Evaluation of machine learning models for the accelerated prediction of Density Functional Theory calculated ^{19}F chemical shifts based on local atomic environments

Sophia Li[1], Emma Wang[1], Leia Pei[1]*, Sourodeep Deb[1]*, Prashanth Prabhala[1], Sai Hruday Reddy Nara[1], Raina Panda[1], Shiven Eltepu[1], Marx Akl[1], Larry McMahan[1], Edward Njoo[2]**

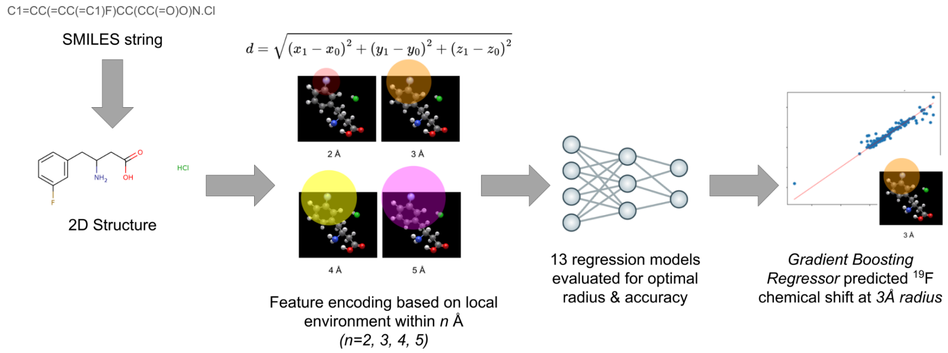
*Authors contributed equally to this work

** Corresponding author

[1]Department of Computer Science & Engineering, Aspiring Scholars Directed Research Program, Fremont, California 94539 USA
[2]Department of Chemistry, Biochemistry and Physics, Aspiring Scholars Directed Research Program, Fremont, California 94539 USA

ARTICLE INFO

Keywords:
Machine learning
^{19}F NMR Spectroscopy
Chemical shifts
Density Functional Theory
Gradient Boosting Regressor
Fluorine



C1=CC(=CC(=C1)F)CC(CC(=O)O)N.Cl
SMILES string
F
NH$_2$
O
OH
HCl
2D Structure
$d = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}$
2 Å
3 Å
4 Å
5 Å
Feature encoding based on local environment within *n* Å *(n=2, 3, 4, 5)*
13 regression models evaluated for optimal radius & accuracy
3 Å
Gradient Boosting Regressor predicted ^{19}F chemical shift at *3Å radius*

ABSTRACT

The introduction of fluorine in compounds plays a crucial role in drug development as it greatly influences their final pharmacokinetic and dynamic properties. Due to the increasing prevalence of fluorine in FDA-approved drugs in recent years, identifying the underlying mechanisms driving their chemical transformations has become crucial in the drug discovery landscape. ^{19}F NMR spectroscopy is a powerful analytical technique that allows for the examination of fluorine-containing compounds, offering valuable information about their structure, dynamics, and reactivity. Consequently, this technique has become a cornerstone in the mechanistic evaluation of fluorine-containing chemical transformations. NMR spectra can be interpreted through the leveraging of Density Functional Theory (DFT), an *ab initio* modeling method that can be harnessed for the prediction of NMR chemical shifts. However, the screening of compounds and discovery of feasible drug candidates is severely limited due to the computational cost of DFT. Here, we present a machine learning approach to accelerate the prediction of DFT-calculated ^{19}F NMR chemical shifts. The fluorine atoms' features in the models were derived from their local three-dimensional structural environments, representing their neighboring atoms within a radius of n Å away from the given fluorine atom in the compound. A comparative analysis of thirteen regression models was conducted using features extracted from 501 fluorinated compounds in our laboratory's chemical inventory. The target chemical shift values were calculated using DFT with the quantum chemistry software ORCA. Among the models evaluated, Gradient Boosting Regression (GBR) exhibited the highest performance, achieving a mean absolute error of 2.89 ppm - 3.73 ppm with a local environment radius of 3 Å. This demonstrates a comparable accuracy to DFT calculations while significantly reducing computational time per compound from several hundred seconds to milliseconds. 3 Å was also found to be the most optimal radius across all models when encoding features for local atomic environments.

1. Introduction

Fluorinated small molecules and materials play an integral role in the modern development of therapeutic agents[1], synthetic polymers[2], agrochemicals[3,4], and industrial materials[5]. Given the unique properties imbued by organofluorine functional groups on a molecule[6] such as enhanced bioavailability and metabolic stability, about one in five small molecules currently approved by the US Food and Drug Administration (FDA) in the last three decades contain one or more fluorinated motifs[7]{Figure 1A}. Such compounds include fluoroquinolone antibiotics such as ciprofloxacin[8], fluorinated nucleoside antivirals such as fluoxetine[9], cholesterol regulators atorvastatin[10], fluorinated corticosteroids such as fluticasone propionate[11], and most recently, the HIV capsid inhibitor lenacapavir, which has recently shown remarkable efficacy in attenuating viral propagation in a late-stage FDA clinical trial{Figure 1B}[12].

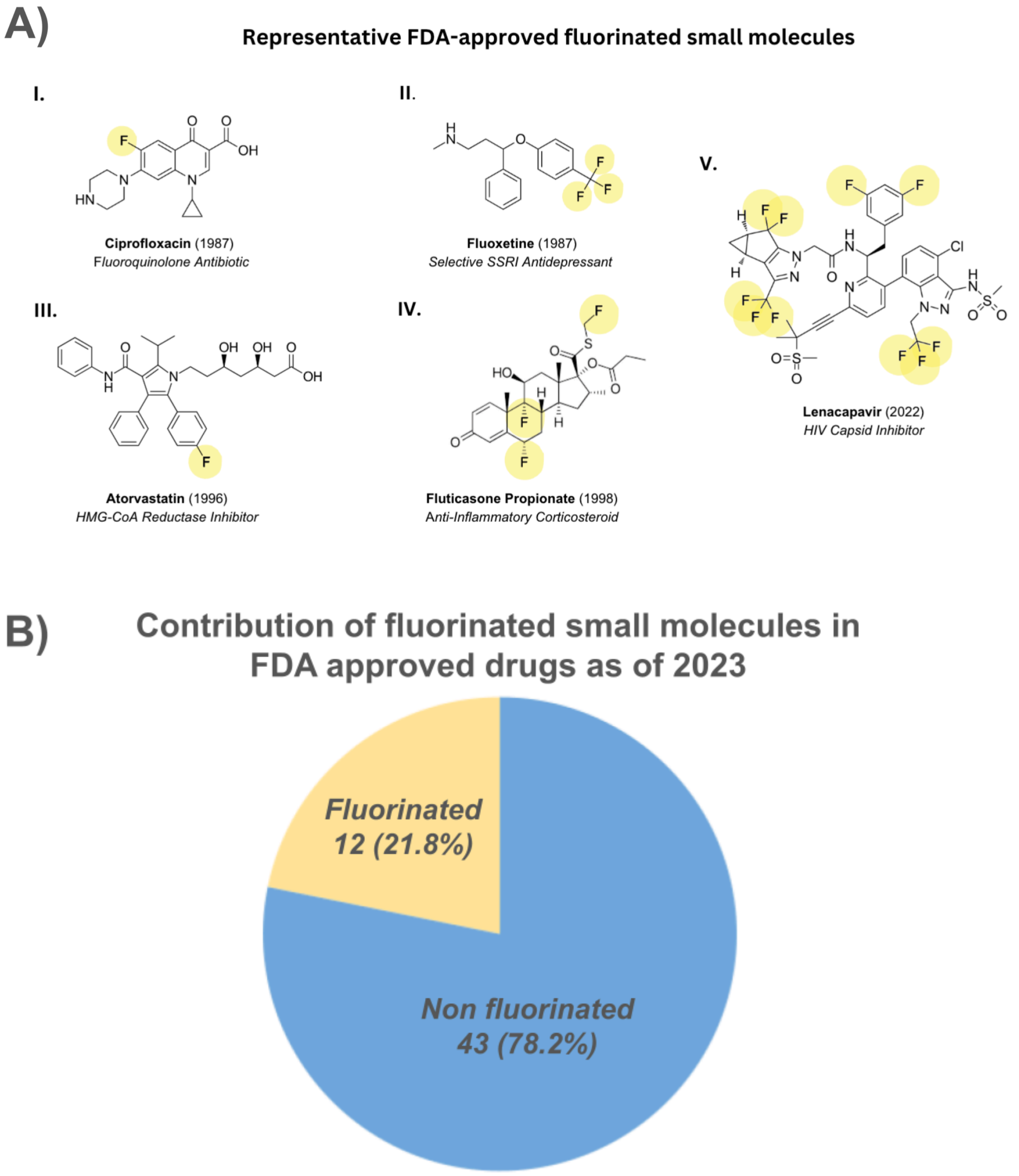

A)
Representative FDA-approved fluorinated small molecules
I.
Ciprofloxacin (1987)
Fluoroquinolone Antibiotic
II.
Fluoxetine (1987)
Selective SSRI Antidepressant
III.
Atorvastatin (1996)
HMG-CoA Reductase Inhibitor
IV.
Fluticasone Propionate (1998)
Anti-Inflammatory Corticosteroid
V.
Lenacapavir (2022)
HIV Capsid Inhibitor
B)
Contribution of fluorinated small molecules in FDA approved drugs as of 2023
Fluorinated 12 (21.8%)
Non fluorinated 43 (78.2%)

