Density Functional Theory and Machine Learning as Tools to Study Fluorine Containing Molecules

Tom Watts

Wolfson College

University of Oxford

A thesis presented for the degree of

Doctor of Philosophy

Trinity 2024

For my gran, who began in the workhouse 101 years ago who could not have imagined seeing me finishing a DPhil in Oxford. For my grandad, who would have been so proud to see me graduate after all this time.

Abstract

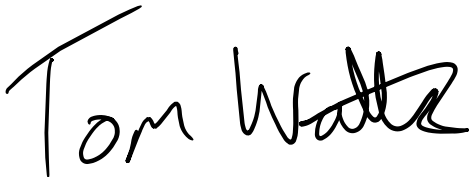
Computational organic chemistry has grown vastly in the last two decades, with the development of larger and more powerful models and more accessible experimental datasets. Advances in both Density Functional Theory (DFT) and Machine Learning (ML) have resulted in the ability to predict reaction selectivities and a range of molecular properties with ever-increasing levels of accuracy.

This thesis focuses on different DFT and ML approaches to predict enantioselectivity, NMR shifts and coupling constants for a range of fluorine-containing small molecules. In Chapter 3, both DFT and ML are used to predict the enantioselectivity of the Hydrogen Bonding Phase Transfer Catalysis (HBPTC) reaction. We then interrogate the model's learnt parameters to suggest a selection of potential catalysts and substrates that maximise the enantioselectivity for the synthesis of alkyl beta-fluoroamines. Chapter 4 further explores the study of the HBPTC complexes by developing an approach to calculate the ^{19}F NMR shifts of the fluoride anion and the $^{1}J_{HF}$ coupling constants across the H-F hydrogen bond. Finally, in Chapter 5, a BERT model for the prediction of ^{19}F NMR in a range of organic molecules is introduced. Integrated Gradients (IGs) are then used to understand if the model has learnt chemistry. Finally, the model is applied to identify which regioisomer is produced from late-stage fluorination reactions purely from ^{19}F NMR.

This thesis shows how a range of different ML and DFT options can be utilised by computational chemists to solve a series of challenging chemical problems, while also giving direction for future research in the field.

Declaration

I, Tom Watts, declare that the Thesis I am submitting is entirely my own work except were indicated in the text, caption, footnote or bibliography. This Thesis has not been submitted in whole or in part for any other academic degree or professional qualification.



Signature

14th May 2025

Date

Acknowledgments

It has taken me a very long time to reach this point over the course of the last seven years, and therefore I have quite a long list of people who have helped and supported me so this will take a while.

My first set of thanks must go out to my supervisors, both Ben and Veronique who let me propose an idea to work on machine learning for catalysis after my rotations and were more than happy to let me pursue the idea. Fernanda, after joining the department was asked to also join the project and was more than willing to help.

The work in this thesis has been influenced by many people who I will try and thank but apologises if I missed anyone. David who looked after me while I initially started to do computational chemistry and would happily answer my questions, no matter how stupid. Francesco and Anna, whose experimental knowledge, and willingness to collaborate and share their thoughts and ideas were key in the development of all my work on the HBPTC catalysts from the VG group. Matina, I must thank you many times for your support and collaboration on the MLHBPTC work, not only did you teach me more about coding and ML, but you were always willing to listen to some of my crazier ideas. As part of the original ML subgroup with me and Matina, Chloe has been a key person to bounce ideas off and suggest different ML approaches for models and is always willing to help with any issues I come across. Finally, to my self-appointed “daughter” Sara, thank you for always being uplifting and supportive when I rant about my work. I also need to thank, Matina, Chloe, Sara, Ewa, Bastian, Tomasz, and Aleksy for proofreading this thesis.

From my time in the VG group, thanks need to be given to Rob, Jimmy, Gulia, Jeron, Joe, and Nathan all of whom welcomed me throughout my rotation and through the first part of my DPhil. Similarly, those in the BGD group who helped me settle in and generated a positive atmosphere to work in; Abarami, Andrew, Adaline, Patrick, Brian, Geroge, Tobi, Alex and

Simon all were a part of that. From the FD group, a special thanks needs to go to Alister, Hanwen, Ally, Tomasz, Aleksy and Henry.

Alongside my research, a large part of the last two years has been my teaching in the CTL, both for enjoyment and maintaining my mental health. For this I wish to thank those other demonstrators who have been an integral part of my teaching experience, Anabell, Ryan, Amy, Diana, Nicoletta, Andrew and all the Matts. To all the technicians making the jobs of the demonstrators easier, Alissa, Ben, Prab, Ollie, Sami, Tom, Louise you have made my teaching time so much more fun. Special thanks go to Jennie for always being someone I can ask for support, or rant about silly things students have done. To Max for always matching my sarcasm, Shaun for teaching me more about inorganic, and answering my dumb questions, and Hannah for teaching me more of Oxfords quirks, you have all been invaluable. Alison, you have always kept me upbeat and have all the (medical) drugs I would ever need! For the teaching staff in the first-year labs, thank you for giving me the chance to teach a new year, and develop my teaching skills, Malcom, Craig, Lucy and Zoe my greatest thanks. For the staff in the second and third-year lab, Rebecca, Ryan, Sam and Megan, you have always supported me this year and provided endless conversations about the latest student shenanigans. A special shout-out must go to Mark, who enjoyed too much trying to break my brain with the actual physical chemistry explanations, rather than the handwavy organic ones I have always used. The final person here who I am left to thank, is the one who has been the most influential on me over the past two years. Andrew has been the ideal mentor for me to further develop my teaching skills and fall back in love with chemistry. Not only was any topic not off-limit, but you were supportive and nurturing of my skills. I cannot thank you enough.

As most people who know me are aware, my mental health has been something that has been a struggle over the whole of my PhD, there has been a lot of support from my GPs over the years which I am completely grateful for, alongside the support of my mentor Julia. The most

helpful, long-term, support I have received has been through groups I have attended through Oxfordshire Mind. All the facilitators and people who attended the groups, especially those in the peer support group, who listened, provided support and hours of laughter I am immensely grateful to all of you.

I would also like to thank all my friends who have supported me over the last few years, not limited to Alex, Amelia, Matt, Dori, Niamh, Maria and Meher. Several people have yet to be thanked but without them and are the most important. Firstly, to Sarah for listening, helping me write, getting me out of my room and being a caring and considerate friend during my worst periods. Not only did you continue to inspire me to work further on science but our conversations on academic Twitter were always a highlight of the day. To Amy, you have always kept me grounded over these past two years, ensuring that I did not forget about the progress and the people supporting me. Plus, who else is going to send me pictures of all the different drinks! The final person on this list is Ewa, you have supported me through ups and downs, letting me rant and emotionally dump all my problems on you, eyerolls notwithstanding. Your patience, constant faces and kindness have been tremendously important for my mental health. My apologies for all the anxious messages you have got, but on the plus side you can now travel abroad without them!

Before I get to the final two people, I would like to thank several people who are not able to see this. To my favourite Auntie Carole, to my champagne-loving grandad and my caring and loving gran, I wish you all were here to celebrate this with me. Finally, to my parents, both of you two have stood by me, supported me when I was down and never given up on me. Your love, support and kindness have been unwavering over the past seven years, and your willingness to listen and advise me on any issue has been invaluable. There is no possible world, where I would be here in this situation if it wasn't for the help, you both have provided over all this time.

Data availability

The data presented in this Thesis, including calculated energies, NMR calculations, Cartesian coordinates, PyTorch models, Python scripts for calculations and data processing have been submitted as Supplementary Material. This data is publicly available on the Oxford University Research Archive (ORA) at https://doi.org/10.5287/ora-z6nq99nqe.

List of Publications

The following manuscripts were in preparation at the time of completion of this degree:

1. Watts, T.; Davis, B. G.; Gouverneur, V.; Duarte, F.* Translating Chemical Structures: Using NMRBERT to Predict ^{19}F NMR Shifts. Manuscript in preparation

Contributing-author articles published or submitted for publication during this degree that do not form part of this Thesis are as follows:

2. Liu, W.[#]; Faulkner, S.[#]; Couturier, A. M.[†]; Watts, T.[†]; Josepheson, B.; Bedwell, A.; Duarte, F.; Davis, B. G.*; Esashi, F.* Intact nucleosomal context enables chromodomain reader MRG15 to distinguish H3K36-me3 from -me2. Manuscript Submitted. A previous version is available as a preprint https://doi.org/10.1101/2020.04.30.070136. [#] and [†]indicate equal contributions.

List of Abbreviations

AEV	Atomic Environment Vector
BERT	Bidirectional Encoder Representations from Transformers
BINAM	1,1′-Binaphthyl-2,2′-diamine
BO	Bond Order
CPCM	Conductor-like Polarizable Continuum Model
CREST	Conformer–Rotamer Sampling Tool
CV	Cross-Validation
DFT	Density Functional Theory
DFTB	Density Functional Tight Binding Theory
DH-DFT	Double Hybrid Density Functional Theory
DNN	Deep Neural Networks
DSO	Diamagnetic Spin-Orbit
e.r	Enantiomeric Ratio
ee	Enantiomeric Excess
EOM-CC	Equation of Motion Coupled Cluster
FC	Fermi Contact
GGA	Generalised Gradient Approximation
GIAO	Gauge Independent Atomic Orbitals
GNN	Graph Neural Networks
HB	Hydrogen Bond
HBC	Hydrogen Bonding Catalysis
HBD	Hydrogen Bond Donors
HBPTC	Hydrogen Bonding Phase Transfer Catalysis
HF	Hartree-Fock
HOMO	Highest Occupied Molecular Orbital
HOESY	Heteronuclear Overhauser spectroscopy
IG	Integrated Gradients
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Local Density Approximation
LIME	Local Interpretable Model-agnostic Explanations
LUMO	Lowest Unoccupied Molecular Orbital
MAE	Mean Absolute Error
MCSCF	Multiconfiguration Self-Consistent Field
MD	Molecular Dynamics
ML	Machine Learning
MLR	Multivariate Linear Regression
MPNN	Message Passing Neural Network
MSE	Mean Square Error
MT	Molecular Transformer
MUF	Mono Urea Fluoride
NHK	Nozaki–Hiyama–Kishi

NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
NN	Neural Network
PES	Potential Energy Surface
PSO	Paramagnetic Spin-Orbit
PTC	Phase Transfer Catalysis
QM	Quantum Mechanics
QSAR	Quantitative Structure Activity Relationship
RDF	Radial Distribution Function
RI	Resolution of Identity Approximation
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
RxnBERT	Reaction Bidirectional Encoder Representations from Transformers
SA	Simulated Annealing
SCF	Self-Consistent Field
SD	Spin Dipole
SHAP	Shapley Additive Explanations
SMILES	Simplified Molecular Input Line Entry System
SOAP	Smooth Overlap of Atomic Positions
SP	Single Point
SSR	Sum of Squared Residuals
SVM	Support Vector Machines
TBA	Tetrabutylammonium
TFA	Trifluoroacetic acid
TMA	Tetramethylammonium
TMAF	Tetramethylammonium fluoride
TMS	Tetramethylsilane
TS	Transition State
UFF	Universal Forcefield
WFT	Wave Functional Theory
ZORA	Zero Order Regular Approximation

Table of Contents

Chapter 1 Overview, Background and Theory for DFT in Organic Chemistry

1.1 Overview

Computational chemistry can be of great aid to experimental chemists, through the identification of reaction paths and transition states and by rationalising selectivity. In particular, computational chemists can work in collaboration with experimentalists to identify and improve enantioselectivity through modelling transition states and suggesting modifications to better differentiate between major and minor transition states.[1-8] Alongside work on aiding reaction development and optimisation, computational chemists can help in the identification of reaction products through the calculation of experimentally measurable properties, including NMR, allowing for the differentiation between diastereomers or reassignment of natural products.[9]

This thesis covers the study of fluorine containing compounds through the use of both traditional DFT-based methods alongside more modern machine learning concepts. Overall, the aim of the work described in the following chapters is to develop workflows and methods that can help experimental chemists to rationalise both selectivity and structure of ^{19}F-containing compounds. Chapters 3 and 4 focus on hydrogen bonded fluoride complexes, while Chapter 5 is broader covering most organofluorine compounds.

Chapter 1 starts with an overview of fluorine-containing molecules and their importance in organic chemistry. This is followed by a discussion on hydrogen bonding phase transfer catalysis before an overview of the prediction of ^{1}H, ^{13}C and ^{19}F NMR using computational means is described. The theory and background of Computational Chemistry, including wave

function theory and density functional theory, are introduced, along with the calculation of molecular properties.

Chapter 2 introduces machine learning techniques to study organic reactivity and properties, looking at the prediction of yield, enantioselectivity and NMR shifts. Before the background theory underpinning machine learning algorithms, data selection and overfitting are discussed.

In Chapter 3 focuses on the development of a workflow to aid in the enantioselective synthesis of trans-β-fluoroamines which is a key target motif for medicinal chemistry and drug discovery. A data mining and machine learning approach was used, utilising both published and unpublished experimental data obtained by the Gouverneur group. This model was trained to predict the enantioselectivity for a range of epi-sulfonium and aziridinium substrates achieving accuracies below 1.1 kJ/mol. By interrogating this model, we were able to generate a linear energy relationship between substrate, catalyst structure, and enantioselectivity, before computationally screening catalysts to identify future potential synthetic targets.

In Chapter 4 we further studied the HBPTC catalysts in solution with the goal to develop a computational model that could predict $^{1}J_{HF}$ coupling constants and ^{19}F NMR shifts. This project required us to first develop a computational workflow and methodology that could both sample the conformational space of the catalyst complex in solution before identifying an appropriate level of DFT theory that could reproduce experimental NMR shifts and coupling constants.

In Chapter 5 we aimed to develop a general machine learning model for the prediction of ^{19}F NMR shifts in a range of fluoro-containing organic molecules. Our approach is based on the RxnBERT (Reaction Bidirectional Encoder Representations from Transformers)[10] neural network architecture, originally developed to predict reaction classes and later adapted to predict reaction yields from SMILES only. We envisaged that this architecture could be adapted to the prediction of NMR shifts, focusing on the prediction of ^{19}F NMR shifts, which

currently have not been reported in the literature. We demonstrate that the model can both predict the ^{19}F NMR shift with low errors and show the model is learning the subtle balancing of electron-donating and withdrawing effects within its prediction. The model is then applied to the identification of regioisomers in late-stage fluorination reactions.

1.2 Fluorine-containing molecules in organic chemistry

1.2.1 Importance of fluorine containing molecules

Fluorine-containing molecules make up a large proportion of both pharmaceuticals and agrochemicals. For the pharmaceutical industry, 18% of all pharmaceuticals and 22% of small-molecule drugs contain at least one fluorine atom,[11] while the agrochemical industry shows a much greater coverage with 54% of chemicals containing at least one fluorine atom.[12] The fact that the C-F bond is one of the strongest bonds in organic chemistry (a bond dissociation energy of 105.4 kcal $mol^{-1,}$)[13] has resulted in it being widely used to replace C-H bonds, and increase the metabolic stability and lifetime of drugs in the body.[14] Fluorine, with its high electronegativity, results in the C-F bond being highly polarised.[13] This polarisation can be harnessed to modulate both the pK_a and the lipophilicities of a drug molecule, which can benefit the potency and uptake of pharmaceuticals.[14] The experimental study and discovery of new methods to synthesise fluorine-containing compounds in a selective and specific manner is therefore of great importance to both academia and industry. Methods that either result in new fluorine-containing building blocks, enantioenriched products, or late-stage mild fluorination methods, are of intense study and interest, especially if those methods are also sustainable and require fewer toxic reagents.[15-25]

Alongside the development of new and novel synthesis applications, the analysis and identification of reaction products are also key in understanding chemical reactivity. One of the major methods for the analysis of fluoro-containing molecules is nuclear magnetic resonance (NMR). ^{19}F is the only naturally occurring isotope of fluorine and has a spin = ½ meaning the nucleus is NMR active. ^{19}F also has a high gyromagnetic ratio, similar to that of ^{1}H,[26] meaning that ^{19}F NMR therefore is both sensitive and produces clear and narrow peaks, resulting in easy-to-identify spectra.[27]

1.2.2 Development of Hydrogen Bonding Phase Transfer Catalysis

In 2018, the Gouverneur group developed an elegant urea-catalysed enantioselective nucleophilic fluorination reaction with metal alkali salts, also known as Hydrogen Bonding Phase Transfer Catalysis (HBPTC).[16] This method, inspired by the fluorinase enzyme,[28, 29] uses an insoluble CsF salt which is brought into solution via the use of a chiral BINAM derived catalyst. This work combined two previously widely used concept in organocatalysis: firstly Hydrogen Bonding Catalysis (HBC), where hydrogen bonds are used to deliver a reactant to a starting material in a enantioselective manner.[30-32] The second concept is Phase Transfer Catalysis (PTC), where organocatalysts move between phases, bringing reactants together and resulting in an enantioselective reaction.[33-37] In HBPTC the fluoride, once brought into solution, is in a chiral environment and so can be added to a range of electrophiles in an enantioselective transformation.

