JOC The Journal of Organic Chemistry

pubs.acs.org/joc Note

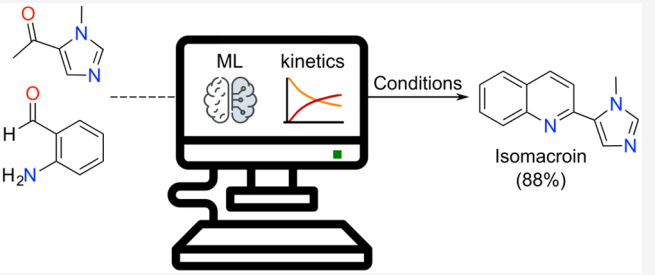
Augmenting Adaptive Machine Learning with Kinetic Modeling for Reaction Optimization

A. Filipa Almeida, Filipe A. P. Ataíde,* Rui M. S. Loureiro, Rui Moreira, and Tiago Rodrigues*

Cite This: https://doi.org/10.1021/acs.joc.1c01038 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We combine random sampling and active machine learning (ML) to optimize the synthesis of isomacroin, executing only 3% of all possible Friedländer reactions. Employing kinetic modeling, we augment machine intuition by extracting mechanistic knowledge and verify that a global optimum was obtained with ML. Our study contributes evidence on the potential of multiscale approaches to expedite the access to chemical matter, further democratizing organic chemistry in a data-motivated fashion.



Synthetic organic chemistry toward high value bioactive entities and materials is key in modern molecular medicine, but often delivers suboptimal processes and insufficient amounts of chemical matter for advanced functional profiling.[1−3] Indeed, the lack of screening materials for biological investigations may curb or delay the identification of potentially life-changing therapies. Natural products (NPs) have long been exploited as therapeutics or as a source of inspiration for molecular design due to the biological prevalidation of their frameworks as protein-binding motifs.[4−6] In fact, ca. 33% of approved small-molecule drugs are either NPs or NP-derived compounds, which highlights their value in translational science.[7]

Fragment-like NPs usually provide simpler and synthetically more accessible architectures that can also be readily adopted for myriad discovery chemistry applications.[8,9] For example, we had unveiled that isomacroin (**1**, Figure 1)—a fragment-like NP from *Macrorungia longistrobus*[10]—presented the blueprints for efficient platelet-derived growth factor receptor alpha kinase (PDGFRα) modulation while using self-organizing maps for target deorphanization.[11,12] The tractability of **1** as a prototype for medicinal chemistry was demonstrated by the development of potent PDGFR[13,14] and IκB kinase[15,16] inhibitors based on the imidazolyl quinoline scaffold—a hinge binding motif (Figure 1). However, synthetic access to **1** through Friedländer quinoline synthesis proved challenging in our hands. Poor (15%) yield[17] was obtained, which limited screening efforts on a wider scale.

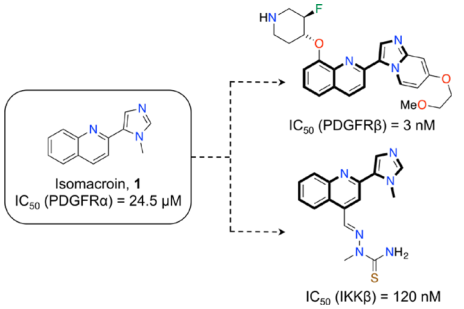


Figure 1. Structure of isomacroin, **1**, and related molecules as human kinase inhibitors. The framework of **1** is highlighted in the NP-derived, bioactive small molecules.

With that problem in mind, we have recently developed LabMate.ML—a self-evolving machine-learning (ML) routine for digitalizing reaction optimizations under low data regimes—with the goal of providing an interpretable, generalizable, and nonexhaustive search space exploration alternative to full/fractional factorial "design-of-experiments".[18] The method leverages information in a small set of random reactions for initialization, and adaptive random forests powered by a bespoke experiment selection policy to iteratively drive the optimization process. In a previous report, and in alignment with related methods,[19−21] we had shown that LabMate.ML not only efficiently modeled real-valued and categorical variables but, more importantly, was competitive with expert intuition in competition studies. Taking advantage of its decision process, interpretable outputs could also be extracted to augment domain intuition.[18]

Notwithstanding the prior success of LabMate.ML, daily practice in process chemistry optimization more frequently

Special Issue: Enabling Techniques for Organic Synthesis

Received: May 3, 2021

exploits kinetic modeling to motivate reaction conditions lead to productive syntheses.[22,23] Ultimately, this approach allows a data-informed elucidation of reaction mechanisms and tailored protocol design.[22,24] Herein, we build on our prior ML workflow and provide a preliminary proof of concept for cascading active learning and kinetic modeling as an approach toward fast reaction optimization, including the identification of a global optimum. Using **1** as the model compound, our ML tool rapidly exploited a discretized Friedländer reaction space. The generated insights were subsequently augmented with kinetic modeling to ascertain a reactivity pathway and corresponding rate/equilibrium constants, which to the best of our knowledge have not been determined.[25] We also confirmed that the ML routine had rapidly converged to a global optimum (88%) with minimal experimental effort. Overall, our multiscale, human-in-loop approach may prove transferable to other chemistries and enable the swift access to high value chemical matter by democratizing organic synthesis.