Fig. 1. A) Representative FDA-approved fluorinated small molecules: I. Ciprofloxacin, II. Fluoxetine, III. Atorvastatin, IV. Fluticasone propionate, V. Lenacapavir. **B) Pie chart**

representation of fluorinated FDA-approved drugs in 2023. This data was taken from Inoue et al.[7]

Consequently, ^{19}F NMR spectroscopy has become fundamental in structural characterization [13], purity analysis[14], and screening[15] of these compounds. This technique has also been used to study biological phenomena, including biomolecular structure and function[16–18], enzymatic mechanisms[19,20], metabolic pathways[21], drug screening[22–24], and medical imaging[25,26]. Moreover, given the broad applicability of fluorinated organic molecules, the application of ^{19}F NMR spectroscopy for the real-time deconvolution of complex reaction mechanisms has become a primary focus for our group and many others[27,28].

Density Functional Theory (DFT) is an *ab initio* computational modeling method applying quantum mechanical principles to accurately describe the distribution of electrons within molecules, enabling the precise predictions of electronic molecular structures[29–31]. DFT calculates the shielding effects experienced by nuclei due to surrounding electron densities to determine NMR chemical shifts[32–36]. However, the computational expense of such calculations can prove prohibitive for researchers, limiting practicality when handling large datasets[37–40]. In order to address this, machine learning techniques have been employed to generate similar predictive values at a fraction of the compute cost. This provides an accelerated shortcut for the normalized process while maintaining a reasonable balance between speed and accuracy[41–45], allowing for the prediction of chemical properties of previously unknown structures. Other advanced modeling techniques have also been recently introduced in the forms of graph neural networks (GNN) to predict ^{19}F and ^{13}C chemical shifts by Li et al. and Rull et al. respectively[46,47]. While many of these models rely on one-dimensional (e.g. SMILES) or two-dimensional (e.g. graphs) chemical representations, they do not contain enough information given the inherent three-dimensional nature of anisotropic and bond-polarization effects that contribute to NMR shielding constants. To address this, we postulate that the simplest representation of each fluorine's local electronic environment can be gathered through encoding three-dimensional structural information of proximal neighboring atoms along with electronegativity and atomic weight. Due to the 3D nature of this input data, machine learning models trained using this dataset will be able to capture local atomic environments more effectively than the aforementioned 2D or 1D representations. Specifically, these models will have the ability to form crucial connections paralleling real time chemical interactions from the information offered through this third dimension within our data. This featural input can then be used to train machine learning models to predict DFT-calculated ^{19}F NMR chemical shifts at a fraction of the computational cost associated with a full DFT calculation.

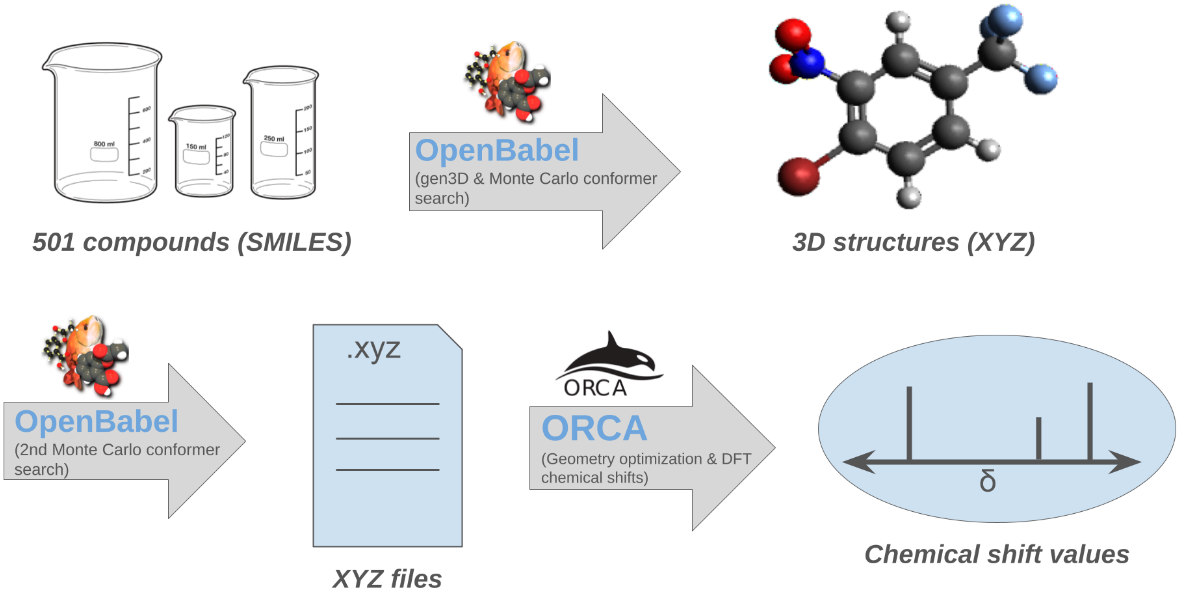

OpenBabel
(gen3D & Monte Carlo conformer search)
501 compounds (SMILES)
3D structures (XYZ)
OpenBabel
(2nd Monte Carlo conformer search)
.xyz
XYZ files
ORCA
ORCA
(Geometry optimization & DFT chemical shifts)
δ
Chemical shift values

Fig. 2. Data preprocessing workflow. This was used to extract chemical shift values from fluorinated compounds. We begin with SMILES strings of fluorinated compounds pulled from our chemical inventory, then convert them into 3D structure form using the gen3D and Monte Carlo search functions in Open Babel. Additionally, Open Babel was used to convert the optimized 3D structures into XYZ coordinate format. Further details can be found in our Electronic Supplementary Information.

