stan van Gisbergen. The amsterdam modeling suite. *The Journal of Chemical Physics*, 162(16):162501, 04 2025.

[67] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, October 1996.

[68] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, March 2011.

[69] Theoretical Chemistry SCM. Plams interfaces: Conformers. `https://www.scm.com/doc/plams/interfaces/conformers.html`, 2023. Software for Chemistry & Materials, Amsterdam, The Netherlands.

[70] Theoretical Chemistry SCM. Plams components: RDKit. `https://www.scm.com/doc/plams/components/mol_rdkit.html`, 2023. Software for Chemistry & Materials, Amsterdam, The Netherlands.

[71] Andreas Klamt. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena. *The Journal of Physical Chemistry*, 99(7):2224–2235, February 1995.

[72] Cory C. Pye and Tom Ziegler. An implementation of the conductor-like screening model of solvation within the amsterdam density functional package. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)*, 101(6):396–408, May 1999.

[73] Ralph G. Pearson. *Chemical Hardness.* Wiley, October 1997.

[74] W Yang and R G Parr. Hardness, softness, and the fukui function in the electronic theory of metals and catalysis. *Proceedings of the National Academy of Sciences*, 82(20):6723–6726, October 1985.

[75] Robert G. Parr, László v. Szentpály, and Shubin Liu. Electrophilicity index. *Journal of the American Chemical Society*, 121(9):1922–1924, February 1999.

[76] Paul W. Ayers and Mel Levy. *Perspective on "Density functional approach to the frontier-electron theory of chemical reactivity"*, page 353–360. Springer Berlin Heidelberg, 2000.

[77] Ravishankar Sundararaman, William A. Goddard, and Tomas A. Arias. Grand canonical electronic density-functional theory: Algorithms and applications to electrochemistry. *The Journal of Chemical Physics*, 146(11), March 2017.

[78] Daniel de las Heras and Matthias Schmidt. Full canonical information from grand-potential density-functional theory. *Physical Review Letters*, 113(23), December 2014.

[79] R. Grincourt, G. Hoffmann, F. Guégan, V. Tognetti, L. Joubert, H. Chermette, A. Toro Labbé, and C. Morell. Title of the article. *J. Chem. Phys.*, 2025. in press.