To delve into the reactivity space and identify an optimal synthesis protocol toward **1**, we initially surveyed the literature and collected preferable conditions that could be experimentally probed[26−30] (Figure 2a). These included different

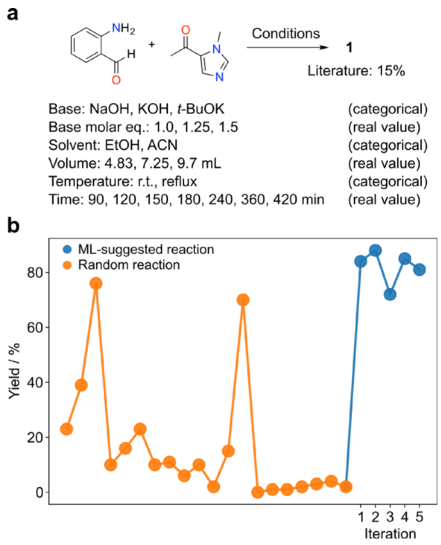


Figure 2. Optimization of reaction conditions toward isomacroin, **1**. (a) Six reaction variables were probed in the optimization of a Friedländer quinoline synthesis. Both real-valued and categorical features were considered, affording a search space of 756 possible reactions. Categorical variables were one-hot encoded for modeling. Real value variables were employed for modeling without any transformation. (b) Optimization routine, following a user-defined number of random reactions for initialization (orange) and an additional five ML-suggested reactions until reasonable convergence was obtained (blue).

solvents and their volume, bases and molar amount, temperature, and reaction time. The reaction variables were encoded as real values (e.g., volume) or strings as in the case of categorical (e.g., base) variables.

For machine interpretation and modeling, categorical variables were then one-hot encoded, denoting either the presence ("1") or absence ("0") of a specified entity. Collectively, 756 reactions were digitalized in high dimensional space, which in this case constitutes more combinations than are reasonably feasible to screen. This search space was subjected to iterative investigation by our adaptive random forest routine—a supervised learning algorithm that harnesses the "wisdom of the crowds" concept by aggregating individual predictions from uncorrelated decision trees. Because each tree analyzes a fraction of data, only weak predictions are individually expected as output. However, averaging those over *n* trees allows for a significantly more realistic prediction and uncertainty estimation. Together, these characteristics endorse the popularity of random forests in the chemical sciences[17,31,32] and make them robust to common pitfalls in ML research, such as outliers, noise, and overfitting.[33,34]

To initialize the optimization campaign, 20 random reactions were performed. Their outcomes (target variable/yield) were collected via a calibration curve that was built from the area under the curve for the required product peak at different concentration values (HPLC−UV/vis−MS traces). Interestingly, one such random reaction corresponded to literature conditions,[11] but afforded a higher yield in this reassessment (76% here vs 15% literature). The obtained yield was still used as benchmark, yet highlights the previously discussed variability in chemical data that can determine the predictive power of statistical learning.[35] All random reaction conditions and their yields (0−76%, Figure 2b) were employed as an initial training set, and a stratified 10-fold cross-validation routine, wherein analyses are repeated with 90% of data used for training and 10% as internal test set, was implemented for a preliminary assessment of the model utility. By harnessing a previously validated[18] "greedy" approach for experiment prioritization, i.e., selecting reactions with high-predicted yield and low uncertainty, our recommender tool rapidly (over five iterations) converged to an optimum (Figure 2b). Relative to the literature conditions, the ML routine elected to substitute the base (KOH to *t*-BuOK) and its molar equivalent amount (from 1.25 to 1.50 molar equiv) to afford **1** in 88% yield (cf. Table S1). Our understanding is that such modifications are reasonable and among the top variables a skilled chemist would also probe, which supports a correct formalization of chemical intuition by the computational tool.

Additionally, shuffling of target (Y) variables resulted in less predictive control models. This evidenced that meaningful, true patterns in the data structure had been disrupted in the Y-randomization process (cf. Table S3). We deemed this control necessary due to the high likelihood of ML heuristics exploiting artifacts and memorizing rather than learning data.[36−39] This is especially critical when dealing with low data in high dimensional space, as in the present case. Finally, we studied if different baseline algorithms, e.g., linear, lasso, and ridge regression, could provide similar or better statistical models relative to our random forests while enforcing simplicity in the decision process. Our method proved more accurate and efficient in navigating the training data, as assessed through RMSE and MAE values (cf. Table S2). Overall, the results attest to the appropriateness and robustness of adaptive random forests for assisting chemists in reaction optimization problems.

It is now established that different ML algorithms can afford expert level predictions in numerous tasks relevant to the chemical sciences.[40−42] However, it is still usual to harness ML as black boxes, wherein no insight into the decision process is provided. This can not only hinder the adoption of potentially useful technologies but also bypass on new knowledge that might otherwise remain hidden to perception.[43] To shed light onto the key variables for our statistical model and infer on their directionality, i.e., desired or undesired, we extracted

https://doi.org/10.1021/acs.joc.1c01038
J. Org. Chem. XXXX, XXX, XXX−XXX

importance values from the regressor (Figure 3a) and employed SHapley Additive ExPlanations (SHAP; Figure 3b)