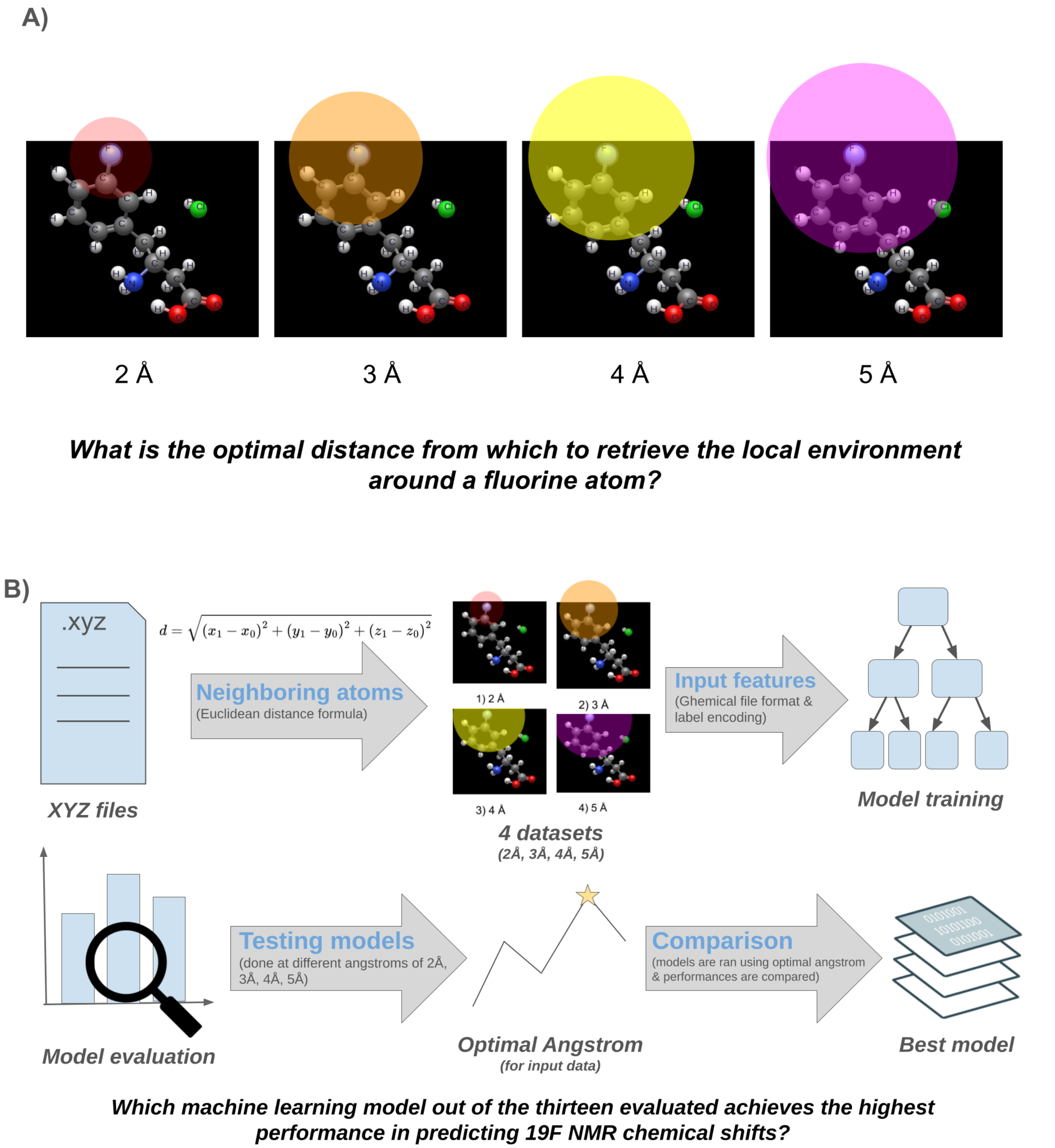

A)
2 Å
3 Å
4 Å
5 Å
What is the optimal distance from which to retrieve the local environment around a fluorine atom?
B)
.xyz
XYZ files
$d=\sqrt{(x_1-x_0)^2+(y_1-y_0)^2+(z_1-z_0)^2}$
Neighboring atoms
(Euclidean distance formula)
1) 2 Å
2) 3 Å
3) 4 Å
4) 5 Å
4 datasets
(2Å, 3Å, 4Å, 5Å)
Input features
(Ghemical file format & label encoding)
Model training
Model evaluation
Testing models
(done at different angstroms of 2Å, 3Å, 4Å, 5Å)
Optimal Angstrom
(for input data)
Comparison
(models are ran using optimal angstrom & performances are compared)
Best model
Which machine learning model out of the thirteen evaluated achieves the highest performance in predicting 19F NMR chemical shifts?

Fig. 3. A) Visualization of varying local environments with an example compound structure. Areas spanning 2.0 Å radius, 3.0 Å radius, 4.0 Å radius, and 5.0 Å radius away from the given fluorine atom. **B) Workflow for model training and testing.** The proximity of neighboring atoms of each fluorine nucleus was calculated with the Euclidean distance formula, forming four datasets of atomic environment information at varying radii. Input features were then extracted for model training from these geometrically optimized coordinates in Ghemical

file format and subsequently encoded. Optimal distance for input data was then determined after model evaluation and testing based on runtime speed, MAE, and R^2. Upon running all the models using the dataset tailored to this optimal atomic radius, we were then able to determine the best performer.

Here, we present a comparative study of thirteen machine learning models in predicting ^{19}F NMR chemical shifts given structural input data. Since different model families exhibit varying levels of applicability to this task due to distinct model architectures, this greatly influenced the specific models evaluated in our project. For instance, ensemble models are well suited for capturing intricate non-linear relationships, making them an attractive candidate. In contrast, simpler linear models may struggle with the complexity of the input data but are less prone to overfitting due to simpler parameters. Due to these factors, we selected models from these two major families in our approach.

Through this, we seek to identify models that accelerate the prediction of chemical shifts while minimizing the loss of accuracy in addition to elucidating the optimal radii when encoding neighboring atoms as model features. Out of the thirteen machine learning models tested, we found that Gradient Boosting Regressor was most suitable for this specific task. Furthermore, we found 3 Å to be the optimal radius when extracting local atomic information of fluorine atoms, as all thirteen models generally performed their best when trained with this dataset.

2. Material and methods

The data curation began with 501 fluorinated organic compounds and their respective SMILES strings from our laboratory's chemical inventory. The SMILES strings were first converted to an initial 3D structure using Openbabel's gen3D function. These initial 3D structures were then optimized using a Monte Carlo conformer search with the MMFF94 force field using Open Babel's *obconformer* module[48]. The resulting optimized conformers' 3D coordinates were extracted into xyz files, which were then used as input for Orca's geometry optimization function[49,50] with the B3LYP functional and 6-31G(d,p) basis set. Using the final optimized 3D structures, the ^{19}F NMR chemical shifts for each compound were calculated at the B3LYP / 6-31G(d,p) level of theory. Subsequently, the DFT-optimized compound's coordinates were converted from xyz format to Ghemical format for additional features, such as atomic number and charge.

To extract the local environment around each fluorine atom within our dataset, we looped through the 3D coordinates of each fluorinated compound to find the Euclidean distance from each atom to a given fluorine atom within the compound according to Equation 1.

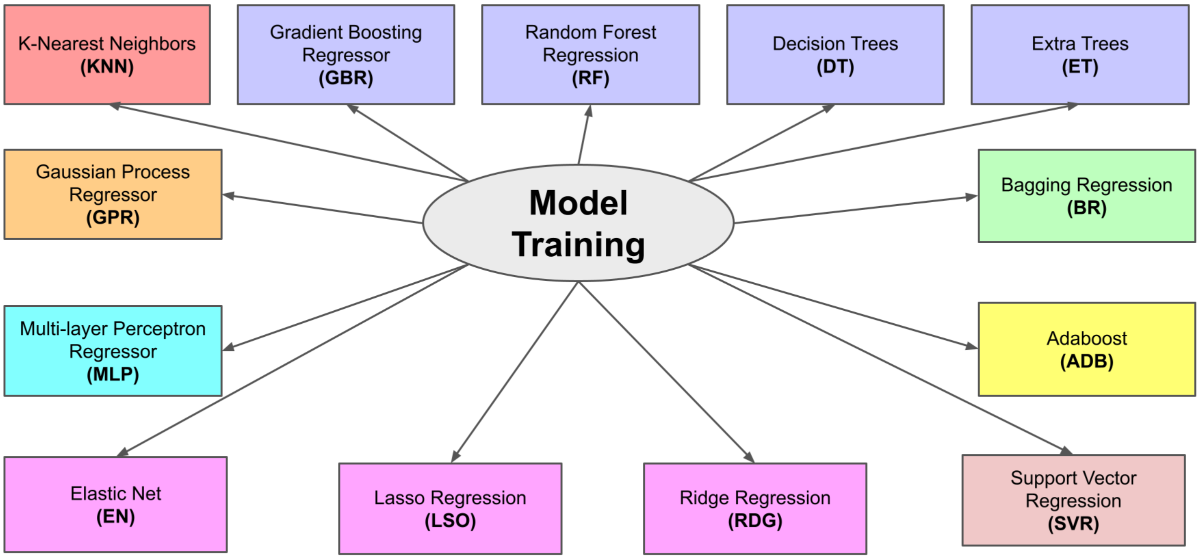
$$d = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} \qquad (1)$$

We define an atom as a "neighbor" of the fluorine atom if its distance from the given fluorine is less than or equal to *n* Å. To determine the optimal distance from which to select "neighbors," we set *n* = 2 Å, 3 Å, 4 Å, and 5 Å in four independent data processing rounds, yielding four different datasets with varying "neighbor" sets for each fluorine atom. Additionally, for every "neighbor," we included its atomic number and charge (from the Ghemical files), electronegativity and mass number (from Python's Mendeleev library), and its element symbol label-encoded into a numeric representation. This leads to each dataset varying in terms of capturing the complexity of the given fluorine's local atomic environment.

The final datasets included 1161 entries, with each entry representing a fluorine atom. The input features included each of the fluorine atoms' neighbors with their atomic number, charge, electronegativity, mass number, and label-encoded symbol. The target values were the fluorine atoms' DFT calculated chemical shifts. To adjust the size of the data for the regression models, missing values were padded with zeros. Each dataset was then split into an 80% training set and a 20% testing set.