The initial discovery was that Schreiner′s Urea (Figure 1A, **1**) could catalyse the addition of fluoride to an in-situ generated *epi*-sulfonium intermediate (Figure 1A). After a large amount of experimental screening, catalyst design, and optimisation to identify an active chiral backbone, the BINAM based catalyst **2a** was identified as a promising candidate. Alongside experimental work Density Functional Theory (DFT) and Molecular Dynamics (MD) simulations were crucial in understanding the underlying reaction pathway along with the key interactions which resulted in a high enantioselectivity. A key computational insight was discovered when using MD to investigate the binding modes of the catalyst to the CsF. When the catalyst is non alkylated R = H (Figure 1A, **2a**), the urea can interchange between the *anti-anti* form and the *anti-syn* conformer (Figure 1C). DFT calculations performed on the two possible conformers suggested that the tri dentate binding (*anti-syn* conformer, **2b**) would be more energetically favourable. When alkylation of the catalyst was carried out experimentally, followed by a series of further optimisations, the enantiomeric ratio increased from 82:18 to

95.5:4.5. To experimentally determine the binding mode of the tridentate catalyst **2b**, the complex with a TBA cation was successfully crystallised and characterised by its X-ray structure, showing the three hydrogen bonds to fluoride.

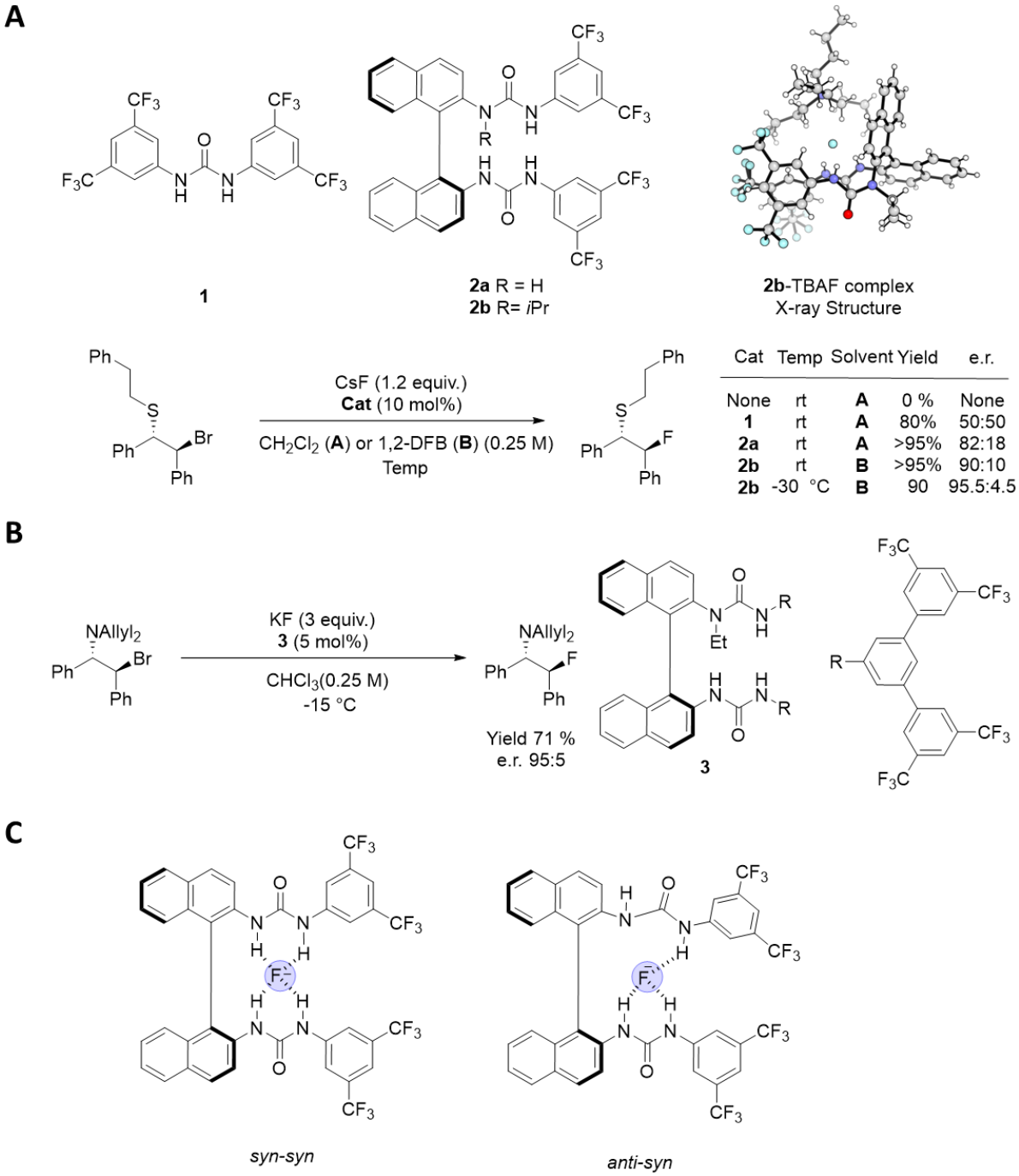


Figure 1A) Overview of the discovery of the HBPTC Reaction developed by the Gouverneur group[16] along with selected catalysts, yields and e.r's. DFB = difluorobenzene B) Extension to the synthesis of beta fluoroamines using HBPTC[38] C*) syn-syn* and *anti-syn* conformer of catalyst **2a**, *anti-syn* conformer showed stronger bonding to fluoride,

To identify the transition state (TS) of the reaction and the major contributions to a high differentiation between the energy of the major and minor TS's, DFT calculations were then carried out using optimisations carried out with the CPCM(CH_2Cl_2)-M062X/def2-SVP(TZVPPD) level of theory, before free energy calculations were performed using the COSMO(CH_2Cl_2)-ωB97X-D3/ma-def2-TZVPP level of theory. Two major noncovalent interactions were identified from the DFT calculations: firstly, a cation-π interaction between the substrate and the BINAM backbone of the catalyst. The other major noncovalent interaction is a CH-π interaction between the hydrogens on the aromatic R1 groups of the catalyst and the stilbene aromatic groups. While these interactions are present in both the major and minor TS, there are shorter distances in the major TS thus in line with preferential binding and a lower TS. Furthermore, only in the major TS can the phenyl rings from the stilbene remain conjugated to the reaction centre which results in a lower energy conformer than in the minor TS. Overall, a combination of these factors results in an excellent differentiation between the major and minor TS, resulting in a high enantioselectivity. This work was extended to the addition of fluoride to aziridinium intermediates using both KF and CsF salts (Figure 1B).[38] Further catalyst optimisation was required, with larger R groups required to reach the high enantioselectivities (**3**).

1.2.3 Studying the catalyst in solution

To understand how the catalyst behaves in solution, the Gouverneur group then performed a series of NMR studies to identify how the fluoride sits in the catalyst complex.[39] In particular they were able to measure both the ^{1}H and ^{19}F NMR shifts of a range of urea- fluoride complexes along with the $^{1}J_{HF}$ coupling constants over the H-F hydrogen bonds (Figure 2A). This allowed for an identification of where the anion was positioned in regard to all the hydrogens in the catalyst along with the strength of each hydrogen bond (HB), giving insight into the resting state of the catalyst in solution. NMR titrations were used to determine the

binding constant for the catalyst to fluoride. This study also identified that the non-alkylated catalyst forms two strong hydrogen bonds to the fluoride, leading to efficient reactivity but low enantioselectivity. However, when the catalyst is alkylated a trifurcated hydrogen-bonded fluoride complex is formed which is rigid, efficient for phase transfer and capable of high enantioselectivities. Using quantitative ^{1}H-^{19}F HOSEY experiments the length of each H-F HBs were able to be experimentally determined, showing that the structures in solution are similar to the crystal structures obtained previously (Figure 2B). From NMR experiments they were also able to identify the presence of a small amount of a catalyst's dimer complex with fluoride, a possible off cycle resting state for the catalytic cycle.

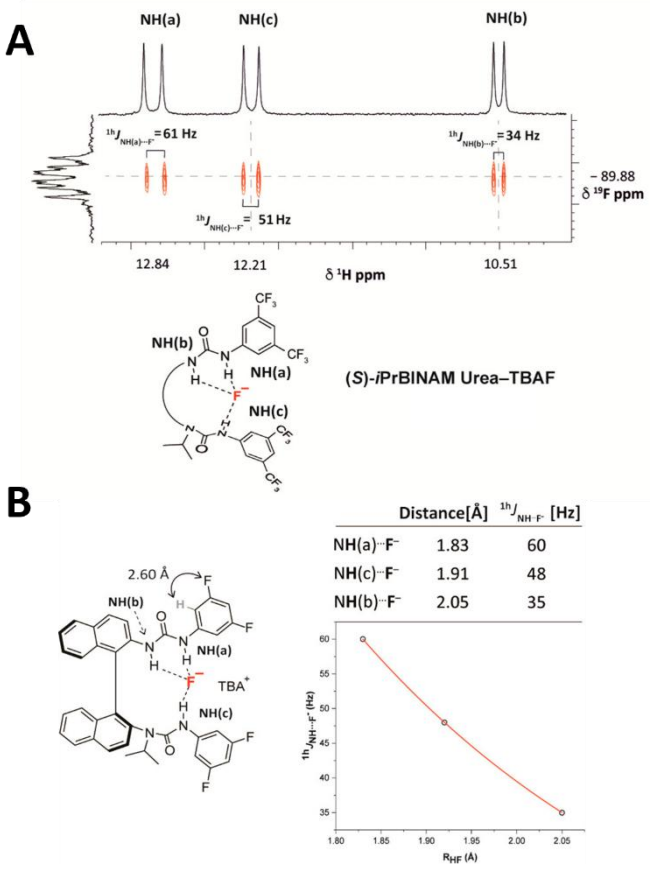


Figure 2 A) Measurement of $^{1}J_{HF}$ coupling constants from the Gouverneur group on catalyst **2a**. B) Measurement and determination of H-F hydrogen bond lengths using quantitative ^{1}H-^{19}F HOSEY experiments. Both images reproduced from Ibba *et al*.[39]

Using 13 variations of the catalyst Ibba *et al.*[39] were able to alter the electronics of the urea catalyst, resulting in the ability to tune the strength of each hydrogen bond and alter the position that the fluoride occupies in the complex. By studying the efficacy of each catalyst, a structure-activity relationship was able to be identified which showed that the contributions of each hydrogen bond to the overall selectivity was not uniform, implying that tuning the strength of individual HBs is important for overall catalyst selectivity.

1.3 Computational chemistry for NMR prediction

1.3.1 ^{1}H and ^{13}C NMR calculations

A large body of work has focused on the accurate prediction of ^{1}H and ^{13}C NMR shielding constants to aid in the structural elucidation of unknown compounds. Large databases of ^{1}H and ^{13}C experimental NMR data have been assembled to benchmark and identify the best computational methods available. From these studies the functionals PBE0, B3LYP, and mPW1PW have been identified as accurately calculating experimental NMR shifts.[9, 40-45] To aid in the prediction of NMR shifts a range of computational workflows such as those from the groups of Goodman[46-49] and Grimme[50, 51] have been developed. These workflows use forcefields or semi-empirical methods to generate conformational ensembles before DFT optimisations and the calculation of the isotropic shielding constants are performed. The NMR shifts from all the conformers are usually then Boltzmann weighted to generate weighted ^{1}H and ^{13}C NMR shifts. Overall the DFT based methods are able to achieve small error for both ^{13}C and ^{1}H NMR with Root Mean Square Errors (RMSEs) of <3 ppm and <0.2 ppm respectively,[40, 49, 52] which corresponds to < 2% error over the range of experimental scale for both ^{1}H and ^{13}C predictions.

An important use of the computational prediction of ^{1}H and ^{13}C NMR is for the elucidation of complex molecules and natural products, in particular where a large number of possible diastereomers could be present. Through combinations of conformational sampling followed by a comparison between predicted values and experimental, a complete assignment of complex molecules including the absolute stereochemistry of complex ring systems is possible.[48, 53, 54] Pioneered by the Goodman group, DP4[48] used Bayesian statistical methods to assign a probability to a selection of different possible structures given reference experimental data. By taking into account computed ^{1}H and ^{13}C NMR shifts, along with the known computational errors, the model returns a probability for each potential structure. In 2022 the Goodman group extended this further with DP5[47] which uses machine learning techniques alongside ^{13}C NMR calculations to predict the probability that the structure is correct based on the experimental data.

1.3.2 ^{19}F NMR prediction

The calculation of ^{19}F NMR spectra has been much more limited than that of ^{1}H and ^{13}C due to a lack of accurate and reliable experimental data. Recently this lack of reproducible ^{19}F NMR data was noted in a seminal study by Togni.[55] The authors note that multiple measurements of the same sample by separate groups can show errors of >1 ppm depending on the lab and machine performing the measurement. Such significant levels of error means that large scale collection of ^{19}F NMR data could result in an intrinsic error which computational calculations would not be able to quantify.

Bagno *et al.*[56, 57] computed ^{19}F NMR signals on a wide range of fluorinated compounds. In their first contribution,[56] they studied a wide range of small organic and inorganic fluoride compounds spanning the entire range of known ^{19}F NMR shifts (ca 1300 ppm, from ClF to FOOF) and so allowed for a broad investigation into the computational methods needed to calculate ^{19}F NMR accurately. They compared a range of DFT functionals from simple GGA

functionals (BP86) to the more complex meta-hybrid functionals (M06). They also performed selected calculations using wave function theory (MP2), but this was only used as a comparative benchmark for a select number of compounds. Relativistic effects were also included using zero order regular approximation (ZORA), given the presence of heavy metals in the dataset. The best performing functional was ZORA-PBE0/TZ2P (Mean Absolute Error (MAE) = 21 ppm: 2% error). The overall differences between all the methods are small with all showing an MAE between 20-40 ppm (2-3% error over the range). However, as noted in the paper, "the overall accuracy values remain significantly lower and may compromise applicability to structural problems".

Table 1 Overview of prediction of ^{19}F NMR shifts. "Method" describes NMR calculation level, "Solvation" indicates if implicit solvation was used during calculations, "Conformational Sampling" indicates whether any sampling was used (sometimes is used when only certain flexible molecules were sampled).

Paper	Method	Solvation	Conformational Sampling	Range (ppm)	Max (ppm)	Min (ppm)	Data Points	Error (ppm)
Bagno *et al*	ZORA-PBE0/TZ2P	X	X	1273	825	-448	75	21 (MAE)
Bagno *et al*	ZORA-BLYP/TZ2P	X	X	295.6	18.6	-277	299	6.2 (MAE)
Tantillo *et al*	B3LYP/6-31+G(d,p)	X	Y	153	-43	-196	100	2.1 (MAE)
Dumon *et al*	ωB97X-D/aug-cc-pVDZ	Y	Sometimes	289.1	16.7	-272.4	83	3.57 (RMSE)
Gerken *et al*	MP2/6-31++G(d, p)	X	X	74	-74	-148	8	10 (MAE)

Bagno *et al.*[57] then followed up on their previous publication by focusing solely on the prediction of ^{19}F NMR for larger organic molecules. ZORA-BLYP/TZ2P was used to predict ^{19}F NMR shifts over the range of ca 300 ppm (TMS-CH_2F -277 ppm to CF_2I_2 18.6 ppm). This dataset includes alkyl monofluorides (primary, secondary, and tertiary), alkyl difluorides (primary and secondary), trifluoromethyl derivatives, and fluorobenzene derivatives. Overall, this approach had an MAE of 35 ppm which was corrected to 6.2 ppm using linear fitting. Conformationally flexible compounds exhibited had the largest error, likely due to a lack of conformational sampling.

Tantillo *et al.*[58] reported a method for the calculation of ^{19}F NMR shifts for a range of fluorinated heterocycles at the B3LYP/6-31+G(d,p)//B3LYP/6-31+G(d,p) level of theory. Covering a range of 152 ppm and a range of mono and multi-fluorinated heterocycles, their approach gave an MAE of 2.1 ppm for the corrected ^{19}F NMR shifts. The inclusion of implicit solvation in the NMR calculation had a negligible impact on the overall accuracy of the model, however they did include conformational sampling, and the predicted values for each compound were weighted according to a Boltzmann distribution of the conformers.

Recently Dumon *et al.*[59] used ωB97X-D/aug-cc-pVDZ to calculate the ^{19}F isotropic shielding constants for a range of small organic molecules, including some boron-fluorine bonds. This method achieves a fitted RMSE of 3.57 ppm. This model also extends well to cationic structures where their error is no larger than the methods RMSE, however, anionic structures perform less well, with some errors larger than 10 ppm. The explanation for this underperformance with anionic species is the lack of the inclusion of the counterion in the structure, which while difficult to model computationally could have a significant effect on the electron density around the anion. For molecules which were conformationally flexible, the ^{19}F NMR shifts were weighted according to a Boltzmann distribution of the conformers.

While the study of ^{19}F NMR has recently become an area of more active research, there is limited work on the study of the fluoride anion in solution and how hydrogen bonding and coordination affect the ^{19}F NMR shift. Gerken *et al.*[60] reported a protocol for the calculation of ^{19}F NMR shifts of tetramethylammonium fluoride (TMAF) complexes in a range of solvents.

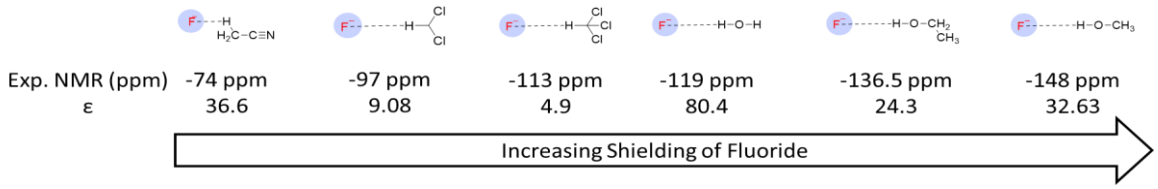


Figure 3 Experimental ^{19}F data of TMAF complexes in a range of solvents, along with the dielectric constant (ε) which not correlate with the NMR shift

Using an MP2/6-31++G(d, p) methodology they generated complexes of one solvent molecule with a fluoride anion before optimising and calculating the ^{19}F NMR with an accuracy of 10 ppm after linear fitting. An important conclusion was that the ^{19}F NMR shielding constant does not correlate with the dielectric constant of the solvent, but that the mono solvation of fluoride gave results which would correlate with experimental data (Figure 3).