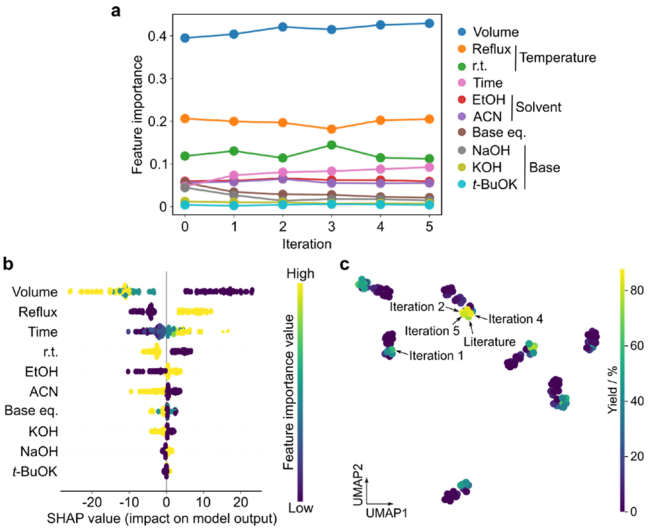


Figure 3. ML interpretation on a global scale. (a) The importance of each feature in the decision process of the adaptive random forest model is provided over five iterations, until reasonable target variable (yield/output) convergence was obtained. Data shows dynamic feature importance fluctuations as more data becomes available in the active learning process. Categorical variables are decoupled in the plot. (b) SHAP[44] was employed to infer on the directionality of each feature for the model/reaction output/yield. Low reaction volumes, higher temperature, and longer reaction times are preferred, together with using EtOH as solvent and NaOH or *t*-BuOK as base. (c) Visualization of the search space with UMAP (Uniform Manifold Approximation and Projection),[45] which preserves the local neighborhood structure in the data set. Data shows specific islets of reactivity and preferred regions leading to the formation of isomacroin.

as an orthogonal, model-agnostic approach. SHAP fits linear models to provide both global (model) and local (individual prediction) data interpretations.[39]

In this Friedländer quinoline synthesis, small reaction volumes were crucial, as was refluxing for a long period of time (Figure 3a,b). Indeed, without encoding any explicit chemical knowledge, our ML routine was able to autonomously formalize unwritten rules of intuition, which corroborates its utility. These interpretations support that reaction optimizations can effectively be configured as data mining problems. Furthermore, with manifold learning (Figure 3c) we confirmed that all reactions cocluster in specific islets of reactivity that were exploited by our algorithm, and that each cluster can afford **1** in moderate yields. This result provides a readily visualizable interpretation of the pursued selection policy. Moreover, it is possible to observe that the iterative search chiefly revolved around the best performing random reactions in the initialization step, without undermining the identification of an optimum (e.g., iteration 2).

With this result in hand, we next wondered if the initial search space discretization step had limited the ability to identify a global optimum and, therefore, if ML served the purpose of a swift yet coarse-grained interrogation of the Friedländer reactivity space with additional room for improvement. To that end, we set up a series of experiments to model the reaction kinetics, identify a global optimum, and establish a mechanism. Specifically, we explored different molar equivalent values for the 2-aminobenzaldehyde starting material (0.94, 1.0, 1.45, 2.0), *t*-BuOK (1.25, 1.5, 2.2, 2.9), and temperature values (40, 53, and 78 °C; cf. Table S4). In these experiments, reaction mixture aliquots were collected at specified time-points and all relevant species were quantified through HPLC−UV/vis−MS (e.g., Figure 4a). This allowed us

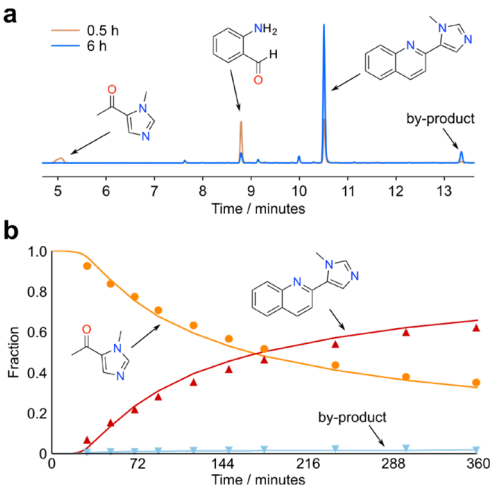


Figure 4. Kinetic modeling as an orthogonal approach to augment machine-learning outputs. (a) Exemplary, overlaid HPLC−UV traces (0.5 and 6 h) showing the formation of isomacroin over time. (b) Exemplary plot showing the evolution of species concentration over time and fitted curve. Kinetic parameters were extracted from a series of similar experiments.

monitoring the concentration increase/decrease of each starting material, product, and byproduct over 360 min of reaction (e.g., Figure 4b). While reaction intermediates may provide an additional layer of information and confidence in downstream data inference, such species were not quantifiable through our analytical method. We thus assumed their transient nature and did not consider them for modeling.

A kinetic model for the Friedländer quinoline synthesis was then developed based on differential equations that describe the reaction rate as a function of concentration change over time for all above-mentioned species. This valuable information was modeled in DynoChem (Scale-up Systems[46]) by minimizing the sum of squares between the experimental data and model predictions using the gradient-based Levenberg−Marquardt algorithm.

The fitted model led us to establish the reaction mechanism toward **1**, in line with the literature[25] (Figure 5). To fully profile the transformation, we extracted the first- and second-order rates (k), equilibrium (K_{eq}) constants, and activation energy (E_a) values at a 95% confidence interval level (cf. Table

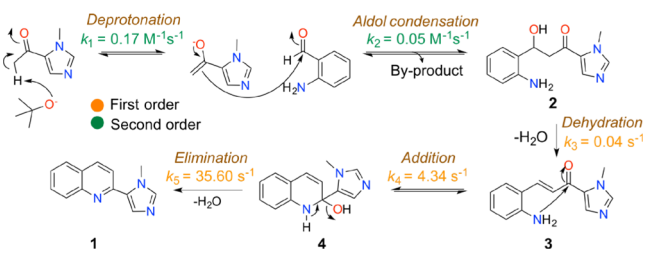


Figure 5. Friedländer chemistry mechanism according to kinetic modeling with DynoChem.