We trained and evaluated the following regression models: Gradient Boosting Regressor (GBR)[51,52], Gaussian Process Regressor (GPR)[53], Decision Trees (DT)[54,55], Random Forest Regression (RF)[56–58], Extra Trees (ET)[59], Adaboost (ADB)[60,61], Bagging Regression (BR)[62], Lasso Regression (LSO)[63,64], Ridge Regression (RDG)[65,66], Elastic Net (EN)[67,68], Multi-layer Perceptron Regressor (MLP)[69], Support Vector Regression

(SVR)[70], and K-Nearest Neighbors (KNN)[71,72] {Figure 2}. All models were implemented in Python using the scikit-learn library[73].



Model Training
K-Nearest Neighbors (KNN)
Gradient Boosting Regressor (GBR)
Random Forest Regression (RF)
Decision Trees (DT)
Extra Trees (ET)
Gaussian Process Regressor (GPR)
Bagging Regression (BR)
Multi-layer Perceptron Regressor (MLP)
Adaboost (ADB)
Elastic Net (EN)
Lasso Regression (LSO)
Ridge Regression (RDG)
Support Vector Regression (SVR)

Fig. 4: Model training diagram of all 13 models, color-coded by model “family” (decision tree-based models are purple; linear regression models with penalties are in pink; all other models of different families were assigned unique colors).

Each model’s hyperparameters were optimized using the Bayesian optimization algorithm[74] from a combination of Python’s Bayesian optimization library[75] and scikit-learn’s Bayes Search CV. Hyperparameter search space bounds were manually defined. In use cases of the Bayesian optimization library, the objective functions were written to optimize by either mean absolute error (MAE) or R^2. The accuracy of each model in predicting DFT-calculated ^{19}F NMR chemical shifts was evaluated using MAE and R^2.

3. Results and Discussion

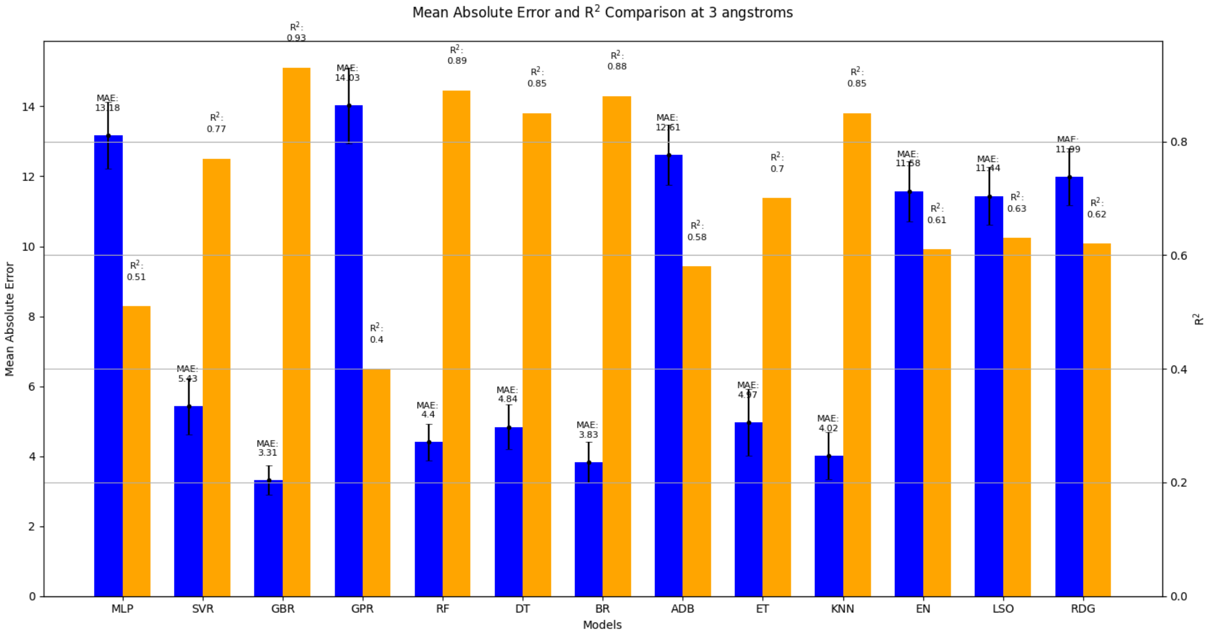


Fig. 5: Performances of various ML algorithms on the 3 Å dataset. Evaluation metrics were MAE and R^2 for all tested models.

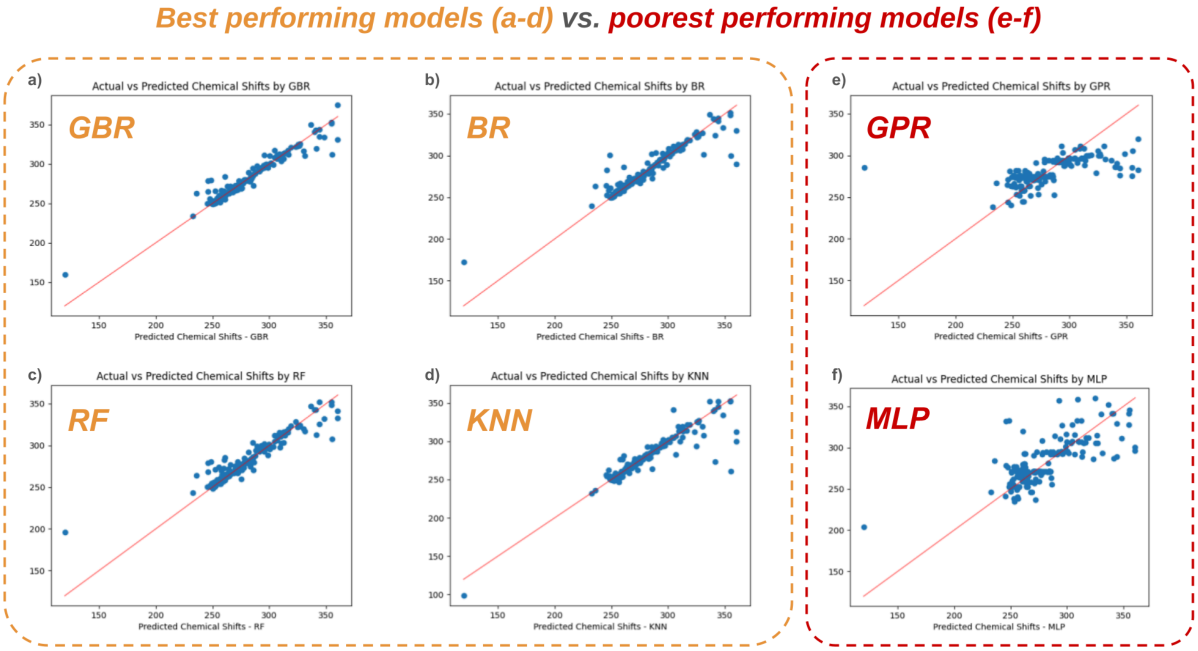


Fig. 6: **Model performance graphs.** Prediction diagrams (predicted values vs actual values) of the four best-performing models: a) GBR, b) BR, c) RF, d) KNN and the two poorest performing models: e) GPR, f) MLP.

Following fine-tuning of all regression models, GBR continued to have the lowest MAE and highest R^2 values, 3.31 ppm, and 0.93, respectively. RF also exhibited high performance with an R^2 of approximately 0.89 and MAE of 4.39 ppm. Similarly, BR showed strong performance, achieving an R^2 of 0.88 and the second lowest MAE of 3.82 ppm compared to all 13 models. Additionally, KNN had a comparable performance, with a slightly lower R^2 of 0.85, but a relatively low MAE of 4.01 ppm. The least accurate models were MLP with an R^2 of 0.51 and MAE of 13.18 ppm, as well as GPR with an R^2 of 0.4 and MAE of 12.95 ppm - 15.11 ppm {Figure 4}. Furthermore, GBR, BR, RF, and KNN demonstrated superior performance by exhibiting close adherence to the ground truth line in the prediction plots. Conversely, the models GPR and MLP displayed a more dispersed distribution, indicating less accurate predictions {Figure 5}.We observed no trends within the performance of these models that could be attributed to inherent differences in the model family.