1.3.3 Computational calculation of $^{1}J_{X\text{-}H}$ coupling constants

NMR prediction of spin-spin coupling has been instrumental in the analysis and assignment of unknown molecules. For example, $^{1}J_{CH}$ $^{3}J_{HH}$ and $^{4}J_{HH}$ couplings have been implemented in a range of methods for the stereochemical assignment of natural products.[52, 61-66] These methods can predict the coupling constants with errors of < 1 Hz. However, heteroatom-hydrogen coupling has been much less studied.

The study of $^{1}J_{HF}$ coupling has focused on the calculation of coupling constants within HF clusters. Work initially carried out by the Limbach group[67] experimentally determined the $^{1}J_{HF}$ coupling constants for FHF^-, $F(HF)_2^-$, $F(HF)_3^-$ and $F(HF)_4^-$ in solution (Figure 4A). Using MP2/6+31+G(p,d) to obtain geometries before using the PW86-P86 functional to calculate NMR shielding constants and coupling constants. While this method does reproduce the shielding constants for the proton spectra, it is unable to predict the coupling constants across the hydrogen bonds, with errors of > 50 Hz. Complete active space calculations did improve the predicted $^{1}J_{HF}$ for the FHF cluster, however the expense of such a calculation prohibited further study.

The extensive work by Bartlett *et al.* on the use of equation of motion coupled cluster (EOM-CC) calculations[68, 69] resulted in a methodology for the calculation of both NMR shielding constants and coupling constants.[70] They expanded this work onto the calculation of $^{1}J_{HF}$ couplings to predict the values obtained by Limbach.[71] Geometry calculations were carried out at the CCSD(T)/aug-cc-pVDZ level before the spin-spin coupling calculations were carried out

using EOM-CCSD/qz2p. This method was able to predict the experimental $^{1}J_{HF}$ coupling constants within 30 Hz of their experimental values and with the correct sign (Figure 4B).

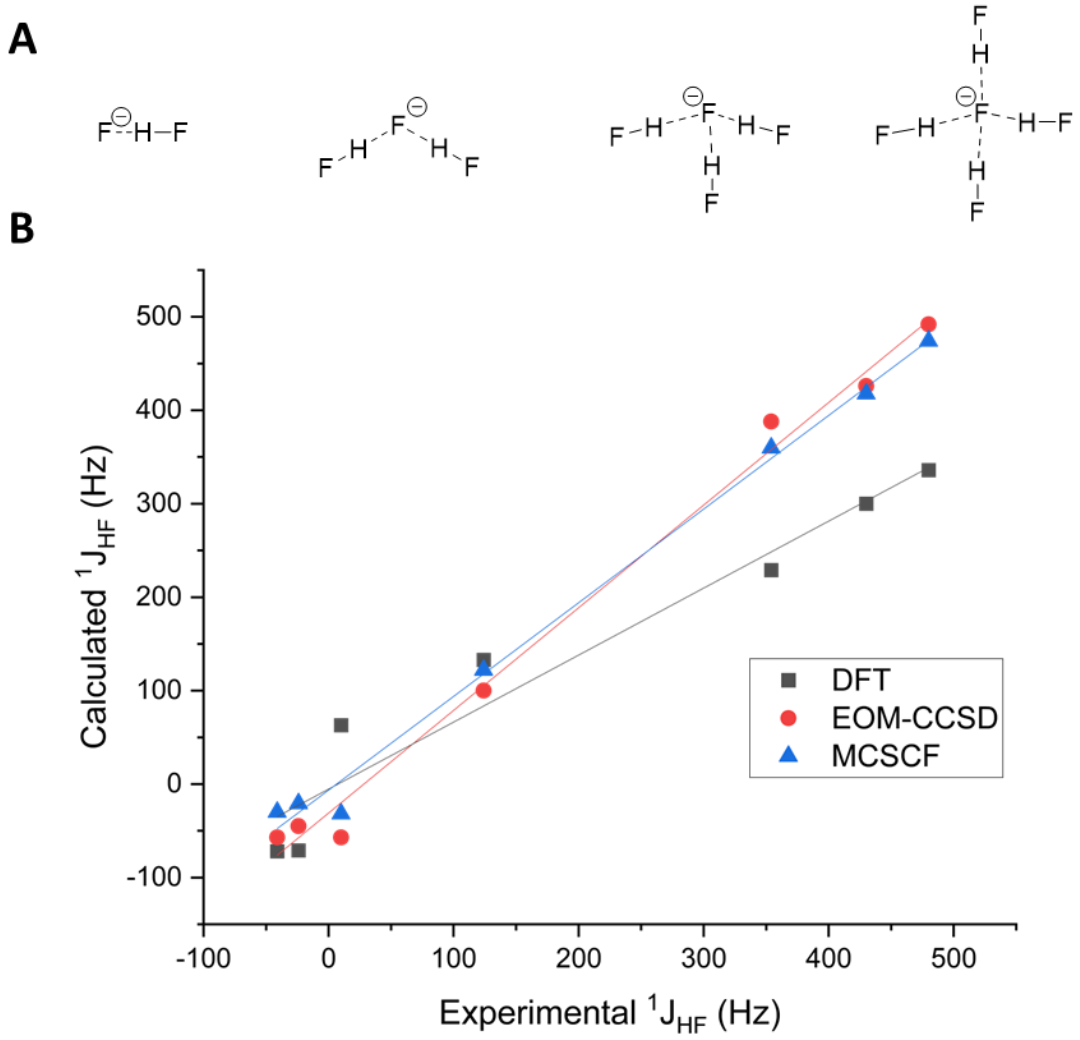


Figure 4 A) Four hydrogen bonded complexes identified by Limback *et al.*[67] B) Example data of experimental vs computed $^{1}J_{HF}$ coupling constants. DFT is work by Limbach *et al.*[67], EOM-CCSD is work by Bartlett *et al.*[71] and MCSCF is work by Leszczynski *et al.*[72]

Leszczynski *et al.* approached the problem of $^{1}J_{HF}$ couplings using a multiconfiguration self-consistent field (MCSCF) approach.[73] To achieve high accuracy they had to use the large aug-cc-pVDZ basis set. For FHF^- and $F(HF)_2^-$ this method is very accurate and within experimental error of 2 Hz, however upon moving to larger structures such as $F(HF)_3^-$ and $F(HF)_4^-$ the error

increases to 10 and 40 Hz respectively. Geometries for this set of calculations were taken from both experimentally determined distances and previous calculations.

Computation of N-H-X couplings across hydrogen bonds has also been studied using DFT methods. Alkorta *et al.*[74] studied the couplings in a series of N-H-N complexes with N donors including pyrrole, cyanide, and ammonia. The authors performed calculations on neutral, cationic, and anionic systems with both symmetrical and asymmetric systems. Geometry optimisations were carried out at the MP2/6-311++G (d, p) level of theory before NMR calculations were performed using three different methods. B3LYP was used in combination with either the 6-311++G (d, p) or aug-cc-pVTZ basis sets. Overall, all three methods performed equally well with only the cyano groups being inaccurately predicted with errors more than 20 Hz. The larger aug-cc-pVTZ basis set did reduce the overall error slightly by <5 Hz, however compared to the smaller Pople basis set this is insignificant as it only corresponds to a < 1% error reduction over the range of coupling constants measured. However, this study was only carried out on the monomers as no experimental data was available for the dimers which show the N-H-N bridge.

Studying the more complex species of an N-H-F hydrogen bond in a $[F(NH_4)_2]^+$ complex, Del Bene and Elguero[75] found that the equilibrium geometry had an error of 8 Hz compared to the similar $(collidineH:F:Hcollidine)^+$ structure. Using the optimum geometry for this system resulted in a slightly worse prediction with an error of 9 Hz. However, fixing the N-H and H-F bond distances at the values report by Limbach[76] resulted in an error of 4 Hz for these complexes. This detailed analysis of the effects of varying both the H−F and N−H distances shows that small fluctuations can change the calculated coupling constant - as the H-F bond length changes from 1.4 Å to 1.0 Å the $^1J_{HF}$ coupling constant varies from -70.5 Hz to 452.6 Hz.

A more recent paper is a collaboration between Alkorta, Elguero, and Limbach.[77] Here they studied intramolecular hydrogens bonds of the type N-H-F in fluorinated molecules that show a weak hydrogen bridge. Both optimisations and NMR calculations were carried out at the B3LYP/6-311++G** level of theory. Only one data point was used to directly study the N-H-F coupling, but this model was still accurate with the coupling constants predicted within 2 Hz of the experimentally determined value.

A large body of work by Cremer[78-82] has focused on a method for the decomposition of the NMR coupling constants into different bonding contributions called J-OC-PSP.[82] This method uses either B3LYP[81] or BLYP(60:40)[82] and a specially designed decontracted basis set to give an improved description of the core electrons. These methods have been used in a variety of systems including proteins and small H−F molecules. In all cases, an average error for the prediction of couplings constants is 23 Hz for the $^{1}J_{HF}$ coupling while the $^{1}J_{NH}$ and $^{3}J_{CN}$ coupling in proteins was found to be lower at < 5 Hz error.

1.4 NMR databases

The choice of NMR dataset is important to ensure the model is both accurate and general to many environments. Databases assembled from QM calculations should be free from misassignment errors and should contain data for every ^{1}H or ^{13}C atom, removing the need to consider symmetry during data preparation. However, the limitations of these databases are threefold: firstly, the database of compounds needs to cover chemical space well, secondly, the compounds need extensive conformational sampling to generate a Boltzmann weighted NMR prediction, and finally, the level of theory needs to be accurate to reproduce experimental NMR errors, and any error from the QM calculations needs to be considered. A QM derived databases from Glezakou[83] use compounds from the QM9 database, which only contains HCNOF atoms, and their lowest energy conformer. Another QM derived database is that from Beran,[84] where the model has an error of only 0.7 ppm compared to computed ^{13}C NMR, however that error increases to 4.8/2.3 ppm when compared to the experimental values in $CDCl_3$/DMSO. This shows that the implications of solvation are not considered in models trained on these datasets. Paton *et al.*[85] have also reported a computational database that contains 8k molecules with over 200k individual NMR shifts for both ^{1}H and ^{13}C. These 8k molecules were selected as a diverse set from the NMRShiftDB,[86] with molecules with a MW > 500 and charged species being excluded. Confirmational sampling was performed before DFT calculations were performed at the mPW1PW91/6-311+G(d,p)//M06-2X/def2-TZVP levels of theory. This generated the complete set of weighted NMR shifts for each molecule.

Experimental datasets remove the need to calculate Boltzmann weighted NMR values as they provide an ensembled-averaged values. They also consider solvent effects which not all computational databases consider. One potential downside of experimental datasets is that overlapping signals can lead to either ambiguous or incorrect assignments which are carried

over into computational predictions. Online databases such as SDBS[87] or NMRShiftDB[86] are available for a wide range of compounds with the associated data attached.

In this thesis we test both experimental and computational datasets, but our final models only use experimental data sets as the error in computational calculations can be negated and the models showed no loss in accuracy.

1.5 Background on computational chemistry

1.5.1 Computational theory

A large part of Computational Chemistry theory is heavily based upon Quantum Mechanics, which was developed throughout the 20th century. The seminal discovery was the Schrödinger equation, first published in 1926,[88] which describes how the energy of a molecule is quantised and can be obtained by using the Hamiltonian operator $\boldsymbol{H}$ on the wavefunction ψ of the molecule (Equation 2.1).

$$\boldsymbol{H}\psi = E\psi \tag{2.1}$$

While simple in appearance the equation is impossible to solve outside a few simple examples. Therefore, for more complex systems, like molecules, approximations must be made to solve the equation.

The first is the Born-Oppenheimer approximation[89] which postulates that the difference in mass between electrons and nuclei means that on the time scale of electrons moving the nuclei can be considered as fixed. This means that the Hamiltonian can be partitioned into separate contributions from the nuclei and the electrons (Equation 2.2):

$$\boldsymbol{H} = \boldsymbol{T}_n + \boldsymbol{T}_e + \boldsymbol{V}_{ne} + \boldsymbol{V}_{ee} + \boldsymbol{V}_{nn} \tag{2.2}$$

where $\boldsymbol{T}$ is the kinetic energy and $\boldsymbol{V}$ is the potential energy and the subscript *n* is for the nuclei contribution and the *e* is the electronic contribution. Given the nuclei are frozen in this approximation, the kinetic energy of the nuclei is zero.

Therefore, the full equation of the Hamiltonian is described in Equation 2.3:

$$\boldsymbol{H} = -\frac{1}{2}\sum_{i=1}^{N}\boldsymbol{\nabla}_i^2 - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{r_{iA}} + \sum_{i=1}^{N}\sum_{j>i}^{N}\frac{1}{r_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{N}\frac{Z_A Z_B}{r_{AB}} \tag{2.3}$$

where *i* and *j* are electron indices, *A* and *B* are nuclear indices, *r* is the distance between two bodies, and Z_A is the atomic number of the nucleus. However, while this equation is solvable

for hydrogenic wavefunctions, i.e. only one electron, the electron-electron Coulomb term is not solvable, as both positions of the two electrons must be known, and therefore other approximations must be used to solve the equation for molecules with more than one electron. Approaches to solve the Hamiltonian for complex molecules can be divided into two types: wave functional theory (WFT) and density functional theory (DFT).

Wave Function Theory

To approximate the many-electron wave function, we can make use of the Hartree-Fock (HF) approximation. In HF electrons are considered as independent particles which only interact with the average field of all the other electrons. The electronic wave function is represented by a Slater determinant of the antisymmetrised product of one-electron wave functions (Equation 2.4).

$$\Psi_{HF} = \frac{1}{\sqrt{N}} \begin{vmatrix} \chi_1(x_1) & \cdots & \chi_N(x_1) \\ \vdots & \ddots & \vdots \\ \chi_1(x_N) & \cdots & \chi_N(x_N) \end{vmatrix} \tag{2.4}$$

Each electron is represented as a spin-orbital, which is expanded as a basis and must be solved variationally until the energy of the system convergences. This method is also known as the Self-Consistent Field (SCF) approach.

The final HF energy (E_{HF}) is calculated according to Equation 2.5:

$$E_{HF} = \sum_{i}^{N} \varepsilon_i - \frac{1}{2} \sum_{ij}^{N} (\boldsymbol{J_{ij}} - \boldsymbol{K_{ij}}) + \boldsymbol{V_{nn}} \tag{2.5}$$

$$\varepsilon_i = -\frac{1}{2} \boldsymbol{\nabla_i^2} - \sum_{A=1}^{M} \frac{Z_A}{r_{iA}} + \sum_{j}^{N} (\boldsymbol{J_{ij}} - \boldsymbol{K_{ij}}) \tag{2.6}$$

Where ε_i is the energy of orbital *i*, Z_A is the atomic number of atom *A*, r_{iA} is the distance between atom *A* and orbital *I* and $\boldsymbol{V_{nn}}$ is the nuclear repulsion energy. Here two important

operators have been included, the Coulomb operator $\boldsymbol{J_{ij}}$ which represents the electron repulsion by the mean field of the remaining electrons, while the $\boldsymbol{K_{ij}}$ operator represents the electron exchange. While this approximation works well, HF does not consider electron correlation and therefore dispersion is poorly modelled.

Post-HF methods such as Møller-Plesset[90] and Coupled Cluster[91] were developed to overcome these limitations. These methods use electron excitations from occupied orbitals to virtual orbitals to further derive corrections to the HF Wavefunction which describes how electronic correlation will affect the energy of the system.

Coupled Cluster calculations can include Single (S), Double (D), Triple (T) or Quadruple (Q) excitations. One of the most common is CCSD(T)[92] where both the S and D excitations are calculated explicitly while the T contribution is calculated using perturbation theory. This has become a Gold Standard in Computational Chemistry,[93-95] however, the scaling factor of N^7 for the number of electrons makes it prohibitionary expensive for systems with > 50 atoms.

Density Functional Theory

Alongside the development of WFT was the development of density functional theory (DFT). DFT is based on the theory by Kohn and Hohenberg that the energy of a molecule is dependent on the electron density of the molecule.[96] In theory, the ground state electron density depends solely on the three spatial coordinates of the system and therefore should scale better than comparable WFT methods. Kohn-Sham further developed this work by proposing a method with non-interacting electrons which would have the equivalent electron density as the system with interacting electrons.[97] While this does result in more accurate energies than orbital-free methods, the reintroduction of orbitals does result in more computationally intensive calculations. This allowed for the partitioning of the energy calculation into the kinetic energy ($\boldsymbol{T}$), potential energy ($\boldsymbol{V}$), Coulomb energy ($\boldsymbol{J}$) and the exchange-correlation energy (E_{XC}) (Equation 2.7).

$$E_{DFT} = \boldsymbol{T} + \boldsymbol{V} + \boldsymbol{J} + \boldsymbol{E_{XC}} \tag{2.7}$$

However, while $\boldsymbol{T}$, $\boldsymbol{V}$ and $\boldsymbol{J}$ can all be calculated explicitly, the exact E_{XC} energy is not known and therefore must be approximated. This has led to a wide range of functionals which all calculate the E_{XC} energy differently, also known as the functional zoo.[98, 99]

One way to describe the approach to DFT functional development is the Jacob's ladder of functionals (Figure 5).[100] The bottom rung is the local density approximation (LDA) where E_{XC} is dependent only on the electron density of an electron gas. The second rung is the generalised gradient approximation (GGA), where E_{XC} is also dependent on the gradient of the electron density. The third rung is meta-GGA which also include dependence on the second derivative of the electron density, which is equivalent to the kinetic energy of the density. The fourth rung is hybrid GGA which include explicit HF exchange in the E_{XC} calculation. The kinetic energy of the density is also sometimes added to generate hybrid-meta-GGA functionals.