https://doi.org/10.1021/acs.joc.1c01038
J. Org. Chem. XXXX, XXX, XXX−XXX

S5). The reaction rate predictions suggest that formation of **1** occurs rapidly, with the dehydration step being rate-limiting. The results also indicate that consumption of 2-aminobenzaldehyde occurs through an aldol condensation toward intermediate **2**. Moreover, **1** is amenable to slow degradation relative to the elimination step when excess 2-aminobenzaldehyde and *t*-BuOK are employed. Domain knowledge informs that formation of a Schiff base is not viable under basic conditions. As an adversarial control to the proposed mechanism, we fitted a model assuming the formation of an imine as first step. The results are in line with established intuition and assert the formation of **2** (k_{imine} = 0.01 M^{-1} s^{-1} vs 0.05 M^{-1} s^{-1}), thus providing an additional layer of confidence on our kinetic model.

On a broader perspective, we confirmed the ML predictions through an orthogonal means, as both methods in our investigation converged to identical reaction protocols and predicted yields. More specifically, kinetic modeling suggests that a global optimum, with minimization of the byproduct, is achieved at 78 °C and utilizing 2 and 1.05 molar equiv of 2-aminobenzaldehyde and *t*-BuOK, respectively, over 593 min of reaction (Figure S15). There is, however, value in coalescing both approaches into a streamlined workflow. The kinetic model augmented the ML-derived intuition by elucidating the reaction mechanism and rate constants. It also informed the impact of the byproduct for the development of an optimized synthesis protocol, albeit requiring time-consuming computation (ca. 5−10 min for ML vs 220−250 min for kinetic model). Together, this substantiates the power of pattern identification through ML. It also provides a motivation for employing computationally expensive methods when added value is warranted.

In conclusion, we implemented a cascaded workflow comprising adaptive learning and kinetic modeling to facilitate the access to **1** via Friedländer chemistry and afford physical chemistry insights. The whole optimization process—including featurization, random selection, hyperparameter tuning, and ML selection—allowed the prioritization of experiments and identification of an optimum while executing only a minute amount (3%) of all possible reactions. This is relevant because ML can equally work as an optimizer or fine-tuner of experiments, using poor or good yields as starting points, respectively. Further, the process also allowed establishing a mechanistic path to the transformation. Our study provides proof of concept for a viable integration of well-established concepts in process chemistry with emerging technologies. Ultimately, it may impact on molecular medicine pipelines by expediting the access to high value chemical matter for screening purposes. We foresee this and similar integrations working as research assistants, promoting probabilistically informed decisions and democratizing organic syntheses in the digital chemistry era.

EXPERIMENTAL SECTION

General Methods. Starting materials and reagents were purchased from Sigma-Aldrich, Alfa Aesar, Fluka, or Acros and used without further purification. Reactions were carried out on a Radleys Carousel 6 Plus Reaction Station. ^{1}H NMR spectra were obtained at on a Bruker Avance 300 MHz in CD_3OD and $(CD_3)_2CO$ with chemical shift values (δ) in parts per million using residual solvent peaks as the internal standard, and ^{13}C NMR spectra were obtained at 75 MHz in the same deuterated solvents. Coupling constants (J) are reported in hertz with the following splitting abbreviations: s = singlet, d = doublet, t = triplet, dd = doublet of doublets, ddd = doublet of doublet of doublets, ddt = doublet of doublet of triplets, and m = multiplet. A high-performance liquid chromatography (HPLC) (Waters Corp., Milford, MA) system was used to quantify the isomacroin, as product, and the samples were collected from the chemical reactions. The chromatographic analysis was performed in Agilent Eclipse XDB-C18 column (150 mm × 4.6 mm i.d. 3.5 μm) at room temperature. The product was detected by ultraviolet (UV) absorbance detection at 254 nm. The analyses were performed under an appropriate gradient of ammonium acetate/acetonitrile (90:10) at a pH of 8.5 and acetonitrile/ammonium acetate (90:10) and a total flow rate of 1 mL/min. The injection volume was 5 μL, and the total run time was 20.1 min. The ultraperformance liquid chromatography (UPLC) analyses were performed on an ACQUITY UPLC system equipped with a photodiode array (PDA) detector (Waters Corp., Milford, MA) coupled to a mass single-quadrupole detector (QDa Waters). This system was used to identify the starting raw materials, product, and byproducts, and the samples were collected from the chemical reactions. All compounds were monitored at 254 nm by PDA detector. The single ion recording (SIR) method was established on a single quadruple mass detector. The QDa conditions were set as follows: a cone voltage of 15 V, a capillary voltage of 0.8 kV, and a source temperature of 600 °C. The data were acquired under the SIR mode. The column, injection volume, flow rate, and solvent managers were already described previously.