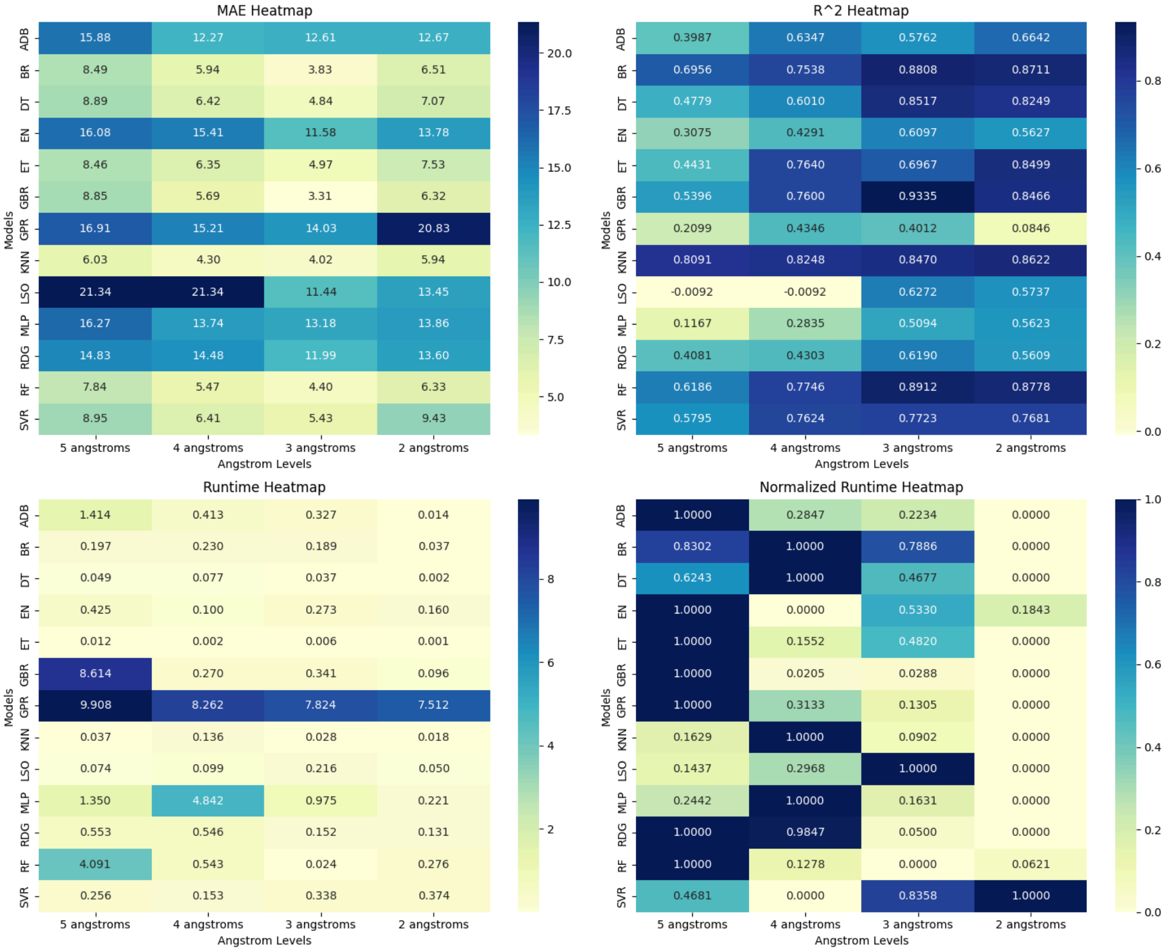


Figure 7: Angstrom comparison visualizations by heatmaps: runtime, normalized runtime, MAE, and R^2 by models trained on extracted neighbor features within n Å radii of each fluorine atom. Runtimes are calculated on the time it takes for the model to predict on the testing data, i.e. around 232 compounds.

To address the question of the optimal distance from each fluorine atom from which to retrieve its neighbors, we ran a comparative analysis of each model's performance on datasets consisting of neighbors taken at 2-5 Å radii. Results are shown in heatmaps of runtime, MAE, and R^2 for each model on the 2 Å, 3 Å, 4 Å, and 5 Å radii datasets {Figure 6}.

For every model, the 3 Å dataset yields the lowest MAE values and the highest R^2 values for 9 models out of the 13. Overall, the best performance across all models and all potential radii was shown by the GBR model at 3 Å with an MAE of 2.89 ppm - 3.73 ppm and R^2 of 0.93.

To address the balance between efficiency and model performance, which we anticipated would vary with each dataset due to differences in atomic environmental complexity, we recorded the runtimes of each models' training and prediction process. Since the runtimes vary from model to model due to differences in architecture, we normalized the plotted runtimes for each model in the heatmap above for ease of interpretation, where we generally observed an exponential decrease between runtime and radii after this normalization. Expectedly, the 2 Å dataset had the fastest runtime overall among all the models. On the other hand, the superior model performance while using the 3 Å dataset (as shown through low MAE and R^2 values) led to the conclusion that the majority of models performed their best when trained on the 3 Å dataset. Therefore, we deduced that the optimal distance when retrieving local atomic environments of each fluorine atom is 3 Å.

In comparing model performance on varying radii, we aimed to identify the optimal balance between providing the model with too much information (with a radius too large) and not enough information (with a radius too small to adequately capture local environmental factors) without sacrificing our predictive accuracy[76,77]. Our results indicated that a radius of 3 Å most effectively addressed this bias-variance tradeoff between the fluorine atom's environmental complexity and the model's predictive accuracy[78–82]. For model performance using datasets of radii of 4 Å and 5 Å, it is possible that too much extraneous information in the input features (that have little true correlation with the NMR shifts) was introduced into the models, potentially leading to overfitting and poor generalization[83].

When tested with the 3 Å dataset, the best-performing model overall was GBR with the lowest MAE value of 3.31196, followed by BR with an MAE of 3.8294 and KNN with 4.0169. The least predictive models were MLP and GPR, with MAE values of 13.1814 and 14.0274, respectively.

In summary, the dataset utilized for model training and testing consisted of atomic numbers, charges, electronegativities, mass numbers, and label-encoded symbols of each fluorine atom's neighbors as input features (X). The compounds' corresponding DFT shifts[84,85] were considered as the output (y). While DFT can shorten the time needed to calculate the shifts of all 501 compounds–in contrast to producing experimental ^{19}F NMR shifts over the course of months or years (depending on dataset length)–DFT has limited predictive power. Imperatively, computational methods will never be equal in accuracy to real experimental methods[86–88]. Specifically, the inability of DFT to fully account for dynamic averaging, relativistic effects, and electron correlation can lead to small discrepancies between predicted and experimentally observed chemical shifts[89–91]. This deviation typically ranges from 3 ppm to 10 ppm for the shifts of fluorinated compounds. Models trained with DFT shifts as output will therefore be less accurate than DFT itself. In our study, the MAE of our best performing model (GBR) was 3.5 ppm when using the DFT calculated shifts as ground truth. Therefore, its overall MAE from experimentally calculated shifts would likely range from 4.5 to 13.5 ppm.

4. Conclusion

In this study, we systematically screened and evaluated thirteen machine learning platforms to identify the most optimal means of predicting DFT-calculated ^{19}F NMR chemical shifts. As in most end-case applications, not all machine learning models perform equally well under the constraints provided. From a dataset of several hundred fluorinated chemical structures, our highest performing model, GBR, achieved an R^2 of 0.93 and MAE of 2.89 ppm - 3.73 ppm, with a runtime (for 3 Å) of 0.34 seconds. As suggested by the results of this model in addition to the other 12 evaluated, a radius of 3 Å is the optimal distance around each fluorine atom of the compound structure from which to extract neighboring atoms as model input features. This allows for consistent high performances and an effective bias-variance tradeoff.

Given the widespread applicability of fluorinated compounds within a broad spectrum of industrial, biopharmaceutical, and other chemical spaces, the rapid and precise prediction of ^{19}F NMR chemical shifts with machine learning presents an attractive opportunity to decrease the computational cost of high throughput screening approaches of fluorinated chemical entities.

Ultimately, the applicability of this and similar approaches of integrating machine learning approaches in predictive modeling of computationally costly calculations will certainly enhance modern efforts in improving efficiency and accuracy of targeted predictions of molecular properties. In the future, integrating Graph Neural Networks (GNNs) to predict ^{19}F NMR chemical shifts, leveraging their ability to learn from the 3D molecular structures, may further improve DFT-level accuracy. Additionally, developing machine learning models trained directly on experimental NMR data could expand the applicability of these models beyond DFT-calculated shifts. Such efforts are currently underway in our group and will be reported in due course of time.

Funding Declaration

The authors state no funding is involved.

Credit authorship contribution statement

Sophia Li: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Emma Wang:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Leia Pei:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration **Sourodeep Deb:** Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **Prashanth Prabhala:** Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization **Sai Hruday Reddy Nara:** Methodology, Software, Validation, Formal analysis, Investigation **Raina Panda:** Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing **Shiven Eltepu:** Methodology, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing **Larry McMahan:** Supervision, Writing - Review & Editing **Marx Akl:** Supervision **Edward Njoo:** Conceptualization, Methodology, Resources, Writing - Review & Editing

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We thank the Aspiring Scholars Directed Research Program for providing us with their ChemInventory data of fluorinated compounds as well as access to their high-performance in-house server cluster. We would also like to thank Dennis Liu for providing valuable feedback on our code.