More recently double hybrid DFT has been developed which mix perturbation theory and DFT calculations.[101] In this case the Exchange energy is calculated using single excitations of electrons which allows for a more accurate description of the energy. Recent advances in computational power have allowed access to these more computationally intensive calculations.

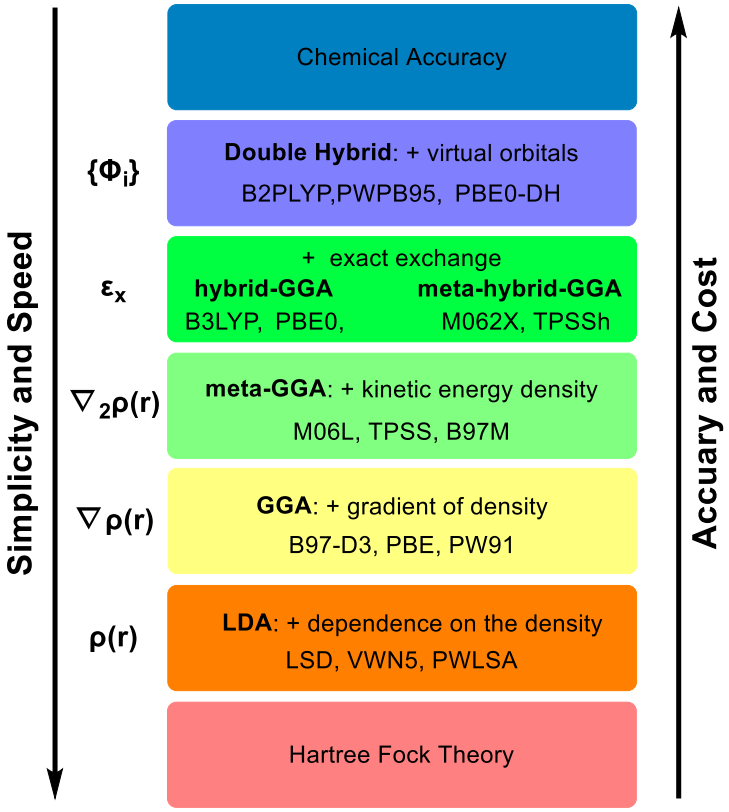


Figure 5 Jacob's ladder of DFT functionals, from HF to chemical accuracy with examples of DFT functionals for each rung on the ladder

While the deployment of more complex functionals has improved the energy calculation of molecules, dispersion energy is still not appropriately accounted for. For that reason, a range of empirical methods have been developed by Grimme *et al.*[102-104] These apply distance-dependant corrections to the energy based on pairwise distances or three body interactions between atoms with no needed to provide bonding structure.

Basis sets

In the case of either DFT or WFT calculations, a basis set must be chosen to represent the electronic orbitals for each atom. While Slater orbitals were originally developed to correctly represent the electronic wave function, they are computationally expensive to calculate integrals for. Therefore, in place of a Slater basis set a Gaussian basis set can be used. While Gaussian functions do not describe the cusp correction and do not exhibit exponential decay, they are much faster to calculate integrals for and therefore multiple Gaussians can be combined to represent one Slater orbital, with minimal loss in accuracy.

Basis sets are usually described according to the number of basis functions per electronic orbital in an atom. Single Zeta basis sets are the simplest but usually are not adequate to describe chemical reactions. Double, Triple or Quadruple Zeta basis sets have multiple basis functions per electron orbital and therefore are more flexible to accurately calculate molecule energy and properties.

More recently the use of split-valence basis sets such as Ahlrichs[105] def2-XVP have been developed where core electronic orbitals are represented by a single-zeta basis set and valence electronic orbitals have double (X = S), triple (X = TZ) or quadruple-zeta (X = QZ) basis sets. Furthermore, specific basis sets for molecule properties such as NMR,[106] coupling constants,[107] and electron spin resonance[108] have been optimised to reproduce experimental calculations.

1.5.2 Analysis tools

With the electron density in hand, it then can be useful to calculate a range of molecular properties, such as charges, bond orders and HOMO/LUMO levels to understand and predict reactivity.

Charges

There are two broad categories of charges which can be calculated from DFT: ones based on the density matrix obtained from DFT calculations (Mullikin[109] and Löwden[110] charges), and

ones which compare the electron density around an atom to a reference density (Hirshfeld[111] charges). The main advantage of Hirshfeld charges is that they are less susceptible to changes in basis set size and are therefore more reliable. The calculation of Hirshfeld atomic charges is shown in Equation 2.8:

$$q_a = Z_a - \int \frac{\rho_A^0(\boldsymbol{r})}{\sum_B^N \rho_B^0(\boldsymbol{r})} \rho(\boldsymbol{r})\, d\boldsymbol{r} \tag{2.8}$$

Where Z_a is the atomic number, ρ_X^0 is the reference density on atom X, $\rho(\boldsymbol{r})$ is the molecular density. All calculated charges in this thesis are Hirshfeld charges.

Bond orders

Alongside charges, another chemical property used to explain reactivity and bonding is the bond order between two atoms. One method used in this thesis is the Mayer bond order[112] which defines the bond order between two atoms (Equation 2.9).

$$B_{AB} = 2 \sum_{\lambda \,\epsilon\, A} \sum_{\omega \,\epsilon\, B} (\boldsymbol{DS})_{\lambda\omega} (\boldsymbol{DS})_{\omega\lambda} \tag{2.9}$$

Where **D** and **S** are the density and overlap matrices respectively.

1.5.3 NMR calculations in DFT

Calculation of NMR shielding

The nuclear shielding parameter σ is an important second-order derivative of the energy of a molecule (Equation 2.10).

$$\sigma_{st} = \frac{\partial^2 E}{\partial B_S \, \partial \mu_t}|_{B=\mu=0} \tag{2.10}$$

Where the shielding constant σ varies as a function of the total electronic energy E with respect to the magnetic field B and the nuclear magnetic moment of nucleus N, μ_N , in a coordinate system *st*. The Hamiltonian is modified to include a dependency upon the field, which can be then used to solve the shielding tensor for all possible 9 interactions between B and μ (Equation 2.11).[113]

$$\boldsymbol{\sigma_{st}} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{yz} & \sigma_{zz} \end{pmatrix} \tag{2.11}$$

The NMR tensor is then diagonalized to give shielding constants with respect to the external laboratory frame of reference (Equation 2.12).

$$\boldsymbol{\sigma_{st}} = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix} \tag{2.12}$$

In solution, the rapid tumbling of the molecule means that these 3 constants are averaged over time to give σ_{iso}(Equation 2.13).

$$\sigma_{iso} = \frac{\sigma_{11} + \sigma_{22} + \sigma_{33}}{3} \tag{2.13}$$

This is the origin of the isotropic shielding constant. One important effect of the modification of the Hamiltonian is its dependence on the origin (or gauge) of the magnetic field. By changing the origin of the magnetic field, there would also be a corresponding change in the isotropic shielding constant, but this is non-physical and needs to be accounted for. One of the most

popular methods is the gauge invariant (including) atomic orbitals (GIAO) method,[114] where the atomic orbitals are modified by the inclusion of a dependence on the magnetic field (Equation 2.14):

$$\Psi_i(\vec{B},\vec{r}) = \sum_{\alpha} c_{i\alpha}(B)\,\omega_\alpha(B,\boldsymbol{r}) \tag{2.14}$$

where:

$$\omega_\alpha(B,\boldsymbol{r}) = exp\left[-\frac{i}{2}(B\ \times R_\alpha)\cdot\boldsymbol{r}\right]\chi_\alpha(\boldsymbol{r}) \tag{2.15}$$

χ_α is a set of atomic orbitals, $\boldsymbol{r}$ is the electronic position operator, c is the expansion coefficients, and R_α is the position vector of the atom where orbital χ_α is located. The modification of the atomic orbitals by the field-dependant phase factor $\omega_\alpha(B,\boldsymbol{r})$ results in the expectation values which are gauge invariant.

This shielding constant is the absolute value of the nuclear magnetic shielding. However, in NMR experiments the observed ppm shift is the shielding of a given molecule compared to a reference, which is TMS for ^{1}H and ^{13}C and $CFCl_3$ for ^{19}F.

$$\Delta\delta = \sigma_{ref} - \sigma_{iso} \tag{2.16}$$

Hence experimental NMR shifts Δδ are related to the isotropic shielding constant (σ_{iso}) according to Equation 2.16 where σ_{ref} is the isotropic shielding constant of the reference molecule, Usually TMS for ^{1}H and ^{13}C NMR and $CFCl_3$ for ^{19}F NMR).

Calculation of coupling constants

When considering the calculation of coupling constants we calculate the nuclear spin-spin coupling tensor $\mathbf{J}_{MN}$ as the second derivative of the energy with respect to the nuclear magnetic moment of each nucleus *M* and *N* (Equation 2.17).

$$\boldsymbol{J}_{MN} = \frac{\partial^2 E}{\partial \vec{\mu}_M \, \partial \vec{\mu}_N} |_{\vec{\mu}_M = \vec{\mu}_N = 0} \tag{2.17}$$

As seen with the σ_{iso} calculations, we only observe the averaged value of the $\mathbf{J}_{MN}$ tensor and therefore the J_{MN} value is calculated in Equation 2.18.

$$J_{MN} = \frac{J_{MN11} + J_{MN22} + J_{MN33}}{3} \tag{2.18}$$

Where J_{MN11}, J_{MN22}, J_{MN33}, are the diagonal elements of the $\mathbf{J}_{MN}$ tensor. The $\mathbf{J}_{MN}$ can be portioned into four contributions widely known as the Ramsey terms[115, 116]: Fermi contact (FC), spin-dipole (SD), diamagnetic spin-orbit (DSO) and paramagnetic spin-orbit (PSO). While a complete description of the terms is beyond the requirements here, a brief description is given, and further details can be found in these papers.[115-118]

The FC term can be considered a direct coupling between electrons and the nucleus and is usually the major contribution to the coupling constants.[119] The SD term couples together the nucleus's magnetic moment with the electrons' spin. The FC and SD terms are linked as they both describe coupling due to the spin polarisation of the system. The FC term considers the effect of the magnetic field inside the nucleus, while the SD term considers the extended dipole field outside the nucleus.[120] The DSO and PSO terms are derived from the interactions between the nuclear magnetic moments and the orbital magnetic moments and are calculated for the diamagnetic and paramagnetic contributions respectively.

One interesting property of the J_{MN} coupling constant is that it can be both negative and positive in value. The J_{MN} is as positive if the coupling stabilises anti-parallel spins, while if parallel spins are stabilised, the coupling constants are negative.

1.5.4 Solvation effects in DFT calculations

Solvent and counterions can also affect the calculated ^{19}F NMR values and their consideration is particularly relevant when studying fluorides. Solvent effects can be modelled computationally using either an explicit or implicit (analytical) description of the solvent environment (Figure 6). In implicit solvation the solvent is treated as a continuum with a certain dielectric and interfacial properties. Two examples of implicit solvent methods are CPCM[121, 122] and SMD.[123] The solute cavity is generated based on the radii of each atom, the surface of this cavity is split into tesserae using the GEPOL algorithm.[124] The surface charges on each tesserae are calculated from atomic partial charges and the interaction between these charges and the dielectric constant of the solvent is the basis for the CPCM methodology. The SMD method adds extra parameters that account for the surface tension between the surface charge and solvent which are specific for each solvent.[123]

The crucial point is that all the charge is contained within the cavity so no charge can "escape".[125] Therefore any stabilisation of the charge by the solvent is directly dependant on the dielectric constant of the solvent, rather than by explicit interactions which remove electron density.

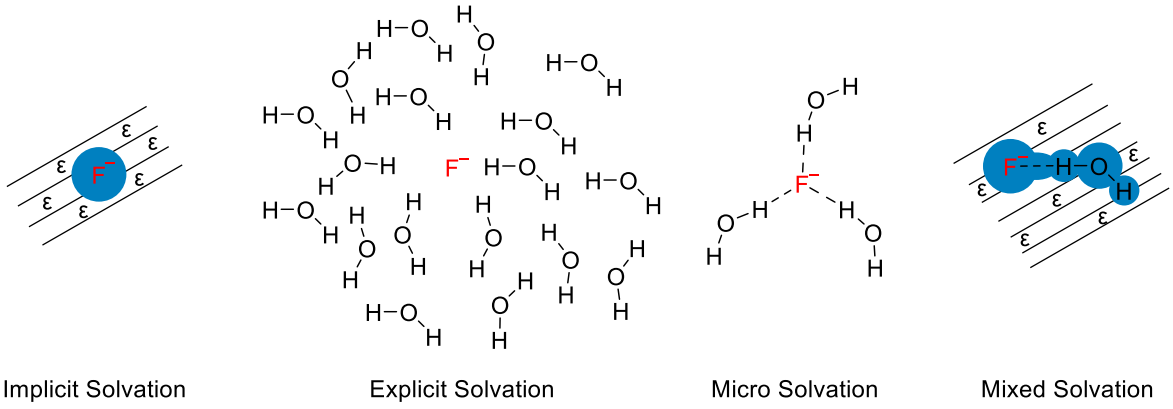


Figure 6 Examples of Implicit, Explicit, Micro and Mixed solvation methods for Fluoride in Water

For explicit solvation methods, solvent molecules are explicitly modelled surrounding a molecule or ion of interest, with the size of the explicit solvation left up to the user to design. Furthermore, large-scale simulations of explicit solvation can be challenging for computational calculations. One possible option is micro solvation, where only a small amount of a molecule is explicitly solvated, and no implicit solvation is used however the challenge is deciding on the number of solvent molecules required, which can be based on chemical intuition or benchmarking. Finally, mixed solvation uses explicit solvation to model solvent effects around the molecule of interest, while implicit solvent is used to treat the remaining bulk solvent effect.

1.5.5 Conformational sampling

Given the wide number of conformers many molecules can exhibit, a method to sample the conformational space is needed. Two methods are described here which are used throughout this thesis.

Semi-empirical methods

Semi-empirical methods are computational techniques which aim to simplify either HF or DFT calculations. Most commonly these simplifications are taken either by reducing the number of integrals to calculate, such as removing three or four electron integrals, and then reducing the complexity of the overlap matrix. The rest of the terms are then parameterised based on element-specific terms, either for atomic energies, atomic orbital overlap, or dispersion interactions. This has led to a variety of methods for the calculation of molecular energies, including AM1,[126] PM3,[127] and PM6.[128]

The other possible approximation that can be made in semi-empirical methods is an approximation in the Fock matrix. This is widely seen in Density Functional Tight Binding theory (DFTB)[129, 130] where three and four electron integrations are set to zero. Furthermore, only the valence electrons are considered, with the core electrons being only included as a repulsion potential, fitted from experimental data. The method used throughout this thesis is

GFN2-xTB,[50] developed by the Grimme group as part of the xTB[50, 51, 131] package. GFN2-xTB is built on DFTB3,[132] including a charge correlation term expanded to the third order. GFN2-xTB uses a minimal valence basis set of Gaussian functions to approximate Slater functions (STO-*m*G) and includes polarisation functions for main group elements. Dispersion interactions are modelled using the D4 model, and only element-specific parameters are used.

The overall energy for a molecule using the GFN2-xTB methods $E_{GFN2-xTB}$ is calculated in Equation 2.19:

$$E_{GFN2-xTB} = E_{Rep} + E_{disp} + E_{IXC} + E_{IES} + E_{AXC} + E_{AES} + G_{Fermi} + E_{EHT} \quad (2.19)$$

Where E_{Rep} is the repulsion energy, E_{disp} is the dispersion energy, E_{IXC} is the isotropic exchange energy, E_{IES} is the isotropic electrostatic energy, E_{AXC} is the anisotropic exchange energy, E_{AES} is the anisotropic electrostatic energy, G_{Fermi} is the Fermi smearing energy and E_{EHT} is the extended Hückel type energy.

As with DFTB3, the model is dependent on atomic distances, orbital overlap, atomic charges, dipole moment, and multipoles. As the latter three variables are not known they must be solved in a self-consistent manner.

Conformational sampling in xTB can be carried out using two methods: simulated annealing (SA) or CREST. SA uses three molecular dynamics simulations, set at three different temperatures for 50 ps where conformations are saved every ps. This gives up to 150 conformers which are optimised by GFN2-xTB and are then sorted and anything above 20 kcal/mol of the lowest energy conformer is excluded.

The CREST workflow is a meta-dynamics workflow, where an RMSD (Root Mean Square Distance) bias is used to explore new conformer space. Multiple different biases are used before MD simulations are carried out around the lowest six conformers. Finally, Z-matrix crossing and geometry optimisations, along with filtering, is used to generate the full conformational ensemble.

Molecular dynamics

Classical molecular dynamics simulations make use of a ball and stick method, where atoms are treated as charged particles, and bonds or interactions between atoms treated using a classical force field mimicking a spring. The force field therefore is key in being able to reproduce and analyse molecular movement over time. A large number of different forcefields have been developed including AMBER,[133, 134] CHARMM,[135] UFF,[136] GAFF,[137] and OPLS-AA.[138, 139] These forcefields separate out the contributions to the overall energy into bonding and non-bonding components. The bonding parameters are for bonds, angles and dihedrals, while the non-bonding terms are the Van der Waals and electrostatic terms (Equation 2.20).

$$E_{MD} = E_{Bond} + E_{Angles} + E_{Dihedral} + E_{VDW} + E_{Electrostatic} \quad (2.20)$$

The fact that the energy can be broken down into such simple parts means that large-scale simulations, with explicit solvent, are feasible with millions of atoms being simulated. However, while efficient the accuracy of the simulations depends solely on the forcefield used, due to the nature of the system bond breaking and forming cannot be modelled. While mixed QM-MD methods are possible, they are not a focus in this work.