Synthesis of 1-(1-Methyl-1H-imidazol-5-yl)ethan-1-one. To a flask containing 1-methyl-1*H*-imidazole (0.485 mL, 0.50 g, 6.0 mmol) and tetrahydrofuran (2.87 mL) at −78 °C was added 1.6 M *n*-BuLi in hexane (4.0 mL, 6.5 mmol) followed by stirring at −78 °C for 40 min. Then chlorotrimethylsilane (0.8 mL, 6.3 mmol) was added slowly, and the mixture was stirred at −78 °C for 1 h. The 1.6 M *n*-BuLi in hexane (4.0 mL, 6.5 mmol) was added, the cooling bath was removed, and stirring was continued until the temperature reached 10 °C. The mixture was recooled to −78 °C, and a solution of *N,N*-dimethylacetamide (0.463 mL, 5.0 mmol) was added. The cooling bath was removed, and the stirring was continued for 40 min at room temperature. The reaction was quenched with a few drops of methanol (1.0 mL), and brine was added. The organic layer was separated, and the aqueous layer was extracted with dichloromethane. The combined organic phases were dried (sodium sulfate anhydrous), filtered, and concentrated under reduced pressure. The crude was filtered through a short silica plug and eluting with ethyl acetate (75 mL). The solvent was evaporated to afford the required compound. Colorless oil, 93% (0.703 g), R_f 0.68 (1/1 v/v ethyl acetate/heptane). ^{1}H NMR (300 MHz, $(CD_3)_2CO$): δ 7.28−7.24 (m, 1H), 7.00 (d, J = 1.0 Hz, 1H), 3.92 (s, 3H), 2.48 (s, 3H). ^{13}C{^{1}H} NMR (75 MHz, $(CD_3)_2CO$): δ 189.7, 128.6, 127.4, 35.2, 26.1. As described in the literature.[47]

Synthesis of 2-Aminobenzaldehyde. A solution of 2-nitrobenzaldehyde (0.50 g, 3.3 mmol) in ethanol (9.4 mL) was stirred for approximately 1 min. Iron powder (0.554 g, 9.9 mmol) and dilute hydrochloric acid (3.3 mL of 1.0 M HCl, 0.5 mmol) were added to the stirred solution, and the reaction was heated to reflux for 2 h in an Asynt DrySyn Single Position Blocks system. The reaction mixture was cooled to room temperature, diluted with ethyl acetate (27.0 mL), and stirred for 5 min before being filtered through a short Celite plug. The filtrate was evaporated under reduced pressure to yield a yellow oil. The product was stored at −20 °C. Yellow oil, 100% (0.40 g), R_f 0.44 (1/5 v/v ethyl acetate/hexane). ^{1}H NMR (300 MHz, $(CD_3)_2CO$) δ 9.88 (d, J = 0.6 Hz, 1H), 7.55 (dd, J = 7.8, 1.6 Hz, 1H), 7.31 (ddd, J = 8.5, 7.0, 1.6 Hz, 1H), 6.81 (dq, J = 8.3, 0.7 Hz, 2H), 6.74−6.66 (m, 1H). ^{13}C{^{1}H} NMR (75 MHz, $(CD_3)_2CO$) δ 193.7, 1508, 135.6, 134.9, 118.6, 115.9, 115.4. As described in literature.[48]

Synthesis of Isomacroin (1). The ketone (0.10 g, 0.80 mmol, 1 molar equiv) and potassium hydroxide (0.057 g, 1.0 mmol, 1.25 molar equiv) were dissolved in ethanol (6 mL/mmol). Then the 2-aminobenzaldehyde (0.098g, 0.8 mmol, 1 molar equiv) was added to the solution and the reaction mixture refluxed 3 h in a Asynt DrySyn Single Position Blocks system. The solvent was evaporated, and the crude was purified by flash column chromatography with heptane/ethyl acetate (3:1) eluent. Yellow solid, 80% (0.135 g), R_f 0.40 (1/1

https://doi.org/10.1021/acs.joc.1c01038
J. Org. Chem. XXXX, XXX, XXX−XXX

v/v ethyl acetate/hexane). ^{1}H NMR (300 MHz, CD_3OD): δ 8.34 (d, J = 8.7 Hz, 1H), 8.18 (dd, J = 8.6, 0.8 Hz, 1H), 8.05 (ddt, J = 8.5, 1.4, 0.7 Hz, 1H), 7.82−7.77 (m, 1H), 7.69 (ddd, J = 8.5, 6.9, 1.5 Hz, 1H), 7.51 (ddd, J = 8.1, 6.9, 1.2 Hz, 1H), 7.18 (d, J = 1.1 Hz, 1H), 7.02 (d, J = 1.1 Hz, 1H), 4.30 (s, 3H). $^{13}C\{^1H\}$ NMR (75 MHz, CD_3OD) δ 150.6, 147.3, 145.0, 136.3, 129.5, 129.4, 128.6, 127.6, 127.2, 126.5, 125.2, 120.6, 36.6.

Machine Learning. We used random forest regressors and probed several hyperparameters [number of trees (100−1000), tree depth (none, 2, 4) and number of features (auto or sqrt)] to create a prediction model. A 10-fold cross-validation was applied, splitting the data set into a training group with 90% of data and test group with the remaining 10%. The selection of best hyperparameters was guided by the calculation of the mean absolute error (MAE). With these parameters, the model predicted the product yield for the remaining possible reaction conditions, and then the next reaction was selected. An exploration approach was applied for the first three iterations to allow model improvement. For the next iterations, an exploitative approach was applied by selecting the reaction with the lowest variance among the predicted top-5 high-yielding reactions. The model develops with each added data point by improving its predictive model. The model and data analyses were fully implemented in Python 3.7.3 using the NumPy 1.16.4, Pandas 0.24.2 and Scikit-learn 0.21.2 libraries and was run (5−10 min) on an HP ProBook G3 (2.40 GHz 2 core processor, 4 Gb RAM). For initialization, 20 random reactions were performed in parallel (Table S1) and analyzed through HPLC to create the first training data set. Using a calibration curve, it was possible to quantify the product yield obtained in each reaction. Based on this data, the machine learning routine selected one additional set of conditions, at a time, for experimental validation. The output for the selected experiment was added to training set and an iterative process involving design−make−test was pursued over five iterations. Code for the optimization routine is accessible at https://github.com/tcorodrigues/LabMate.ML.