Data availability

Our data and code can be found in our Github repository. Documentation of source data and code can be found at our Electronic Supporting Information.

References

[1] K.L. Kirk, Fluorine in medicinal chemistry: Recent therapeutic applications of fluorinated small molecules, J. Fluor. Chem. 127 (2006) 1013–1029. https://doi.org/10.1016/j.jfluchem.2006.06.007.

[2] V.F. Cardoso, D.M. Correia, C. Ribeiro, M.M. Fernandes, S. Lanceros-Méndez, Fluorinated Polymers as Smart Materials for Advanced Biomedical Applications, Polymers 10 (2018) 161. https://doi.org/10.3390/polym10020161.

[3] T. Fujiwara, D. O'Hagan, Successful fluorine-containing herbicide agrochemicals, J. Fluor. Chem. 167 (2014) 16–29. https://doi.org/10.1016/j.jfluchem.2014.06.014.

[4] Y. Ogawa, E. Tokunaga, O. Kobayashi, K. Hirai, N. Shibata, Current Contributions of Organofluorine Compounds to the Agrochemical Industry, iScience 23 (2020) 101467. https://doi.org/10.1016/j.isci.2020.101467.

[5] J. Gardiner, Fluoropolymers: Origin, Production, and Industrial and Commercial Applications, Aust. J. Chem. 68 (2014) 13–22. https://doi.org/10.1071/CH14165.

[6] E.P. Gillis, K.J. Eastman, M.D. Hill, D.J. Donnelly, N.A. Meanwell, Applications of Fluorine in Medicinal Chemistry, J. Med. Chem. 58 (2015) 8315–8359. https://doi.org/10.1021/acs.jmedchem.5b00258.

[7] M. Inoue, Y. Sumii, N. Shibata, Contribution of Organofluorine Compounds to Pharmaceuticals, ACS Omega 5 (2020) 10633–10640. https://doi.org/10.1021/acsomega.0c00830.

[8] P.C. Sharma, A. Jain, S. Jain, R. Pahwa, M.S. Yar, Ciprofloxacin: review on developments in synthetic, analytical, and medicinal aspects, J. Enzyme Inhib. Med. Chem. 25 (2010) 577–589. https://doi.org/10.3109/14756360903373350.

[9] W.-D. Meng, F.-L. Qing, Fluorinated nucleosides as antiviral and antitumor agents, Curr. Top. Med. Chem. 6 (2006) 1499–1528. https://doi.org/10.2174/156802606777951082.

[10] A.P. Lea, D. McTavish, Atorvastatin, Drugs 53 (1997) 828–847. https://doi.org/10.2165/00003495-199753050-00011.

[11] Y.A. Jasem, T. Thiemann, L. Gano, M.C. Oliveira, Fluorinated steroids and their derivatives, J. Fluor. Chem. 185 (2016) 48–85. https://doi.org/10.1016/j.jfluchem.2016.03.009.

[12] J.O. Link, M.S. Rhee, W.C. Tse, J. Zheng, J.R. Somoza, W. Rowe, R. Begley, A. Chiu, A. Mulato, D. Hansen, E. Singer, L.K. Tsai, R.A. Bam, C.-H. Chou, E. Canales, G. Brizgys, J.R. Zhang, J. Li, M. Graupe, P. Morganelli, Q. Liu, Q. Wu, R.L. Halcomb, R.D. Saito, S.D. Schroeder, S.E. Lazerwith, S. Bondy, D. Jin, M. Hung, N. Novikov, X. Liu, A.G. Villaseñor, C.E. Cannizzaro, E.Y. Hu, R.L. Anderson, T.C. Appleby, B. Lu, J. Mwangi, A. Liclican, A. Niedziela-Majka, G.A. Papalia, M.H. Wong, S.A. Leavitt, Y. Xu, D. Koditek, G.J. Stepan, H. Yu, N. Pagratis, S. Clancy, S. Ahmadyar, T.Z. Cai, S. Sellers, S.A. Wolckenhauer, J. Ling, C. Callebaut, N. Margot, R.R. Ram, Y.-P. Liu, R. Hyland, G.I. Sinclair, P.J. Ruane, G.E. Crofoot, C.K. McDonald, D.M. Brainard, L. Lad, S. Swaminathan, W.I. Sundquist, R. Sakowicz, A.E. Chester, W.E. Lee, E.S. Daar, S.R. Yant, T. Cihlar, Clinical targeting of HIV capsid protein with a long-acting small molecule, Nature 584 (2020) 614–618. https://doi.org/10.1038/s41586-020-2443-1.

[13] J. Jiang, L. Wen, H. Wang, X. Chen, Y. Zhao, X. Wang, Detection and identification of amphetamine-type stimulants and analogs via recognition-enabled “chromatographic” ^{19}F NMR, J. Fluor. Chem. 266 (2023) 110085. https://doi.org/10.1016/j.jfluchem.2023.110085.

[14] N. Mistry, I.M. Ismail, R. Duncan Farrant, M. Liu, J.K. Nicholson, J.C. Lindon, Impurity profiling in bulk pharmaceutical batches using ^{19}F NMR spectroscopy and distinction between monomeric and dimeric impurities by NMR-based diffusion measurements, J. Pharm. Biomed. Anal. 19 (1999) 511–517. https://doi.org/10.1016/S0731-7085(98)00247-7.

[15] A. Lingel, A. Vulpetti, T. Reinsperger, A. Proudfoot, R. Denay, A. Frommlet, C. Henry, U. Hommel, A.D. Gossert, B. Luy, A.O. Frank, Comprehensive and High-Throughput Exploration of Chemical Space Using Broadband ^{19}F NMR-Based Screening, Angew. Chem. Int. Ed. 59 (2020) 14809–14817. https://doi.org/10.1002/anie.202002463.

[16] G. Papeo, P. Giordano, M.G. Brasca, F. Buzzo, D. Caronni, F. Ciprandi, N. Mongelli, M. Veronesi, A. Vulpetti, C. Dalvit, Polyfluorinated Amino Acids for Sensitive ^{19}F NMR-Based Screening and Kinetic Measurements, J. Am. Chem. Soc. 129 (2007) 5665–5672. https://doi.org/10.1021/ja069128s.

[17] L. Yu, P.J. Hajduk, J. Mack, E.T. Olejniczak, Structural studies of Bcl-xL/ligand complexes using ^{19}F NMR, J. Biomol. NMR 34 (2006) 221–227. https://doi.org/10.1007/s10858-006-0005-y.

[18] D. Gimenez, A. Phelan, C.D. Murphy, S.L. Cobb, ^{19}F NMR as a tool in chemical biology, Beilstein J. Org. Chem. 17 (2021) 293–318. https://doi.org/10.3762/bjoc.17.28.

[19] H.W. Lee, J.H. Sohn, B.I. Yeh, J.W. Choi, S. Jung, H.W. Kim, ^{19}F NMR investigation of F(1)-ATPase of Escherichia coli using fluorotryptophan labeling, J. Biochem. (Tokyo) 127 (2000) 1053–1056. https://doi.org/10.1093/oxfordjournals.jbchem.a022697.

[20] T.A. Sales, M.A. Gonçalves, T.C. Ramalho, Structural Parameters of the Interaction between Ciprofloxacin and Human Topoisomerase-II β Enzyme: Toward New ^{19}F NMR Chemical Shift Probes, Magnetochemistry 8 (2022) 181.

https://doi.org/10.3390/magnetochemistry8120181.

[21] T. Nakada, I.L. Kwee, P.J. Card, N.A. Matwiyoff, B.V. Griffey, R.H. Griffey, Fluorine-19 NMR imaging of glucose metabolism, Magn. Reson. Med. 6 (1988) 307–313. https://doi.org/10.1002/mrm.1910060309.

[22] R.S. Prosser, A beginner's guide to ^{19}F NMR and its role in drug screening, Can. J. Chem. 101 (2023) 758–764. https://doi.org/10.1139/cjc-2023-0028.

[23] F.-F. Zhang, M.-H. Jiang, L.-L. Sun, F. Zheng, L. Dong, V. Shah, W.-B. Shen, Y. Ding, Quantitative analysis of sitagliptin using the ^{19}F-NMR method: a universal technique for fluorinated compound detection, Analyst 140 (2014) 280–286. https://doi.org/10.1039/C4AN01681E.