Chapter 2 Machine Learning: Background and Theory

2.1 Machine learning in organic chemistry

Fast, accurate and efficient prediction of chemical properties is a central goal of computational chemistry. Significant advances in computing power have enabled the use of high-level electronic structure methods, such as Double Hybrid Density Functional Theory (DH-DFT) and Coupled Cluster calculations (CCSD(T)), which approach chemical accuracy in a wide range of situations.[140-145] However, the cost of scaling of these methods prohibit their application to systems with larger than 150 atoms for energy calculations and even less for gradient calculations. Machine Learning (ML) models promise to bridge this gap, reaching chemical accuracy at a fraction of the computational cost of DFT and CCSD(T) methods. ML methods have been applied to a range of chemical problems including reactivity,[146-151] reaction yield,[152-156] chemical selectivity,[146, 157-161] energy barriers,[162-172] frequency calculations[173-177] and NMR spectroscopy.[87, 178-185]

2.1.1 Machine learning models

To accurately predict a range of chemically relevant qualities a selection of different ML models has been employed. In a range of Quantitative Structure Activity Relationship (QSAR) papers, linear regression models have been utilised to predict enantioselectivity and yield,[186] or a range of biological properties.[187-191] As datasets have increased in size more complex models have been used such as support vector machines, decision trees, and kernel-based methods. These have been used to create models to predict a range of chemically relevant values such as predicting energy and forces for molecular simulations,[192-195] yield,[154, 155, 159] and protein binding.[196-199] The next level of complexity comes from Neural Network (NN) based models. NNs have been widely used in MD to replace force field-based models.[200-205] NNs have also

found wide use in the prediction of molecular properties[146, 180, 206, 207] and binding affinity of different proteins.[208, 209] Graph Neural Networks (GNN), which either use the 2D or 3D structure of a molecule, have found wide applicability in chemistry, with uses for predicting bond dissociation energies,[166, 167] retrosynthesis[210-212] and molecular properties[213, 214] such as NMR.[85, 215]

More recently models originally developed for Natural Language Processing (NLP) have been utilised in a range of retrosynthesis planning, yield prediction, and classification tasks. The majority of these models use the Transformer[216] as their underlying architecture: a model which is used widely in text based tasks such as translation to generate readable, sensible, and understandable text based on a given input.[217, 218] The aim of the Transformer is to learn how words in sentences relate to each other rather than just based on the previous word. This therefore gives better context for the word in a sentence and allows for better prediction for text-based tasks such as translation, chat bots, and search functions.

In chemistry these were originally utilised for retrosynthesis prediction where SMILES (Simplified Molecular Input Line Entry System) strings are used as the text input. The concept of learning the importance of words in a sentence is converted into learning the importance of atoms and bonds in a molecule. The translation task is similarly changed into a retrosynthesis step, as that is a translation from products back to the original reactants. These models are now the leading models for reaction and retrosynthesis prediction and with the first known as the Molecular Transformer.[149] This work has now been adapted to be able to predict the yield of a reaction[10] or classify the type of reaction[219, 220] and to suggest methods to synthesis a molecule in response to a question.[221]

Recently Natural Language Processing techniques have been applied to a range of chemical problems such retrosynthesis prediction,[149, 222] molecule generation,[207, 223-228] the prediction of

pharmacokinetic properties,[207, 209, 229-231] solvation free energy and solubility,[232] and reaction classification.[219, 220, 233, 234]

2.1.2 Machine learning to predict enantioselectivity and yield

A large body of research has been carried out in prediction of reaction selectivity and yield. Work from the Sigman group has focused on developing multivariate linear regression models to predict enantioselectivities for a wide range of systems,[186, 235] including the desymmetrisation of bisphenyls,[236] allylations of aceophenone,[236] kinetic resolution of benzyl alcohols and synthesis of allylic fluorines,[237] and a range of organocatalysed asymmetric reactions.[238-241] Usually, these models use descriptors derived from quantum mechanics (QM) calculations, however Sigman *et al.* have also developed a range of parameters to describe non-covalent interactions and steric bulk. The Stermoil parameters are one such descriptor which represent steric bulk through three parameters B_1, B_5, and L.[236] For a given bond X-Y B_1 represents the shortest distance perpendicular from the primary axis of attachment (i.e. the minimum width of a substituent), B_5 represents the longest distance (maximum width of a substituent), and L is the total distance following the primary axis of attachment (length). These methods are able to predict the enantioselectivity of a particular transformation with an RMSE of below 0.2 kcal/mol and R^2 of greater than 0.95.[186, 236]

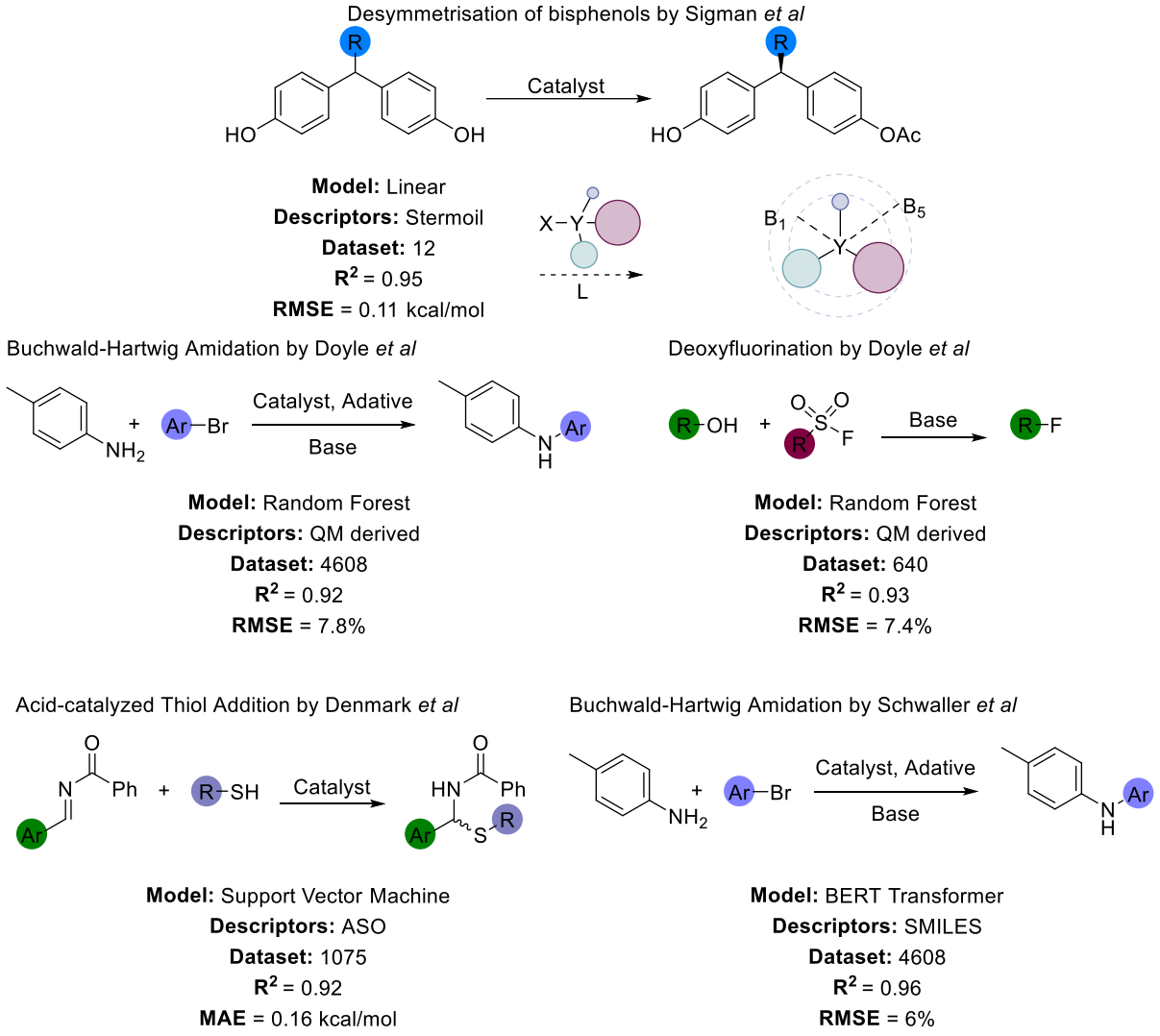


Figure 7 Selected examples of machine learning methods for the prediction of yield and enantioselectivity.

More recently high throughput experiments have allowed for the generation of large diverse datasets and the building of more complex models. Doyle *et al.* have used QM-derived descriptors and Random Forest models to predict the yield of Buchwald-Hartwig couplings[155] and deoxyfluorination reactions.[159] Both of these two methods achieved RMSEs of 7.8 % and 7.4 % with an R^2 of 0.92 and 0.93 respectively. Denmark *et al*. took the concept further by developing their own descriptors based on the positions of atoms in 3D space for a range of phosphoric acid catalysts. These descriptors are utilised by support vector machines to predict the enantioselectivity for the nucleophilic addition of thiols to *N*-acyl imines.[158, 161, 242, 243] This

method was able to predict the enantioselectivity of the reaction with a MAE of 0.16 kcal/mol and a R^2 of 0.92.

In 2021 Schwaller *et al.* published YieldBert,[10] a Transformer-based model which could be used to predict the yield of Buchwald Hartwig couplings based on the dataset generated by Doyle *et al.*[155] This model used a Transformer that was pre-trained on a large amount of data from the USPTO database, before an additional regression layer was finetuned to predict yield. This pretraining allowed the model to learn about chemical space in a non-supervised manner and then apply that knowledge during finetuning to identify what parts of the molecule would contribute to a high or low yield. This method was very successful, achieving an error of 6% (RMSE) and a R^2 of 0.96. This work has also been reproduced with other datasets to predict the yield of Suzuki reactions.[10, 220]

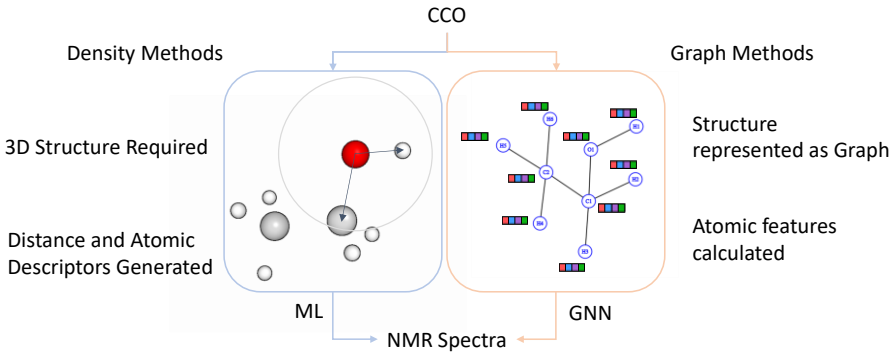
2.1.3 Prediction of ^{1}H and ^{13}C NMR shifts using machine learning

The widespread use of NMR spectroscopy in structural elucidation has led to increasing interest in the use of ML to facilitate the prediction of ^{1}H and ^{13}C NMR shifts. ML methods that can optimise this process and match computed spectra to experimentally determined structures would significantly speed up experimental discovery and structure confirmation when compared to DFT based methods.

ML-NMR has focused on the prediction of ^{1}H or ^{13}C spectra and follows two main approaches: density-based, and graph-based methods (Figure 8). We define density-based methods as those which take atomic coordinates and generate descriptors based on the distance between atoms and the atom type, This therefore can be considered as a representation of the electronic density around each atom.[84, 184, 244] Butts *et al.*[185] have employed smooth overlap of atomic positions (SOAP) descriptors,[195] in combination with kernel ridge regression, to predict ^{1}H and ^{13}C NMR

shifts on a DFT derived dataset. Overall, this method worked very well with RMSEs of 0.35 ppm and 3.88 ppm for ^{1}H and ^{13}C predictions compared to their DFT calculations.

Graph-based methods instead use graph representations of the molecules, where atoms are described as nodes and bonds as edges on a graph,[182, 245] with node descriptors such as atomic number, number of neighbours, and charge.



Paper	Model	Descriptors	Dataset	^{1}H Dataset	^{13}C Dataset	Train:Test	^{1}H Error (ppm)	^{13}C Error (ppm)
Butts *et al*	KRR	SOAP	DFT	18 383	17 081	410*	0.35 (RMSE)	3.88 (RMSE)
Kuhn *et al*	GNN	Calculated Features	Experimental	32,538*	32,538*	4:1	0.4 (RMSE)	1.3 (RMSE)
Kang *et al*	MPNN	Calculated Features	Experimental	12,800*	26,859*	4:1	0.22 (MAE)	1.35 (MAE)
Kang *et al*	MPNN	Learnt Features	Experimental	12,800*	26,859*	4:1	0.22 (MAE)	1.36 (MAE)
Paton *et al*	GNN	RBF	DFT	120 000	100 000	500*	0.16 (RMSE)	2.15 (RMSE)
Zhang *et al*	NN	Calculated Features and NMR Shift	Experimental	270	476	9:1	0.18 (RMSE)	2.10 (RMSE)
Beran *et al*	NN	AEV + NMR Shift	DFT	298,081	496,275	3780*	0.11 (RMSE)	0.70 (RMSE)

Figure 8 Overview of machine learning predictions for ^{1}H and ^{13}C NMR. All dataset sizes are reported as number of shifts, expect when signified by a * which indicates total number of molecules. Train:Test shows the split of the Dataset (numbers indicated a separate set of molecules).

Kuhn *et al*,[182] used a GNN with atomic features calculated in RDKit[246] to predict the ^{1}H and ^{13}C NMR shifts of a range of molecules taken from the NMRShift[86] database. The model also predicted a confidence value for each NMR shift prediction. Overall, the model worked well, with RMSEs of 0.4 ppm for ^{1}H NMR and 1.3 ppm for ^{13}C NMR. The work of Kang *et al*.[180,

[206] used a message-passing neural network (MPNN), a type of GNN, to predict both ^{1}H and ^{13}C NMR shifts. Their first contribution[180] used a 2D graph with a range of descriptors for both the bonds and atoms in the graph. For atoms, these include atomic numbers, formal charges, implicit valence, and ring size, while bonds include descriptors such as bond type if the bond is aromatic or in a ring. The follow-up paper from Kang *et al.*[206] changed the underlying method of the network so that the MPNN learnt atomic features during training rather than having to be calculated beforehand. Therefore, only the structure is needed to create the 2D graph and no further descriptors are calculated. Both two methods perform equally well with MAEs of 0.22 ppm and 0.22 ppm for ^{1}H NMR and 1.35 ppm and 1.36 ppm for ^{13}C NMR, respectively.

Paton *et al.*[85] used a GNN similar to that in SchNet[247] where a 3D graph is generated from 3D atomic coordinates. Atomic descriptors are calculated using radial basis functions between all atoms within 5 Å of each other. These are used as inputs within the GNN which is then used to predict both ^{1}H and ^{13}C NMR, with RMSEs of 0.16 ppm and 2.15 ppm for ^{1}H and ^{13}C NMR respectively.

The final type of models are Δ-ML approaches. This is where an ML method takes the molecular structure and NMR prediction at the low level of theory and predicts the difference between the low level of theory and the NMR shifts if they had been calculated using higher levels of theory or experimental data. The work of Zhang *et al.*[83] used an NN which takes hand-selected descriptors from the surrounding atoms. The lower level of theory is taken from QM calculations at the B3LYP/cc-pVDZ//B3LYP/cc-pVDZ level of theory, and the NN predicted the difference between this method and the experimentally obtained values. This NN predicted the experimentally obtained NMR shifts for each atom with RMSEs of 0.18 ppm and 2.10 ppm for ^{1}H and ^{13}C respectively.

Beran *et al.*[84] use a similar approach, however, the input to the NN was the atomic environment vector (AEV)[248] of the atom of interest which was a set of calculated descriptors based on the

atom's surrounding environment. This model corrects between two levels of DFT theory: PBE0/6-31G to PBE0/6-311+G(2d,p). This model performed exceptionally well with RMSEs of 0.11 ppm and 0.70 ppm for ^{1}H and ^{13}C respectively.

2.2 Background on machine learning

2.2.1 Molecular representations in machine learning

To develop a ML model for a chemical process firstly the molecules need to be represented in a way in which the algorithm can process the data. While multiple methods exist to generate molecule representations they can loosely be broken down into 4 types, 1D, 2D and 3D descriptors and Molecular fingerprints (Figure 9).