Machine Learning Controls. To evaluate the robustness of the method, other models, such as linear regression, ridge, and lasso, were computed. In parallel, we performed y-randomization studies to rule out artifacts in the decision process of our random forests. Error metrics, such as the root mean square error (RMSE), mean absolute errors (MAE), and coefficient of determination (r^2) were calculated and used for comparison between the different methods.

Kinetic Method. The DynoChem software (Scale-Up Systems; Build: 1.3.20b340, Data Version: 1.0) was used for reaction modeling. The software includes model templates, provides simulator, fitting, and optimization add-in options to run the models. In this work, we used a single model template to fit all experiments and calculate kinetic parameters. We employed the Levenberg−Marquardt algorithm for fitting, during minimization of SSQ, and the Backward Euler (BE) integration method to solve the mass balance differential and algebraic equations associated with the reaction. The kinetic parameters were calculated at a defined reference temperature (78 °C). All reactions were run for 360 min and aliquots collected at specified time points. Analyses were performed in HPLC and based on calibration curves.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.joc.1c01038.

^{1}H and ^{13}C NMR spectra for all compounds, UPLC-QDa chromatograms for all compounds, data and conditions used in machine learning model and kinetic model (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Filipe A. P. Ataíde − *R&D, Process Chemistry Development, Hovione FarmaCiência S.A, 1649-038 Lisboa, Portugal;* orcid.org/0000-0002-8288-2152; Email: fataide@hovione.com

Tiago Rodrigues − *Research Institute for Medicines (iMed.Ulisboa), Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal;* orcid.org/0000-0002-1581-5654; Email: tiago.rodrigues@ff.ulisboa.pt

Authors

A. Filipa Almeida − *R&D, Process Chemistry Development, Hovione FarmaCiência S.A, 1649-038 Lisboa, Portugal; Research Institute for Medicines (iMed.Ulisboa), Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal*

Rui M. S. Loureiro − *R&D, Process Chemistry Development, Hovione FarmaCiência S.A, 1649-038 Lisboa, Portugal*

Rui Moreira − *Research Institute for Medicines (iMed.Ulisboa), Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.joc.1c01038

Author Contributions

A.F.A. performed chemistry and modeling. All authors designed experiments and analyzed data. A.F.A and T.R. wrote the manuscript with contributions from the remaining authors. F.A.P.A., R.M.S.L., R.M., and T.R. supervised the research. F.A.P.A. and T.R. coordinated the study. All authors have given approval to the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): T.R. is co-founder and shareholder of TargTex S.A.

■ ACKNOWLEDGMENTS

A.F.A. and T.R. acknowledge Fundação para a Ciência e Tecnologia (FCT) Portugal for financial support through PD/BD/143125/2019 and CEECIND/00684/2018, respectively. This research received funding from European Structural & Investment Funds through the COMPETE Program—Programa Operacional Regional de Lisboa—Program Grant LISBOA-01-0145- FEDER-016405, and from National Funds through the FCT—Fundação para a Ciência e a Tecnologia—Program Grant SAICTPAC/0019/2015. FCT is also acknowledged for support of the MedChemTrain PhD programme (PD147-2013-AA). The authors acknowledge Ricardo Gonçalves (Hovione) for great support in designing the analytical method to quantify all species.

■ REFERENCES

(1) Hayashi, Y. Time Economy in Total Synthesis. *J. Org. Chem.* **2021**, *86* (1), 1−23.

(2) Peiretti, F.; Brunel, J. M. Artificial Intelligence: The Future for Organic Chemistry? *ACS Omega* **2018**, *3* (10), 13263−13266.

(3) Pflüger, P. M.; Glorius, F. Molecular Machine Learning: The Future of Synthetic Chemistry? *Angew. Chem., Int. Ed.* **2020**, *59* (43), 18860−18865.

(4) Harvey, A. L. Natural Products in Drug Discovery. *Drug Discovery Today* **2008**, *13* (19−20), 894−901.

(5) Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Orhan, I. E.; Banach, M.; Rollinger, J. M.; Barreca, D.; Weckwerth, W.; Bauer, R.;

Bayer, E. A.; Majeed, M.; Bishayee, A.; Bochkov, V.; Bonn, G. K.; Braidy, N.; Bucar, F.; Cifuentes, A.; D'Onofrio, G.; Bodkin, M.; Diederich, M.; Dinkova-Kostova, A. T.; Efferth, T.; El Bairi, K.; Arkells, N.; Fan, T.-P.; Fiebich, B. L.; Freissmuth, M.; Georgiev, M. I.; Gibbons, S.; Godfrey, K. M.; Gruber, C. W.; Heer, J.; Huber, L. A.; Ibanez, E.; Kijjoa, A.; Kiss, A. K.; Lu, A.; Macias, F. A.; Miller, M. J. S.; Mocan, A.; Müller, R.; Nicoletti, F.; Perry, G.; Pittalà, V.; Rastrelli, L.; Ristow, M.; Russo, G. L.; Silva, A. S.; Schuster, D.; Sheridan, H.; Skalicka-Woźniak, K.; Skaltsounis, L.; Sobarzo-Sánchez, E.; Bredt, D. S.; Stuppner, H.; Sureda, A.; Tzvetkov, N. T.; Vacca, R. A.; Aggarwal, B. B.; Battino, M.; Giampieri, F.; Wink, M.; Wolfender, J.-L.; Xiao, J.; Yeung, A. W. K.; Lizard, G.; Popp, M. A.; Heinrich, M.; Berindan-Neagoe, I.; Stadler, M.; Daglia, M.; Verpoorte, R.; Supuran, C. T. Taskforce, the I. N. P. S. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discovery* **2021**, *20*, 200−216.