[24] C.R. Buchholz, W.C.K. Pomerantz, ^{19}F NMR viewed through two different lenses: ligand-observed and protein-observed ^{19}F NMR applications for fragment-based drug discovery, RSC Chem. Biol. 2 (n.d.) 1312–1330. https://doi.org/10.1039/d1cb00085c.

[25] G.N. Holland, P.A. Bottomley, W.S. Hinshaw, ^{19}F magnetic resonance imaging, J. Magn. Reson. 1969 28 (1977) 133–136. https://doi.org/10.1016/0022-2364(77)90263-3.

[26] I. Tirotta, V. Dichiarante, C. Pigliacelli, G. Cavallo, G. Terraneo, F.B. Bombelli, P. Metrangolo, G. Resnati, ^{19}F Magnetic Resonance Imaging (MRI): From Design of Materials to Clinical Applications, Chem. Rev. 115 (2015) 1106–1129. https://doi.org/10.1021/cr500286d.

[27] X. Wang, J. Vu, C. Luk, E. Njoo, Benchtop ^{19}F nuclear magnetic resonance spectroscopy enabled kinetic studies and optimization of the synthesis of carmofur, Can. J. Chem. 101 (2023) 518–524. https://doi.org/10.1139/cjc-2022-0266.

[28] R. Chen, P. Singh, S. Su, S. Kocalar, X. Wang, N. Mandava, S. Venkatesan, A. Ferguson, A. Rao, E. Le, C. Rojas, E. Njoo, Benchtop ^{19}F Nuclear Magnetic Resonance (NMR) Spectroscopy Provides Mechanistic Insight into the Biginelli Condensation toward the Chemical Synthesis of Novel Trifluorinated Dihydro- and Tetrahydropyrimidinones as Antiproliferative Agents, ACS Omega 8 (2023) 10545–10554. https://doi.org/10.1021/acsomega.3c00290.

[29] T. van Mourik, M. Bühl, M.-P. Gaigeot, Density functional theory across chemistry, physics and biology, Philos. Transact. A Math. Phys. Eng. Sci. 372 (2014) 20120488. https://doi.org/10.1098/rsta.2012.0488.

[30] V. Butera, Density functional theory methods applied to homogeneous and heterogeneous catalysis: a short review and a practical user guide, Phys. Chem. Chem. Phys. 26 (2024) 7950–7970. https://doi.org/10.1039/D4CP00266K.

[31] E. Sim, S. Song, S. Vuckovic, K. Burke, Improving Results by Improving Densities: Density-Corrected Density Functional Theory, J. Am. Chem. Soc. 144 (2022) 6625–6639. https://doi.org/10.1021/jacs.1c11506.

[32] Z.S. Safi, N. Wazzan, DFT calculations of 1H- and 13C-NMR chemical shifts of 3-methyl-1-phenyl-4-(phenyldiazenyl)-1H-pyrazol-5-amine in solution, Sci. Rep. 12 (2022) 17798. https://doi.org/10.1038/s41598-022-22900-y.

[33] V. Butera, L. D'Anna, S. Rubino, R. Bonsignore, A. Spinello, A. Terenzi, G. Barone, How the Metal Ion Affects the 1H NMR Chemical Shift Values of Schiff Base Metal Complexes: Rationalization by DFT Calculations, J. Phys. Chem. A 127 (2023) 9283–9290. https://doi.org/10.1021/acs.jpca.3c05653.

[34] P. Geerlings, F. De Proft, W. Langenaeker, Conceptual Density Functional Theory, Chem. Rev. 103 (2003) 1793–1874. https://doi.org/10.1021/cr990029p.

[35] W. Kohn, A.D. Becke, R.G. Parr, Density Functional Theory of Electronic Structure, J. Phys. Chem. 100 (1996) 12974–12980. https://doi.org/10.1021/jp960669l.

[36] R.O. Jones, Density functional theory: Its origins, rise to prominence, and future, Rev. Mod. Phys. 87 (2015) 897–923. https://doi.org/10.1103/RevModPhys.87.897.

[37] L. Fiedler, K. Shah, M. Bussmann, A. Cangi, Deep dive into machine learning density functional theory for materials science and chemistry, Phys. Rev. Mater. 6 (2022) 040301. https://doi.org/10.1103/PhysRevMaterials.6.040301.

[38] S. Wieser, E. Zojer, Machine learned force-fields for an Ab-initio quality description of metal-organic frameworks, Npj Comput. Mater. 10 (2024) 1–18. https://doi.org/10.1038/s41524-024-01205-w.

[39] Y.Y. Rusakov, V.A. Semenov, I.L. Rusakova, On the Efficiency of the Density Functional Theory (DFT)-Based Computational Protocol for 1H and 13C Nuclear Magnetic Resonance (NMR) Chemical Shifts of Natural Products: Studying the Accuracy of the pecS-n (n = 1, 2) Basis Sets, Int. J. Mol. Sci. 24 (2023) 14623. https://doi.org/10.3390/ijms241914623.

[40] B. Lu, Y. Xia, Y. Ren, M. Xie, L. Zhou, G. Vinai, S.A. Morton, A.T.S. Wee, W.G. van der Wiel, W. Zhang, P.K.J. Wong, When Machine Learning Meets 2D Materials: A Review, Adv. Sci. 11 (2024) 2305277. https://doi.org/10.1002/advs.202305277.

[41] X. Wan, Z. Zhang, W. Yu, Y. Guo, A density-functional-theory-based and machine-learning-accelerated hybrid method for intricate system catalysis, Mater. Rep. Energy 1 (2021) 100046. https://doi.org/10.1016/j.matre.2021.100046.

[42] Y. Kwon, D. Lee, Y.-S. Choi, M. Kang, S. Kang, Neural Message Passing for NMR Chemical Shift Prediction, J. Chem. Inf. Model. 60 (2020) 2024–2030. https://doi.org/10.1021/acs.jcim.0c00195.

[43] M. Haghighatlari, J. Li, F. Heidar-Zadeh, Y. Liu, X. Guan, T. Head-Gordon, Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods, Chem 6 (2020) 1527–1542. https://doi.org/10.1016/j.chempr.2020.05.014.

[44] M. Cordova, E.A. Engel, A. Stefaniuk, F. Paruzzo, A. Hofstetter, M. Ceriotti, L. Emsley, A Machine Learning Model of Chemical Shifts for Chemically and Structurally Diverse Molecular Solids, J. Phys. Chem. C 126 (2022) 16710–16720. https://doi.org/10.1021/acs.jpcc.2c03854.

[45] P.A. Unzueta, C.S. Greenwell, G.J.O. Beran, Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning, J. Chem. Theory Comput. 17 (2021) 826–840. https://doi.org/10.1021/acs.jctc.0c00979.

[46] Y. Li, W.-S. Huang, L. Zhang, D. Su, H. Xu, X.-S. Xue, Prediction of ^{19}F NMR chemical shift by machine learning, Artif. Intell. Chem. 2 (2024) 100043. https://doi.org/10.1016/j.aichem.2024.100043.

[47] H. Rull, M. Fischer, S. Kuhn, NMR shift prediction from small data quantities, J. Cheminformatics 15 (2023) 114. https://doi.org/10.1186/s13321-023-00785-x.

[48] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open Babel: An open chemical toolbox, J. Cheminformatics 3 (2011) 33. https://doi.org/10.1186/1758-2946-3-33.

[49] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, The ORCA quantum chemistry program package, J. Chem. Phys. 152 (2020) 224108. https://doi.org/10.1063/5.0004608.

[50] F. Neese, Software update: The ORCA program system—Version 5.0, WIREs Comput. Mol. Sci. 12 (2022) e1606. https://doi.org/10.1002/wcms.1606.

[51] J.H. Friedman, Greedy function approximation: A gradient boosting machine., Ann. Stat. 29 (2001) 1189–1232. https://doi.org/10.1214/aos/1013203451.

[52] L. Mason, J. Baxter, P. Bartlett, M. Frean, Boosting Algorithms as Gradient Descent, in: Adv. Neural Inf. Process. Syst., MIT Press, 1999. https://proceedings.neurips.cc/paper_files/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html (accessed June 20, 2024).

[53] C. Williams, C. Rasmussen, Gaussian Processes for Regression, in: Adv. Neural Inf. Process. Syst., MIT Press, 1995. https://papers.nips.cc/paper_files/paper/1995/hash/7cce53cf90577442771720a370c3c723-Abstract.html (accessed June 20, 2024).

[54] Y. SONG, Y. LU, Decision tree methods: applications for classification and prediction, Shanghai Arch. Psychiatry 27 (2015) 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.

[55] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1986) 81–106. https://doi.org/10.1007/BF00116251.

[56] L. Breiman, Random Forests, Mach. Learn. 45 (2001) 5–32. https://doi.org/10.1023/A:1010933404324.