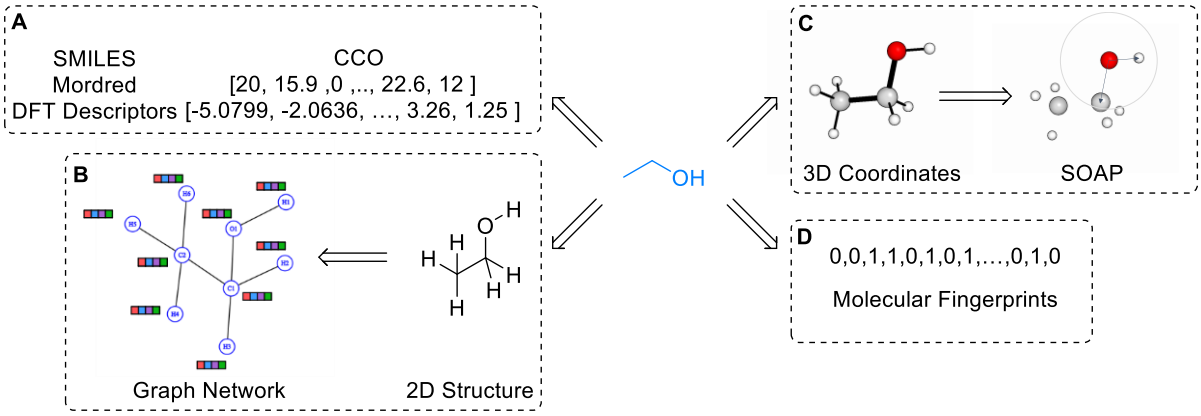


Figure 9 Generation of molecular descriptors A) 1D descriptors such as SMILES, Mordred and DFT descriptors B) 2D descriptors using a 2D graph network C) 3D descriptors generate 3D coordinates before deriving either graphs or atomic descriptors such as SOAP. D) Fingerprint generation from an input structure.

1D descriptors take the form of vectors or strings, SMILES (Simplified Molecular-Input Line-Entry System[249]) strings are an example of this. SMILES is a way to represent molecules using text characters. Atoms are signified using their elemental symbols (H, C, N, As), with hydrogen generally being implied based on the valence of each atom. Lower case atom letters (c,n) show that those atoms are aromatic. Adjacent letters in a SMILES string are joined by single bonds, while = is used to represent a double bond and # represents a triple bond. Chirality is

represented using the @ symbol, where @ indicates the R stereocentre and @@ the S. Numbers are used to indicate the closure of a ring. These simple building blocks are used to create strings of characters that can describe complex molecules such as amoxicillin as seen in Figure 10.

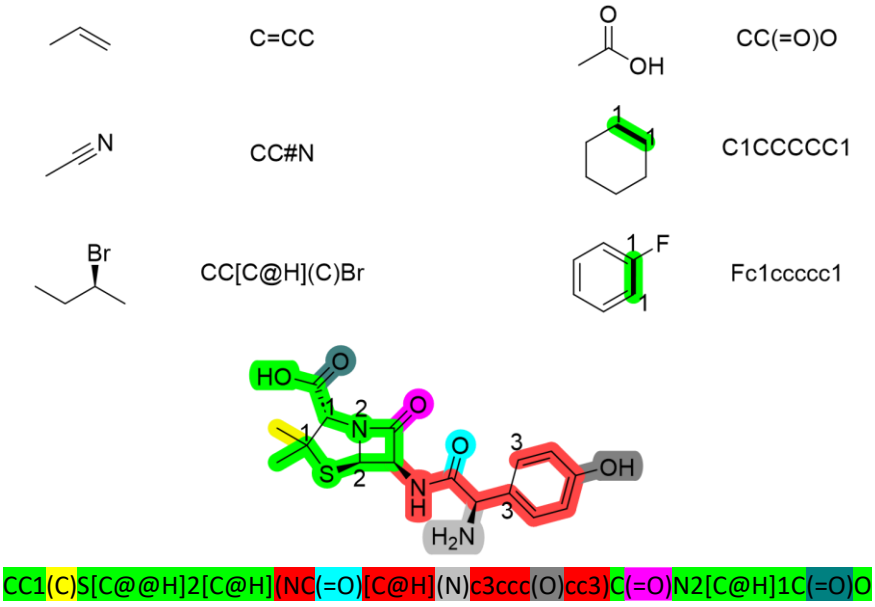


Figure 10 SMILES Representation of molecules, showing ChemDraw representation and the corresponding SMILES String. Amoxicillin is shown with the structures and atoms highlighted in SMILES string

Other 1D representations of molecules can include lists of physical organic properties such as Log P, polar surface area, number of polar hydrogens, etc., or those derived from QM calculations such as HOMO, LUMO, atomic charges, or bond vibrations. The Mordred Python package[250], calculates a wide range (1800) of 1D descriptors including SlogP, the number of rings, molecular weight, molecular volume and atom counts. These descriptors can be screened to identify those which correlate well with experimentally observed parameters.

2D representations start with a 2D representation of a molecular structure, such as a ChemDraw structure, and either create a 2D adjacently matrix or a molecular graph consisting of atomic nodes and bond edges. For graph-based neural networks, alongside the graph structure each node and edge also takes a range of descriptors such as atom type, valency, and types of bonds.

3D representations take a 3D structure, such as those obtained from force fields or DFT calculations, and either create a graph in 3D or take atomic positions as inputs. Smooth Overlap of Atomic Positions (SOAP) are a type of 3D representations which take the local environments of each atom to create a descriptor for each atom in the molecule.

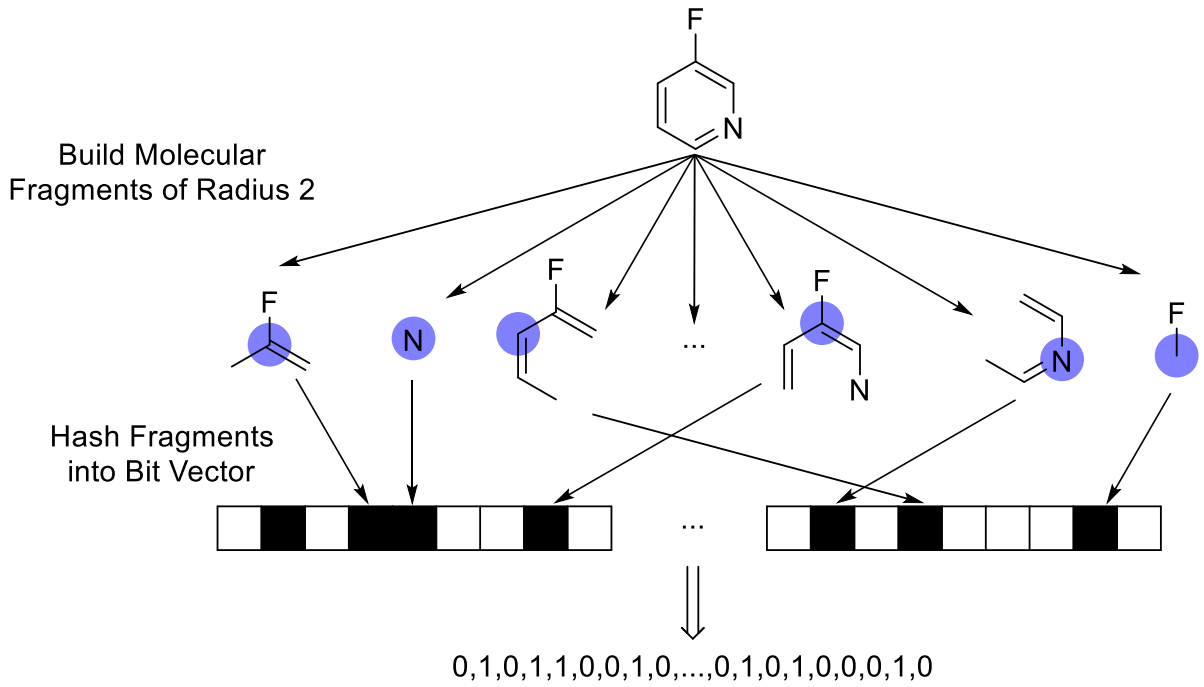


Figure 11 Generation of Molecular Fingerprints using the Morgan method, blue circle shows centre of each fragment

The final type of molecular representation is molecular fingerprints. The main type is the Morgan fingerprint,[251] Which breaks the molecule down into fragments with a set maximum radius. These fragments are then hashed to create a bit vector of a specific length. The bit vector therefore encodes a fragment of a molecule either as a 0 or a 1. One possible issue in the generation of fingerprints is a bit clash, where multiple fragments are hashed into the same bit. This can mean that when comparing fingerprints, it can be difficult for models to learn a direct relationship between structures and activities.

One use of fingerprints is in the calculation of molecular similarity. The Tanimoto similarity score[252, 253] is used within this work to calculate the similarity between two fingerprints (Equation 2.21).

$$Tanimoto\ Similarity = \frac{|\boldsymbol{A} \cap \boldsymbol{B}|}{|\boldsymbol{A}| + |\boldsymbol{B}| - |\boldsymbol{A} \cap \boldsymbol{B}|} \tag{2.21}$$

Where A and B are bit vectors, and the |A∩B| is the intersection of A and B, i.e. the number of bits in common.

2.2.2 Regression models

Linear Regression

Regression models are focused on the prediction of numerical outputs with a continuous range. The models attempt to identify a relationship between the input variables and the output value. The simplest form of this is simple linear regression:

$$y = mx\ + c \tag{2.22}$$

where y is the output, x in the input, m is the gradient of the slope and c is the y-intercept. Multiple linear regression results as the linear combination of multiple input variables:

$$y = \ m_1x_1 + m_2x_2 + \cdots + m_nx_n + c. \tag{2.23}$$

When trying to fit a linear model to the data the model aims to minimise the difference between the predicted output values of the model ($\tilde{y}$) and the true values of the output (y). This takes the form of the sum of squared residuals (SSR):

$$SSR = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n}(y_i - mx_i - c)^2 \tag{2.24}$$

Therefore, to obtain estimates for m and c SSR must be minimised:

$$(\hat{m}, \hat{c}) = argmin(SSR) \tag{2.25}$$

where $\hat{m}$ and $\hat{c}$ are the best estimates for m and c respectively. To obtain these Equations 2.26 and 2.27 can be used:

$$\hat{c} = \bar{y} - (\hat{m}\bar{x}) \tag{2.26}$$

$$\hat{m} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.27}$$

where $\bar{x}$ and $\bar{y}$ are the average of the x and y data respectively.

Lasso Regression

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a method which includes regularisation in the minimisation function. The assumption is that the coefficients m are sparse and close to zero, adding a penalty for large coefficients in the model. This means LASSO is also carrying out feature selection; by setting some coefficients to zero, they are being removed from the model:

$$(\hat{m}^{lasso}, \hat{c}) = argmin\left(SSR + \lambda \sum_{i=1}^{n} |m_i|\right) \tag{2.28}$$

Where the λ parameter represents the shrinkage of the model and is tuneable, either set by the user or optimised using cross-validation. When λ = 0 the model is equivalent to linear regression and when λ = ∞ all the coefficients would be reduced to zero.

2.2.3 Neural networks

Networks

In the past 30 years since the discovery of the backpropagation algorithm,[254] Neural Networks (NNs) have been successfully applied to tasks in finance, medical imaging, social networks, search engines and translation. At its core NNs are based upon two fundamental requirements: that a complex differentiable function can be defined that maps inputs to outputs, and that the model can form a Directed Acyclic Graph (DAG).

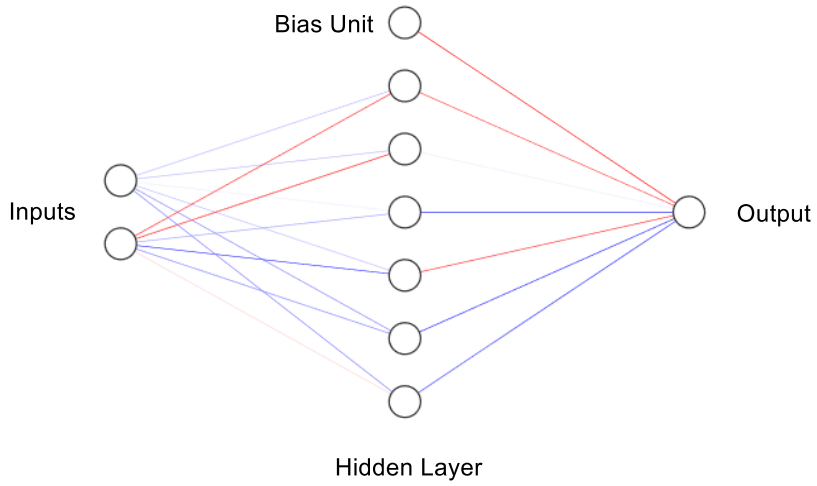


Figure 12 A Simple Neural Network which takes 2 inputs and returns 1 prediction. The network has 1 hidden layer with 6 nodes

For example, consider the toy neural network in Figure 12. This is a simple feedforward network where two inputs are mapped to one output. The centre contains one hidden layer of size six and a bias unit which is a constant similar to c in Equation 2.22 . This is a fully connected network as all the nodes are connected to the hidden layers. We can write this NN in a mathematical formula which allows for a simpler explanation for much deeper networks.

For each layer in our NN we describe we can get the formula for that layer :

$$f_l(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta_l}) = W_l \phi_l\left(\boldsymbol{x}; \boldsymbol{\theta_l^2}\right) + \boldsymbol{b} \tag{2.29}$$

Where $\phi_l\left(\boldsymbol{x}; \boldsymbol{\theta_l^2}\right)$ is the activation function of the layer where $\boldsymbol{\theta_l^2}$ contains the parameters which describe the function performed on the inputs $\boldsymbol{x}$. W_l are the weights associated with the layer and $\boldsymbol{b}$ is the bias added to each layer.

To package this more efficiently, $\boldsymbol{\theta_l}$ contains all the parameters for layer l according to Equations 2.30 and 2.31:

$$\boldsymbol{\theta_l} = \left(\boldsymbol{\theta_l^1}, \boldsymbol{\theta_l^2}\right) \tag{2.30}$$

$$\boldsymbol{\theta_l^1} = (\boldsymbol{W}, \boldsymbol{b}) \tag{2.31}$$

Thus for the toy network in Figure 12 the network can be written as:

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = f_H\left(f_I(\boldsymbol{x})\right) \tag{2.32}$$

where H is the hidden layer, I is the input layer and $\boldsymbol{\theta}$ contains all the parameters of each layer $\boldsymbol{\theta_l}$.

The power and flexibility of NNs come from the wide variety of activating functions; linear, softmax, ReLU, Tanh are a few of the many examples of simple functions which can be put together in a NN to generate a complex overall function able to predict nonlinear relationships.

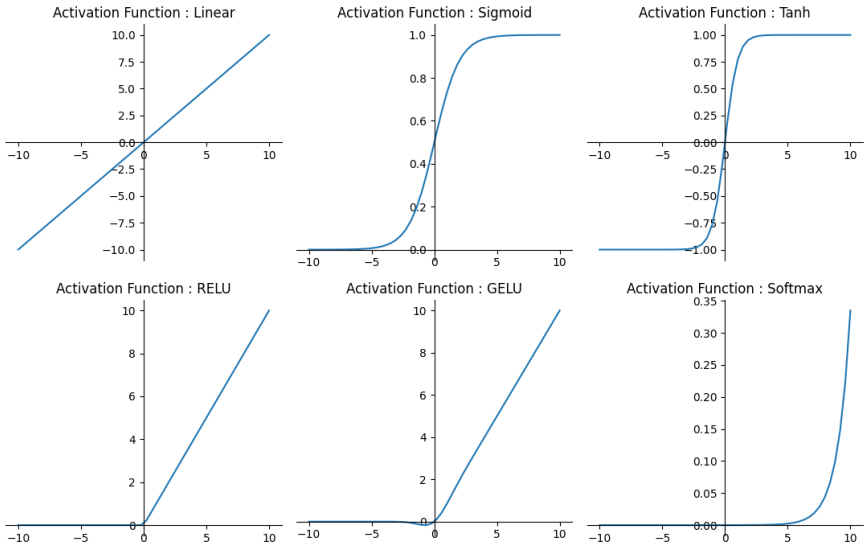


Figure 13 An example set of activation functions used in NNs

The term Deep Neural Networks (DNNs) refers to a subset of NNs which have a large number of hidden layers and can be generally written as:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = f_L\left(f_{L-1}\left(\cdots\left(f_I(\boldsymbol{x})\right)\cdots\right)\right) \tag{2.33}$$

Backpropagation

Backpropagation is the method by which the NNs update the weights for each layer during training. By optimising the weights, the model can converge on the best weights to predict the target. Firstly, one needs to define a loss function (also known as a cost function) which is to be minimised using the mean square error (MSE, Equation 2.34). For non-regression tasks more complex functions such as cross-entropy loss or K-L Divergence can be used.

$$Loss = \frac{1}{n}\|y - \overline{y}\|^2 \tag{2.34}$$

Where n is the number of samples, y is the true values and $\overline{y}$ are the predicted values. Both y and $\overline{y}$ are vectors of length n. To get a more accurate model one needs to calculate the gradient of the loss function with respect to all the parameters in the NN. This is where the backpropagation algorithm comes into play as the chain rule can be used to simplify the maths. Using our toy model in Figure 12 the model can be split up into constituent mathematical

operations. These are either activation functions or multiplication by the weights, and the input layer has only weights and no activation function. The NN for the toy model is therefore written in Equation 2.35:

$$\overline{y} = \boldsymbol{W_2}\phi_H(\boldsymbol{W_1}\boldsymbol{x}) \tag{2.35}$$

The loss of the NN therefore can be written as composition of functions where each function is one mathematical operation:

$$Loss = \boldsymbol{f_4} \circ \boldsymbol{f_3} \circ \boldsymbol{f_2} \circ \boldsymbol{f_1} \tag{2.36}$$

where:

$$\boldsymbol{x_2} = f_1(\boldsymbol{x}; \boldsymbol{\theta_1}) = W_1\boldsymbol{x} \tag{2.37}$$

$$\boldsymbol{x_3} = f_2(\boldsymbol{x_2}; \boldsymbol{\theta_2}) = \phi_I(\boldsymbol{x_2}) \tag{2.38}$$

$$\boldsymbol{x_4} = f_3(\boldsymbol{x_3}; \boldsymbol{\theta_3}) = W_2\boldsymbol{x_3} \tag{2.39}$$

$$Loss = f_4(\boldsymbol{x_4}; \boldsymbol{y}) = \frac{1}{n}\|\boldsymbol{x_4} - \boldsymbol{y}\|^2 \tag{2.40}$$

The chain rule is then used to calculate the gradient for each of the separate parameters $\boldsymbol{\theta_k}$. While the gradient for the final layer directly ($\frac{\partial Loss}{\partial \boldsymbol{x_4}}$) can be calculated directly, for the rest the chain rule is used to calculate the separate derivatives:

$$\frac{\partial Loss}{\partial \boldsymbol{\theta_3}} = \frac{\partial Loss}{\partial \boldsymbol{x_4}}\frac{\partial \boldsymbol{x_4}}{\partial \boldsymbol{\theta_3}} \tag{2.41}$$

$$\frac{\partial Loss}{\partial \boldsymbol{\theta_2}} = \frac{\partial Loss}{\partial \boldsymbol{x_4}}\frac{\partial \boldsymbol{x_4}}{\partial \boldsymbol{x_3}}\frac{\partial \boldsymbol{x_3}}{\partial \boldsymbol{\theta_3}} \tag{2.42}$$

$$\frac{\partial Loss}{\partial \boldsymbol{\theta_1}} = \frac{\partial Loss}{\partial \boldsymbol{x_4}}\frac{\partial \boldsymbol{x_4}}{\partial \boldsymbol{x_3}}\frac{\partial \boldsymbol{x_3}}{\partial \boldsymbol{x_2}}\frac{\partial \boldsymbol{x_2}}{\partial \boldsymbol{\theta_2}} \tag{2.43}$$

The derivative for each layer i.e., $\frac{\partial \boldsymbol{x_2}}{\partial \boldsymbol{\theta_2}}$ is dependent on the architecture of the layer, however as the structure of the network is known it is simply trivial to identify the needed formula and apply that to each derivative.