(6) Karageorgis, G.; Foley, D. J.; Laraia, L.; Waldmann, H. Principle and Design of Pseudo-Natural Products. *Nat. Chem.* **2020**, *12* (3), 227−235.

(7) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770−803.

(8) Over, B.; Wetzel, S.; Grütter, C.; Nakai, Y.; Renner, S.; Rauh, D.; Waldmann, H. Natural-Product-Derived Fragments for Fragment-Based Ligand Discovery. *Nat. Chem.* **2013**, *5* (1), 21−28.

(9) Rodrigues, T.; Reker, D.; Schneider, P.; Schneider, G. Counting on Natural Products for Drug Design. *Nat. Chem.* **2016**, *8* (6), 531−541.

(10) Arndt, R. R.; Jordaan, A.; Joynt, V. P. Alkaloids of Macrorungia Longistrobus C.B. Cl. *J. Chem. Soc.* **1964**, 5969−5975.

(11) Rodrigues, T.; Reker, D.; Kunze, J.; Schneider, P.; Schneider, G. Revealing the Macromolecular Targets of Fragment-Like Natural Products. *Angew. Chem., Int. Ed.* **2015**, *54*, 10516−10520.

(12) Reker, D.; Rodrigues, T.; Schneider, P.; Schneider, G. Identifying the Macromolecular Targets of de Novo-Designed Chemical Entities through Self-Organizing Map Consensus. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111* (11), 4067−4072.

(13) Hicken, E. J.; Marmsater, F. P.; Munson, M. C.; Schlachter, S. T.; Robinson, J. E.; Allen, S.; Burgess, L. E.; DeLisle, R. K.; Rizzi, J. P.; Topalov, G. T.; Zhao, Q.; Hicks, J. M.; Kallan, N. C.; Tarlton, E.; Allen, A.; Callejo, M.; Cox, A.; Rana, S.; Klopfenstein, N.; Woessner, R.; Lyssikatos, J. P. Discovery of a Novel Class of Imidazo[1,2- a]Pyridines with Potent PDGFR Activity and Oral Bioavailability. *ACS Med. Chem. Lett.* **2014**, *5* (1), 78−83.

(14) Effendi, N.; Mishiro, K.; Takarada, T.; Yamada, D.; Nishii, R.; Shiba, K.; Kinuya, S.; Odani, A.; Ogawa, K. Design, Synthesis, and Biological Evaluation of Radioiodinated Benzo[d]Imidazole-Quinoline Derivatives for Platelet-Derived Growth Factor Receptor β (PDGFRβ) Imaging. *Bioorg. Med. Chem.* **2019**, *27* (2), 383−393.

(15) Cushing, T. D.; Baichwal, V.; Berry, K.; Billedeau, R.; Bordunov, V.; Broka, C.; Cardozo, M.; Cheng, P.; Clark, D.; Dalrymple, S.; DeGraffenreid, M.; Gill, A.; Hao, X.; Hawley, R. C.; He, X.; Jaen, J. C.; Labadie, S. S.; Labelle, M.; Lehel, C.; Lu, P.-P.; McIntosh, J.; Miao, S.; Parast, C.; Shin, Y.; Sjogren, E. B.; Smith, M.-L.; Talamas, F. X.; Tonn, G.; Walker, K. M.; Walker, N. P. C.; Wesche, H.; Whitehead, C.; Wright, M.; Browner, M. F. A Novel Series of IKKβ Inhibitors Part I: Initial SAR Studies of a HTS Hit. *Bioorg. Med. Chem. Lett.* **2011**, *21* (1), 417−422.

(16) Cushing, T. D.; Baichwal, V.; Berry, K.; Billedeau, R.; Bordunov, V.; Broka, C.; Browner, M. F.; Cardozo, M.; Cheng, P.; Clark, D.; Dalrymple, S.; DeGraffenreid, M.; Gill, A.; Hao, X.; Hawley, R. C.; He, X.; Labadie, S. S.; Labelle, M.; Lehel, C.; Lu, P.-P.; McIntosh, J.; Miao, S.; Parast, C.; Shin, Y.; Sjogren, E. B.; Smith, M.-L.; Talamas, F. X.; Tonn, G.; Walker, K. M.; Walker, N. P. C.; Wesche, H.; Whitehead, C.; Wright, M.; Jaen, J. C. A Novel Series of IKKβ Inhibitors Part II: Description of a Potent and Pharmacologically Active Series of Analogs. *Bioorg. Med. Chem. Lett.* **2011**, *21* (1), 423−426.

(17) Rodrigues, T.; Reker, D.; Welin, M.; Caldera, M.; Brunner, C.; Gabernet, G.; Schneider, P.; Walse, B.; Schneider, G. De Novo Fragment Design for Drug Discovery and Chemical Biology. *Angew. Chem., Int. Ed.* **2015**, *54* (50), 15079−15083.

(18) Reker, D.; Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. Adaptive Optimization of Chemical Reactions with Minimal Experimental Information. *Cell Reports Phys. Sci.* **2020**, *1* (11), 100247.

(19) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4* (9), 1134−1145.

(20) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154−6168.

(21) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89−96.