[57] G. Louppe, Understanding Random Forests: From Theory to Practice, (2015). https://doi.org/10.48550/arXiv.1407.7502.

[58] S.K. Chandy, K. Raghavachari, MIM-ML: A Novel Quantum Chemical Fragment-Based Random Forest Model for Accurate Prediction of NMR Chemical Shifts of Nucleic Acids, J. Chem. Theory Comput. 19 (2023) 6632–6642. https://doi.org/10.1021/acs.jctc.3c00563.

[59] The Prediction of Dam Displacement Time Series Using STL, Extra-Trees, and Stacked LSTM Neural Network | IEEE Journals & Magazine | IEEE Xplore, (n.d.). https://ieeexplore.ieee.org/abstract/document/9096332 (accessed June 20, 2024).

[60] R.E. Schapire, Explaining AdaBoost, in: B. Schölkopf, Z. Luo, V. Vovk (Eds.), Empir. Inference Festschr. Honor Vladimir N Vapnik, Springer, Berlin, Heidelberg, 2013: pp. 37–52. https://doi.org/10.1007/978-3-642-41136-6_5.

[61] T. Chengsheng, L. Huacheng, X. Bing, AdaBoost typical Algorithm and its application research, MATEC Web Conf. 139 (2017) 00222. https://doi.org/10.1051/matecconf/201713900222.

[62] S.B. Kotsiantis, D. Kanellopoulos, I.D. Zaharakis, Bagged Averaging of Regression Models, in: I. Maglogiannis, K. Karpouzis, M. Bramer (Eds.), Artif. Intell. Appl. Innov., Springer US, Boston, MA, 2006: pp. 53–60. https://doi.org/10.1007/0-387-34224-9_7.

[63] J.H. Lee, Z. Shi, Z. Gao, On LASSO for predictive regression, J. Econom. 229 (2022) 322–349. https://doi.org/10.1016/j.jeconom.2021.02.002.

[64] R.J. Tibshirani, The lasso problem and uniqueness, Electron. J. Stat. 7 (2013) 1456–1490. https://doi.org/10.1214/13-EJS815.

[65] D.N. Schreiber-Gregory, Ridge Regression and multicollinearity: An in-depth review, Model Assist. Stat. Appl. 13 (2018) 359–365. https://doi.org/10.3233/MAS-180446.

[66] A. Lettink, M. Chinapaw, W.N. van Wieringen, Two-dimensional fused targeted ridge regression for health indicator prediction from accelerometer data, J. R. Stat. Soc. Ser. C Appl. Stat. 72 (2023) 1064–1078. https://doi.org/10.1093/jrsssc/qlad041.

[67] Regularization and Variable Selection Via the Elastic Net | Journal of the Royal Statistical Society Series B: Statistical Methodology | Oxford Academic, (n.d.). https://academic.oup.com/jrsssb/article/67/2/301/7109482 (accessed June 20, 2024).

[68] J.K. Tay, B. Narasimhan, T. Hastie, Elastic Net Regularization Paths for All Generalized Linear Models, J. Stat. Softw. 106 (2023) 1. https://doi.org/10.18637/jss.v106.i01.

[69] F. Murtagh, Multilayer perceptrons for classification and regression, Neurocomputing 2 (1991) 183–197. https://doi.org/10.1016/0925-2312(91)90023-5.

[70] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support Vector Regression Machines, in: Adv. Neural Inf. Process. Syst., MIT Press, 1996. https://proceedings.neurips.cc/paper_files/paper/1996/hash/d38901788c533e8286cb6400b40b386d-Abstract.html (accessed June 20, 2024).

[71] Z. Zhang, Introduction to machine learning: k-nearest neighbors, Ann. Transl. Med. 4 (2016) 218. https://doi.org/10.21037/atm.2016.03.37.

[72] K. Taunk, S. De, S. Verma, A. Swetapadma, A Brief Review of Nearest Neighbor Algorithm for Learning and Classification, in: 2019 Int. Conf. Intell. Comput. Control Syst. ICCS, 2019: pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[74] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian Optimization of Machine Learning Algorithms, in: Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2012. https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html (accessed June 20, 2024).

[75] F. Nogueira, Bayesian Optimization: Open source constrained global optimization tool for Python, (2014).

[76] S. Wang, Y. Dai, J. Shen, J. Xuan, Research on expansion and classification of imbalanced data based on SMOTE algorithm, Sci. Rep. 11 (2021) 24039. https://doi.org/10.1038/s41598-021-03430-5.

[77] B. Krawczyk, Learning from imbalanced data: open challenges and future directions, Prog. Artif. Intell. 5 (2016) 221–232. https://doi.org/10.1007/s13748-016-0094-0.

[78] S. Geman, E. Bienenstock, R. Doursat, Neural Networks and the Bias/Variance Dilemma, Neural Comput. 4 (1992) 1–58. https://doi.org/10.1162/neco.1992.4.1.1.

[79] X. Guan, H. Burton, Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications, Structures 46 (2022) 17–30. https://doi.org/10.1016/j.istruc.2022.10.004.

[80] M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, Proc. Natl. Acad. Sci. 116 (2019) 15849–15854. https://doi.org/10.1073/pnas.1903070116.

[81] P. Chakraborty, S.S. Rafiammal, C. Tharini, D.N. Jamal, Influence of Bias and Variance in Selection of Machine Learning Classifiers for Biomedical Applications, in: R. Asokan, D.P. Ruiz, Z.A. Baig, S. Piramuthu (Eds.), Smart Data Intell., Springer Nature, Singapore, 2022: pp. 459–472. https://doi.org/10.1007/978-981-19-3311-0_39.

[82] P. Mehta, M. Bukov, C.-H. Wang, A.G.R. Day, C. Richardson, C.K. Fisher, D.J. Schwab, A high-bias, low-variance introduction to Machine Learning for physicists, Phys. Rep. 810 (2019) 1–124. https://doi.org/10.1016/j.physrep.2019.03.001.

[83] X. Ying, An Overview of Overfitting and its Solutions, J. Phys. Conf. Ser. 1168 (2019) 022022. https://doi.org/10.1088/1742-6596/1168/2/022022.

[84] J. Westermayr, P. Marquetand, Machine Learning for Electronically Excited States of

Molecules, Chem. Rev. 121 (2021) 9873–9926. https://doi.org/10.1021/acs.chemrev.0c00749.

[85] B. Kalita, L. Li, R.J. McCarty, K. Burke, Learning to Approximate Density Functionals, Acc. Chem. Res. 54 (2021) 818–826. https://doi.org/10.1021/acs.accounts.0c00742.

[86] M. Bursch, J.-M. Mewes, A. Hansen, S. Grimme, Best-Practice DFT Protocols for Basic Molecular Computational Chemistry**, Angew. Chem. 134 (2022) e202205735. https://doi.org/10.1002/ange.202205735.

[87] D. Jha, V. Gupta, W. Liao, A. Choudhary, A. Agrawal, Moving closer to experimental level materials property prediction using AI, Sci. Rep. 12 (2022) 11953. https://doi.org/10.1038/s41598-022-15816-0.

[88] D. Xin, C.A. Sader, U. Fischer, K. Wagner, P.-J. Jones, M. Xing, K.R. Fandrick, N.C. Gonnella, Systematic investigation of DFT-GIAO 15N NMR chemical shift prediction using B3LYP/cc-pVDZ: application to studies of regioisomers, tautomers, protonation states and N-oxides, Org. Biomol. Chem. 15 (2017) 928–936. https://doi.org/10.1039/C6OB02450E.

[89] C. Saunders, M.B. Khaled, J.D.I. Weaver, D.J. Tantillo, Prediction of ^{19}F NMR Chemical Shifts for Fluorinated Aromatic Compounds, J. Org. Chem. 83 (2018) 3220–3225. https://doi.org/10.1021/acs.joc.8b00104.

[90] W.C.I. Isley, A.K. Urick, W.C.K. Pomerantz, C.J. Cramer, Prediction of ^{19}F NMR Chemical Shifts in Labeled Proteins: Computational Protocol and Case Study, Mol. Pharm. 13 (2016) 2376–2386. https://doi.org/10.1021/acs.molpharmaceut.6b00137.

[91] J. Huang, J. Xue, M. Li, Y. Cheng, Z. Lai, J. Hu, F. Zhou, N. Qu, Y. Liu, J. Zhu, Exploration of Solid Solutions and the Strengthening of Aluminum Substrates by Alloying Atoms: Machine Learning Accelerated Density Functional Theory Calculations, Materials 16 (2023) 6757. https://doi.org/10.3390/ma16206757.