Optimiser

Once all the gradients are calculated, they are packaged into one vector which is the gradient of the loss with respect to the hyperparameters:

$$\frac{\partial Loss}{\partial \boldsymbol{\theta}} = \boldsymbol{g} \tag{2.44}$$

A variety of different optimisation methods such as Adam,[255] RMSProp[256] or Adagrad[257] are used to update the parameters. These all follow the general formula:

$$\theta_{t+1} = \theta_t - \Delta\boldsymbol{\theta_t} \tag{2.45}$$

where t is the time point, and θ are the parameters. $\Delta\boldsymbol{\theta_t}$ is the update vector to the parameters and can vary depending on the optimiser. In gradient descent, this is simply:

$$\Delta\boldsymbol{\theta_t} = \alpha \boldsymbol{g_t} \tag{2.46}$$

where $\boldsymbol{g_t}$ is the gradient obtained through backpropagation and α is the learning rate of the model. In the case of Adam, this update vector is much more complex as it includes both momentum and squared gradients of previous steps.

Momentum and squared gradients are calculated as follows:

$$\boldsymbol{m_t} = \beta_1 \boldsymbol{m_{t-1}} + (1 - \beta_1)\boldsymbol{g_t} \tag{2.47}$$

$$\boldsymbol{s_t} = \beta_2 \boldsymbol{s_{t-1}} + (1 - \beta_2)\boldsymbol{g_t}^{\boldsymbol{2}} \tag{2.48}$$

The update vector is then calculated as:

$$\Delta\boldsymbol{\theta_t} = \alpha_t \frac{1}{\sqrt{\boldsymbol{s_t}} + \epsilon} \boldsymbol{m_t} \tag{2.49}$$

where α_t is the value of α at time step t .Usually, the constants are set as $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-6}$, however, these can be changed if different convergence criteria as required. During training the learning rate can be adapted, typical variants are linear or exponential decay over the training cycles.

2.2.4 Transformer architecture

The transformer architecture was introduced in 2016 by Vaswani *et al.*[258] The concept is based on the idea that the context of a word in a sentence is critical to understand its meaning, and that context is not just limited to the preceding word. The architecture of a transformer network can be broken into three sections: an embedding layer, an attention layer and a feedforward layer.

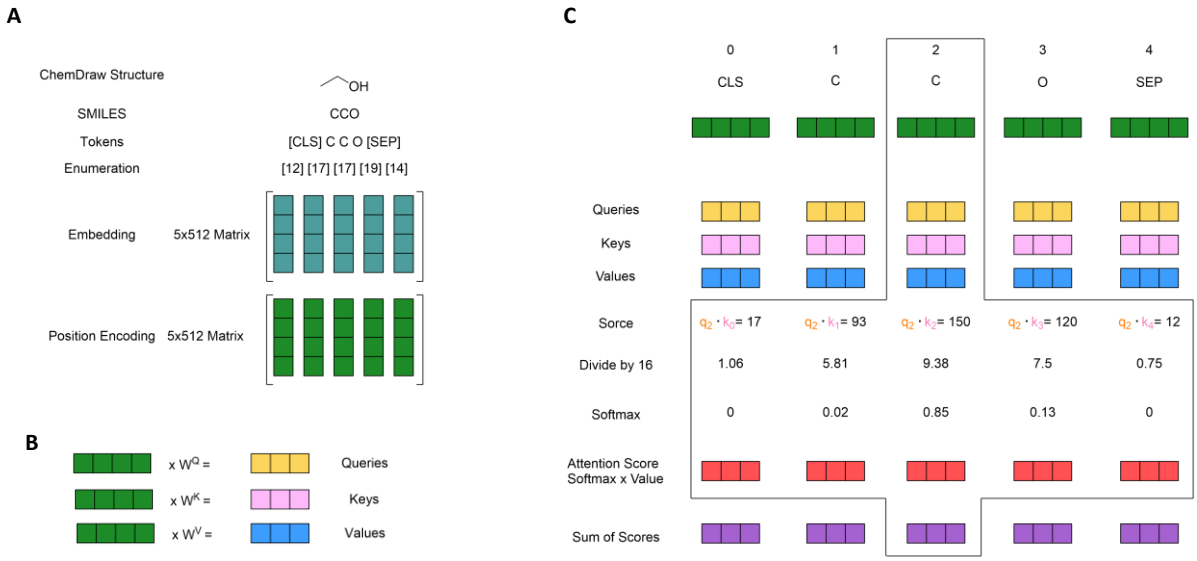


Figure 14 A) Tokenisation and embedding layer of the BERT model with ethanol as an example B) Learnt Weights for the queries, keys and values C) Attention mechanism shown for the secondary C atom in ethanol

In the embedding layer (Figure 14A), a series of operations are performed on each character in a string or word in a sentence to generate the embedded representation of the sentence. Firstly, the string is tokenised according to a predefined library, in our case SMILES characters. A start token (CLS) is added to the beginning of the SMILES string and a SEP Token to the end of the smiles string. Padding is also used for batch processing to ensure that all strings are the same length. The second step is enumeration where each token is converted into a numerical value. The next step is embedding where each token is converted into a vector of length M, this produces a matrix with a size NxM, where N is the length of the modified SMILES string. Finally, positional encoding is added to identify where each token is in the string.

The embedded string is then the input for the attention layer, which is the key architectural development in transformers. An attention layer generates three separate representations of each token during training: a query (**q**), a key (**k**) and a value (**v**) vector. These are learnt representations, meaning that during training three separate weight matrixes are trained to learn a general transformation from the embedded string into each of the three representations (Figure 14B).

For each token in a string the following process is used to generate the output of the attention layer. Firstly, the dot product of the token's query vector and every key vector is calculated, producing a score for each token (Figure 14C). The score is scaled through division by the root of length of the key vector (16) and then passed through a SoftMax layer to get a weighted importance between the query token and all others in the string. This SoftMax value is then multiplied by the value vector for each token and summed to generate the output of the attention layer for the input token. This process is repeated for each token in the string to generate the output of the attention layer.

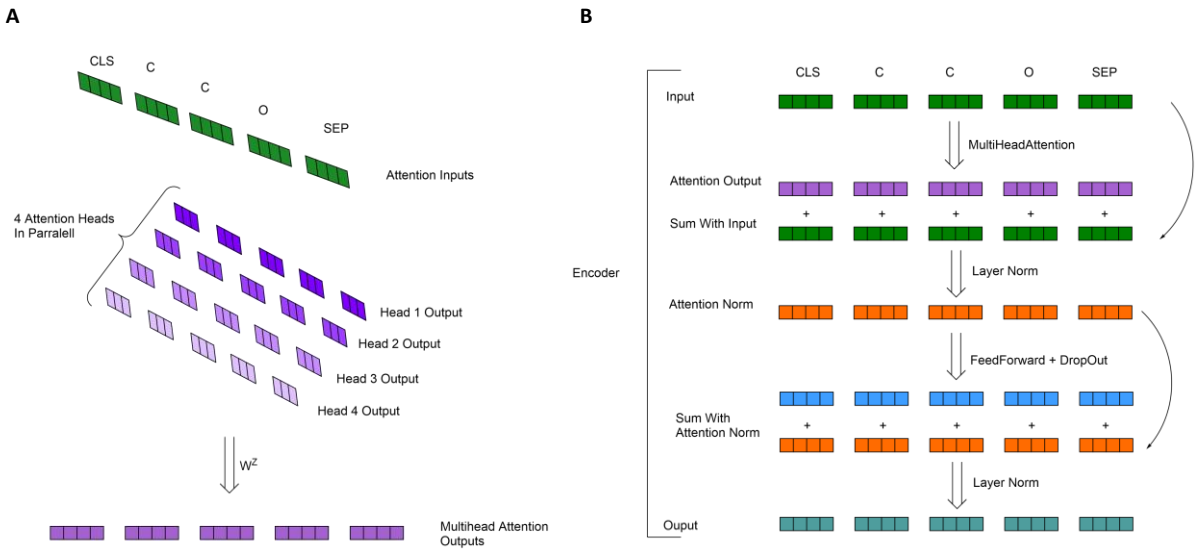


Figure 15 A) Multiheaded Attention in the transformer model with four heads B) Encoder network for the transformer model

This attention process is carried out four times in parallel with separate weights and learnt parameters, a process known as multi-head attention (Figure 15A). The output of these four heads is then transformed into the final output of the multiheaded attention layer by a final matrix, which again is learnt during training. A series of feed-forward neural network layers with both layer normalisation and drop out layers are then utilised to generate the output for the encoder block (Figure 15B).

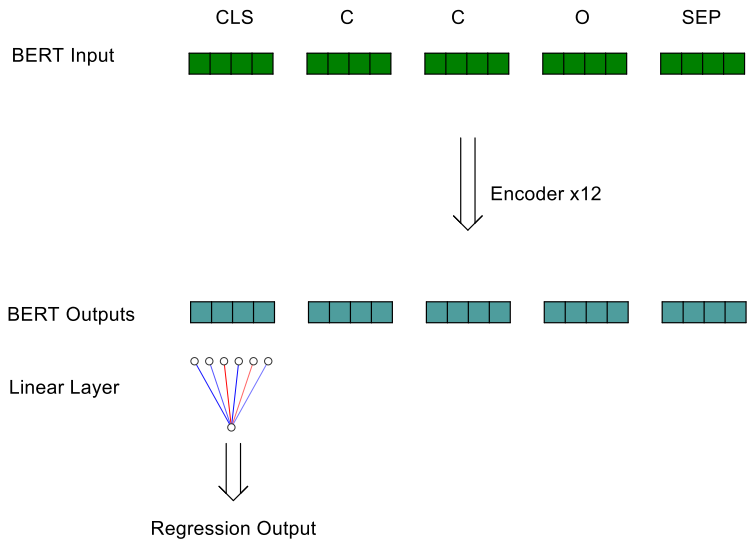


Figure 16 Overall Bert Architecture used for the prediction of Regression tasks

Overall, 12 encoders are used in a stack to get the final attended values for each token in the string (Figure 16). The attended output of the CLS token (placed at the start of the SMILES string) is then used as a global hidden representation for the complete structure. A simple linear feed-forward network is then used to generate the single linear regression output from the hidden representation of the CLS Token. A sigmoid layer can be used instead for the analogue classification problem.

2.2.5 Overfitting

When training an ML model, the data may be learned so efficiently that it has effectively memorised the data, wherein the model is very accurate on the training data but performs poorly on test data. This problem is known as overfitting. Reducing the tendency of a model to overfit is of great importance when trying to develop a model which can generalise to new, unseen data.

The simplest way to reduce overfitting is to use the least complex model possible, such as linear model, low-level polynomials, small decision trees or a small neural network. If the use of a small model, however, doesn't produce the level of accuracy needed, more sophisticated methods must be applied. The first type is regularisation, such as that applied in LASSO seen in section 2.2.2 which includes a penalty for any descriptors which are not set to zero.

During the training of neural networks, it is possible to include a dropout rate, where a percentage of the descriptors are set to zero at any given time. This reduces the chance that the model will become overly dependent on one particular descriptor and therefore reduces overfitting. Another option for neural networks is the use of early stopping, where a model stops training if the validation error increases while the training error continues to decrease.

For the reduction of overfitting in this thesis, we use ensemble methods where multiple models are trained on slightly different datasets and then averaged to reduce potential overfitting. To divide up the data, three methods were tested: cross-validation, data splitting, and shuffling.

Firstly, is the generation of a test set which is unseen during the training process and therefore is removed from the dataset. This can be carried out either through random selection or another criterion, such as scaffold type or Tanimoto similarity. This ensures no transfer of information between the model and test data throughout the training process (Figure 17A).

Figure 17 A) Dataset is firstly split into training and test sets either randomly or using another splitting strategy B) Training data is then divided into training and validation datasets either using cross-validation, data splitting or shuffling.

Cross-validation (CV) is one of the most common methods and is sometimes denoted as *k*-fold CV, where *k* is the number of folds. In the example shown in Figure 12A, *k* is equal to ten. In CV the training data is split into 10 folds, and then 10 models are trained on different combinations of 9 folds, where the remaining fold is an internal validation set (Figure 12A).

Data splits can be used in some cases where complete separation between each fold is required. In this case the model is trained on one-fold instead of training on nine folds where the remaining nine are validation datasets, therefore the opposite of CV.

Finally, data shuffling is when the data is shuffled multiple times and random selections are used to make 10 different training and validation sets.

In each of these cases, the final model will contain ten different models all trained on slightly different datasets, whose outputs are then averaged to give the final prediction from the model.

2.2.6 Quantification of regression models performance

To evaluate and compare different models there must be a method to understand the prediction error. One of the most common statistical measurements used in ML is the correlation coefficient R^2, which provides values between 0 and 1. A value of 0 means there is no correlation between the predicted values and the true values, and 1 means perfect correlation. R^2 is calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.50}$$

where y_i, $\hat{y}_i$ and $\bar{y}$ are the true value, the predicted value of the i dependent variable, and the average of the dependent variables. n is the number of data points.

Another common method of showing the model's predictive power is the error between the predicted output and the true output. The mean absolute error (MAE) is the average absolute difference between the predicted and true values (Equation 2.51).

$$MAE = \frac{\sum_{i=0}^{n}|y_i - \hat{y}_i|}{n} \tag{2.51}$$

Alongside the MAE two related measurements are also widely used. The mean square error (MSE) and the root mean square error (RMSE). The MAE measures the square difference between the predicted and true values while the RMSE is the root of the MSE.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{2.52}$$

$$RMSE = \sqrt{MSE} \tag{2.53}$$

2.2.7 Methods to Interpret Machine Learning Models

Linear Regression

Interpreting linear regression models is rather simple if we consider Equation 2.23:

The coefficients m multiplied by their dependent variable x will give the contribution of that variable to the overall predictor y. Therefore, understanding how the descriptor x changes as a function of chemical structure will allow for an interpretation of how important the model understands a change in structures to change the model's prediction.

Neural Networks

Understanding how and why a NN makes certain decisions is central for its application in science. Recent efforts have focused on techniques to investigate what a NN has learnt during training, including counterfactuals,[259, 260] adversarial examples,[261-263] Shapley additive explanations (SHAP),[264] anchoring[265] and integrated gradients (IGs).[266-268] These techniques can roughly be separated into three different categories: i) white box models, which aim to develop completely open and interpretable models, like Local Interpretable Model-agnostic Explanations (LIME)[269] ii) those which post-hoc explain the reasons behind a black box decision, like SHAP and IGs, iii) those which test the sensitivity of a model, like counterfactuals, adversarial examples and anchoring.

IGs in particular have become widely used in text-based NLP tasks to understand and interpret how words lead to certain outputs in NNs. IGs are based on the fundamental concept that the importance of a token or pixel is directly related to how much the output of an NN changes as you change that part of the input. In mathematical terms, the importance of a token is directly related to the gradient of the output with respect to the inputs, and large gradients are therefore indicative of an important input.

However, studying NN gradients is challenging due to saturation, which occurs when many small magnitude gradients have a large impact on the predictive properties, making them hard

to identify. IGs circumvent this problem by asking "how does the model go from predicting nothing (i.e. noise) to predicting the correct output?" The question can mathematically be formulated as a path integral from a baseline to the input. Therefore, the sum of the gradients along this path gives information on the importance of each part of the input:

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F\big(x' + \alpha(x - x')\big)}{\partial x_i} d\alpha \tag{2.54}$$

In the context of the Molecular Transformer (MT), Lee *et al.*[75] used IGs to explain what the MT was learning about chemical reactions. In the case of an S_N2 reaction with two possible reaction centres, IGs were used to show the importance of each. The Br leaving group was found to be more important than the Cl leaving group, which agrees with fundamental chemistry principles. This result highlights that MT has learned the importance of competing reactions to predict the correct product.