(22) Blackmond, D. G. Reaction Progress Kinetic Analysis: A Powerful Methodology for Mechanistic Studies of Complex Catalytic Reactions. *Angew. Chem., Int. Ed.* **2005**, *44* (28), 4302−4320.

(23) Wang, K.; Han, L.; Mustakis, J.; Li, B.; Magano, J.; Damon, D. B.; Dion, A.; Maloney, M. T.; Post, R.; Li, R. Kinetic and Data-Driven Reaction Analysis for Pharmaceutical Process Development. *Ind. Eng. Chem. Res.* **2020**, *59* (6), 2409−2421.

(24) Nielsen, C. D.-T.; Burés, J. Visual Kinetic Analysis. *Chem. Sci.* **2019**, *10* (2), 348−353.

(25) Marco-Contelles, J.; Pérez-Mayoral, E.; Samadi, A.; Carreiras, M.; do, C.; Soriano, E. Recent Advances in the Friedländer Reaction. *Chem. Rev.* **2009**, *109* (6), 2652−2671.

(26) Lan, X.-B.; Ye, Z.; Huang, M.; Liu, J.; Liu, Y.; Ke, Z. Nonbifunctional Outer-Sphere Strategy Achieved Highly Active α-Alkylation of Ketones with Alcohols by N-Heterocyclic Carbene Manganese (NHC-Mn). *Org. Lett.* **2019**, *21* (19), 8065−8070.

(27) Kumar, P.; Garg, V.; Kumar, M.; Verma, A. K. Rh(Iii)-Catalyzed Alkynylation: Synthesis of Functionalized Quinolines from Aminohydrazones. *Chem. Commun.* **2019**, *55* (81), 12168−12171.

(28) Martínez, R.; Ramón, D. J.; Yus, M. Transition-Metal-Free Indirect Friedländer Synthesis of Quinolines from Alcohols. *J. Org. Chem.* **2008**, *73* (24), 9778−9780.

(29) Harry, N. A.; Ujwaldev, S. M.; Anilkumar, G. Recent Advances and Prospects in the Metal-Free Synthesis of Quinolines. *Org. Biomol. Chem.* **2020**, *18* (48), 9775−9790.

(30) Ghobadi, N.; Nazari, N.; Gholamzadeh, P. *The Friedländer Reaction: A Powerful Strategy for the Synthesis of Heterocycles*; Scriven, E. F. V, Ramsden, C. A. B. T.-A. H. C., Eds.; Academic Press, 2020; Vol. *132*, Chapter 2, pp 85−134.

(31) Chen, Y.; Stork, C.; Hirte, S.; Kirchmair, J. NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* **2019**, *9* (2), 43.

(32) Reker, D.; Schneider, P.; Schneider, G. Multi-Objective Active Machine Learning Rapidly Improves Structure-Activity Models and Reveals New Protein-Protein Interaction Inhibitors. *Chem. Sci.* **2016**, *7* (6), 3919−3927.

(33) Herrera, V. M.; Khoshgoftaar, T. M.; Villanustre, F.; Furht, B. Random Forest Implementation and Optimization for Big Data Analytics on LexisNexis's High Performance Computing Cluster Platform. *J. Big Data* **2019**, *6* (1), 68.

(34) Ao, Y.; Li, H.; Zhu, L.; Ali, S.; Yang, Z. The Linear Random Forest Algorithm and Its Advantages in Machine Learning Assisted Logging Regression Modeling. *J. Pet. Sci. Eng.* **2019**, *174*, 776−789.

(35) Rodrigues, T. The Good, the Bad, and the Ugly in Chemical and Biological Data for Machine Learning. *Drug Discovery Today: Technol.* **2019**, *32*−*33*, 3−8.

(36) Kaneko, H. Estimation of Predictive Performance for Test Data in Applicability Domains Using Y-Randomization. *J. Chemom.* **2019**, *33* (9), 1−12.

https://doi.org/10.1021/acs.joc.1c01038
J. Org. Chem. XXXX, XXX, XXX−XXX

(37) Rücker, C.; Rücker, G.; Meringer, M. Y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47* (6), 2345−2357.
(38) Chuang, K. V.; Keiser, M. J. Adversarial Controls for Scientific Machine Learning. *ACS Chem. Biol.* **2018**, *13* (10), 2819−2821.
(39) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, No. Section 2, 4766−4775.
(40) Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849−860.
(41) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 49−69.
(42) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem.* **2020**, *6* (7), 1527−1542.
(43) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2* (10), 573−584.
(44) Lundberg, S. M.; Nair, B.; Vavilala, M. S.; Horibe, M.; Eisses, M. J.; Adams, T.; Liston, D. E.; Low, D. K.-W.; Newman, S.-F.; Kim, J.; Lee, S.-I. Explainable Machine-Learning Predictions for the Prevention of Hypoxaemia during Surgery. *Nat. Biomed. Eng.* **2018**, *2* (10), 749−760.
(45) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **2018**, *3* (29), 861.
(46) Scale-up Systems, https://www.scale-up.com/dynochem.
(47) Tillekeratne, V.; Al-Hamashi, A.; Dlamini, S.; Alqahtani, A. S.; Karaj, E. Imidazole-Based Anticancer Agents and Derivatives Thereof, And Methods of Making and Using Same. WO2019036607, US 16638587, 2018.
(48) Zhu, H.; Zhang, M.; Xia, L.; Wang, B. Process for Synthesis of Quinoline Containing Fluorescent Core Compound. CN 107417609, 2017.

https://doi.org/10.1021/acs.joc.1c01038
J. Org. Chem. XXXX, XXX, XXX−XXX