One issue with the use of IGs for the MT is that while canonicalized SMILES strings are a unique representation of a molecule, not all tokens are attributable to atoms. For example, tokens such as "1" or "2" are used to indicate ring closure points, but the sum over all the tokens in a functional group is taken to generate importance for the functional group as a whole. e.g all 8 tokens in a benzene ring. Therefore, any specific information about atoms in the ring is lost.

Chapter 3 Machine Learning and DFT as Tools to Aid in the Enantioselective Synthesis of Alkyl β-fluoroamines

3.1 Abstract

Computational Chemistry, in tandem with experimentation can both reduce the amount of time and screening needed to optimise a given reaction by screening reactants and identifying key interactions in the transition state. Machine Learning (ML) models are slowly becoming more used within organic chemistry for both the prediction of enantioselectivities and reaction yield. These models aim to identify key structural motifs which will lead to higher selectivity, or efficacy and therefore can result in better catalyst and substrate design.

In this work, we introduce a computational workflow to predict the enantiomeric excess (*ee*) of the hydrogen bonding phase transfer catalyst (HBPTC) reaction developed by the Gouverneur group. The workflow uses a Multivariate Linear Regression (MLR) Model, density functional theory (DFT) derived descriptors, and our own modified Sterimol values. Our model learns from a range of reactions on both epi-sulfonium and aziridinium data and can predict the $\Delta\Delta G^{\ddagger}$ of the reactions with errors below 1.1 kJ/mol for both internal and external validation.

We then apply this model to the prediction of enantioselectivity for the synthesis of alkyl β-fluoroamines, a class of chemical motif studied due to their prevalence in drug targets. We utilise this model to predict the enantioselectivity for a range of computationally generated substrates and catalysts with the aim of identifying possible future synthetic targets.

3.2 Chapter Overview

This chapter details the development and application of a workflow for the prediction of enantioselectivity for Hydrogen Bonding Phase Transfer Catalysis (HBPTC). This work was part of a joint project with Stamatina Zavitsanou. Python code for this project was developed by both of us, and DFT calculations for Model 2A were also performed by Stamatina. All other calculations and analyses reported here were carried out by me.

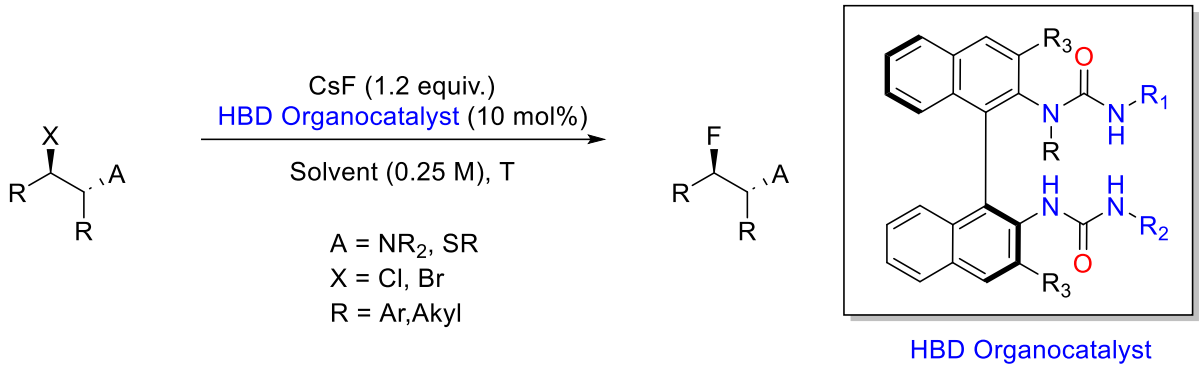


Figure 18 Overview of the HBPTC reaction developed by the Gouverneur Group, where A determines whether the reaction is for the epi-sulfonium or azaradinium substrate.

In the HBPTC reaction, the insoluble CsF salt is brought into solution by the HBPTC catalyst through coordination between the hydrogen bonds on the urea catalyst and the fluoride anion. This generates a soluble chiral fluoride reagent. At the same time, the substrate autoionises to form either the epi-sulfonium or aziridinium intermediate. The chiral fluoride can then attack the intermediate, opening up the 3-membered ring in an enantioselective manner, resulting in the final product.

Inspired by the previous work on the use of ML to predict $\Delta\Delta G^{\ddagger}$ and yield, as introduced in Chapter 2 we wanted to develop a workflow for the generation of descriptors using a cheap and efficient method for conformational sampling and descriptor calculation that could be used in tandem with experimental chemists to improve the selectivity for the HBPTC reaction. .

3.3 Workflow for molecular descriptors generation

The workflow developed in this project exists as a series of Python scripts that generate descriptors either from SMILES or from DFT calculations in ORCA. A brief overview of the methods for generation for different models is given below.

3.3.1 Fingerprints and Mordred descriptors

For the generation of molecular descriptors, both fingerprints and Mordred descriptors can be used within the provided environment. Both descriptors are calculated from .smi files, which are text files where each line contains the SMILES string for each substrate and catalyst obtained from ChemDraw. The final information that the user must give is a list of catalyst and substrate combinations that were experimentally tested, this is a .csv file in the order of catalysts, substrates, temperature, and solvent.

In the Fingerprints.py script, users choose both the size of the fingerprint radius and the overall length of the fingerprint. The script then generates fingerprints for both the Substrates and Catalysts before assembling the Reaction Matrix, which contains the fingerprint for each Substrate and Catalysts in the reaction. This Matrix can then be used for screening and training a range of Machine Learning algorithms.

In the Mordred.py script, the same three user-supplied files are used. The script generates 3D coordinates for each molecule before calculating the Mordred descriptors, and, as before, a reaction matrix is generated at the end of the script.

3.3.2 DFT derived descriptors

For descriptors derived from DFT calculations, two workflows are available for users, differing on the methods used to generate 3D structures. The first, referred to as "fitting to core" workflow, aligns a structure to a known core. The second one uses conformational

sampling using xTB. In both workflows, the Input.py script generates ORCA v4.2 input files from the corresponding .xyz files, allowing users to choose from a range of ORCA options, including optimisations, single points, basis sets, mixed basis sets on heteroatoms, DFT approximations, and NMR calculations.

Fitting to a core

In the "fitting to core" workflow users need to supply the SMILES string of the catalyst along with an .sdf file containing the 3D structure of the core to which the catalyst is being fitted (Figure 19). A crystal structure is an example of such a core. The FittoCore.py script aligns the SMILES string to the core and then generate the corresponding 3D structure. When the catalyst is bonded to an anion, such as fluoride or cyanide, the addF.py script adds the corresponding anion to the structure (Figure 19). Users need to supply coordinates for the anion, relative to its positioning in the core. This will then generate the completed hydrogen-bonded complex.

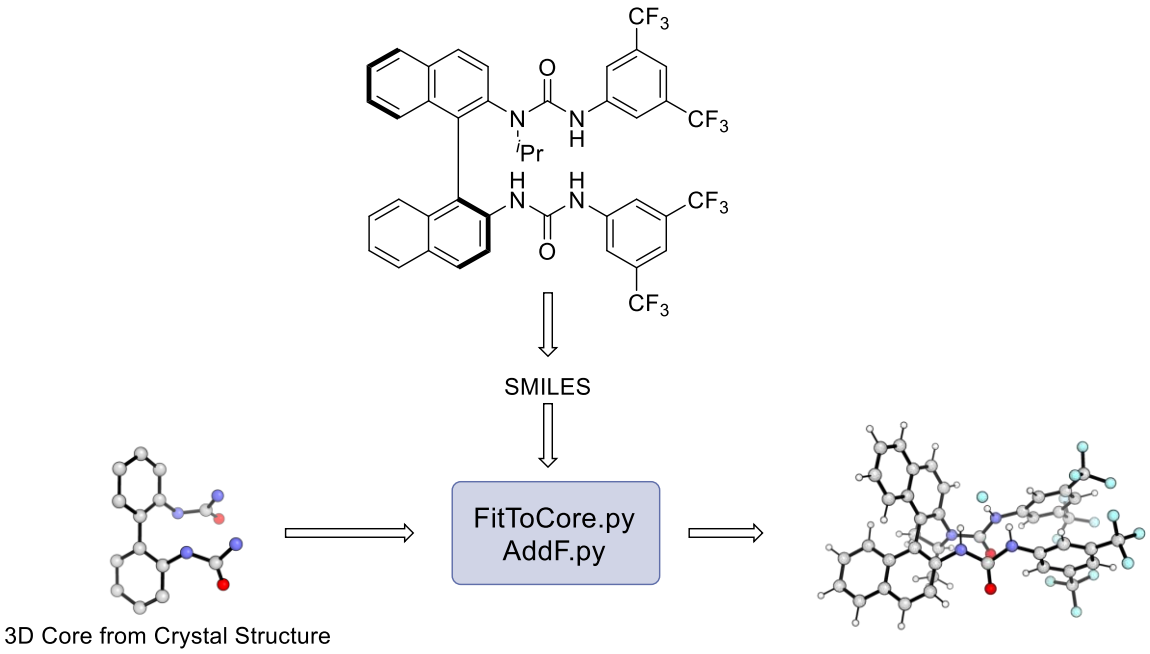


Figure 19 Workflow for the generation of 3D structures using RDKIT Fit to Core Function. Catalyst smiles are fitted to the crystal structure core before the fluoride is added using the AddF.py script.

Conformational sampling

The second workflow employs conformational sampling before DFT optimisations or single points to identify the lowest energy conformer of a catalyst anion complex. The sampling is carried out in the xTB software package[50, 131] using simulated annealing, while DFT calculations are carried out in ORCA v 4.2.[270] The xTB Simulated Annealing process produces 150 conformations for each complex, the script ExcludeandTransform.py is then used to process these conformers. First, it converts output from the xTB sampling (.coord files) to .xyz files, before calling accesibleF.py, which excludes conformers where the fluoride is not bound to the Hydrogen Bond Donors (HBDs), this step was added due to in some situations complex dissociation had occurred during conformational sampling. Finally, ExcludeandTransform.py runs an RMSD calculation using the autodE package,[271] with a cut-off of 1 Å on heavy atoms only (Figure 20). Each step generates a new folder where the accepted conformers are copied, i.e. exclude and rmsd.

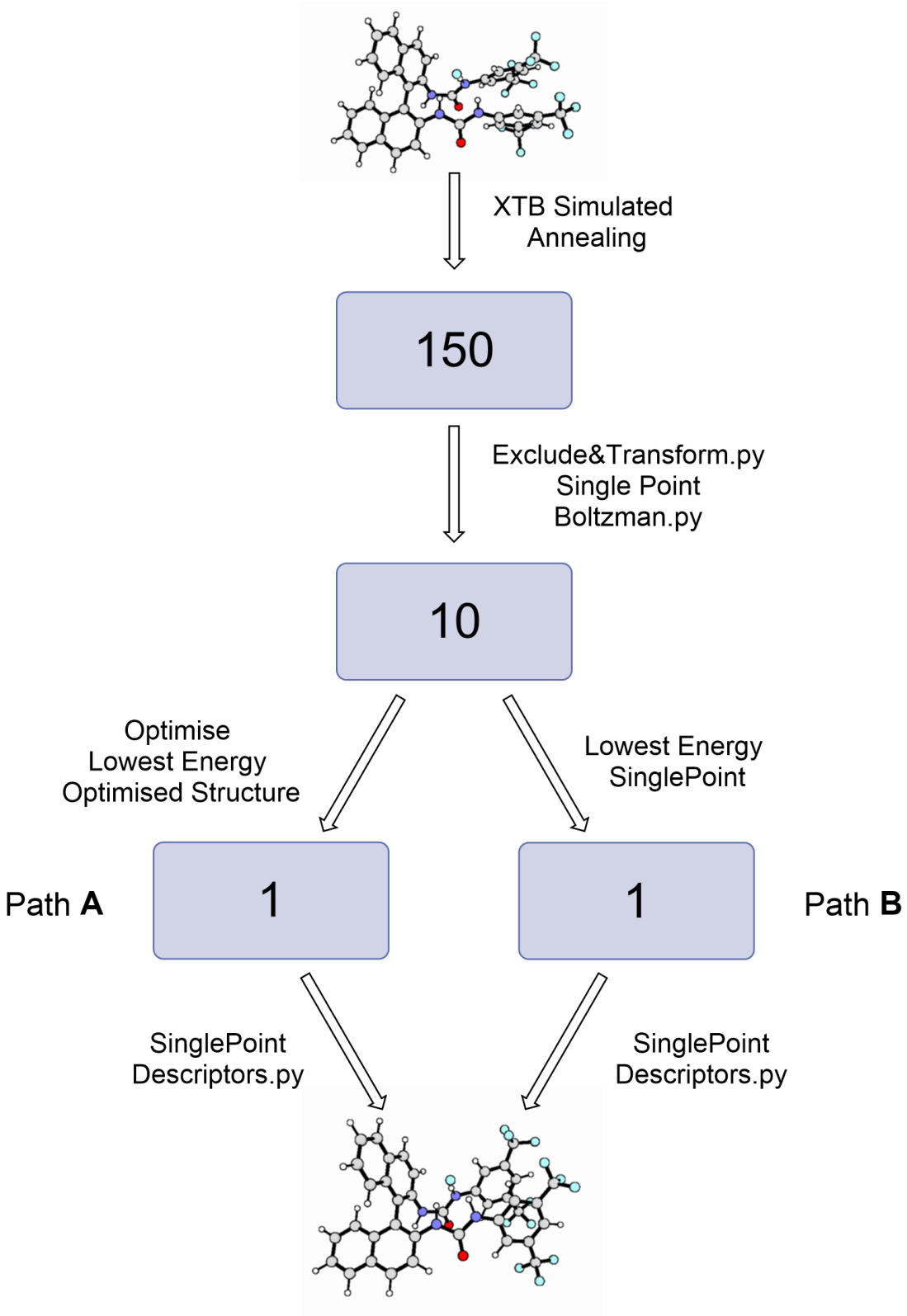


Figure 20 Workflow for the generation of catalyst and substrate descriptors using xTB simulations and ORCA calculations. Each number shows an example number of conformers generated. Two different paths are available where one includes extra DFT optimisations

For the remaining conformers a Single Point (SP) energy calculation is performed at the level of theory chosen by the user. Post DFT single point calculation, the Boltzman.py script is called where the contribution for each conformer is weighted using a Boltzmann distribution at 298 K, and those which sum up to 90% are transferred to the Boltzmann folder for further calculations.

At this point, users choose between two different paths; Path **A** takes the Boltzmann weighted complexes and performs a geometry optimisation, after which the LowestEnergy.py script selects the lowest energy conformer. A final single-point energy calculation is then performed on the lowest energy conformer for each catalyst that including the calculation of properties such as NMR shifts or atomic charges. In Path **B** the lowest energy conformer from the xTB simulation, ranked using DFT single points, is used directly, without further geometry optimisations. In an analogous method to Path **A**, the lowest energy conformer for each complex is subjected to a final single-point energy calculation to generate DFT descriptors.

The choice between Path **A** and Path **B** was created as we were interested if DFT optimisations were needed to generate a predictive model. The reduction in computational time and expense from Path B would therefore making screening future catalysts significantly faster.

Descriptor extraction

After DFT calculations have been performed the calculated descriptors were extracted from the ORCA output files using the descriptorsfree.py script. This script extracts, HOMO and LUMO energies, dipole moment, ^{1}H, ^{13}C, and ^{19}F NMR shifts for the urea moiety and the fluoride ion, bond order (BO) of the hydrogen-fluoride hydrogen bonds (BO_{HF}), BO of the urea N-H bonds (BO_{NH}), and BO of the urea N-C atoms (BO_{NC}) and atomic charges of the fluoride ion and the urea hydrogens (Figure 21). The script then identifies the type of core, i.e. mono-

and di-urea catalysts, and aligns them to the core before extracting the descriptors in an ordered manner to ensure that each column contains the same descriptor no matter the catalyst structure.

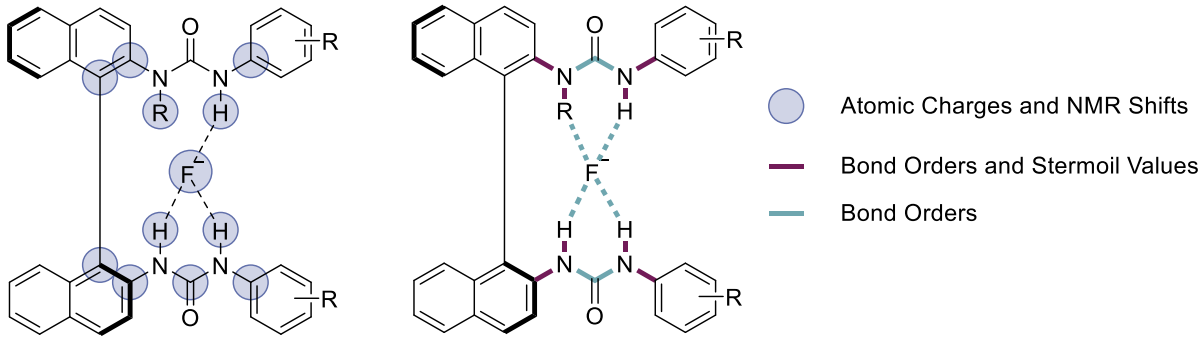


Figure 21 Illustration of the DFT descriptors calculated and used in ML models.

Sterimol generation

Sterimol descriptors, used to represent sterics, were computed using the Sterimol.py script. This is a modified version of the CalcSterimol function obtained from wSterimol developed by Paton *et al.*[272] Users need to provide a text file identifying which atoms to calculate the Sterimol values for, along with the structure of the molecule, the Descriptors.py script generates the atom lists automatically from the structure. This script then performs the calculations for each bond selected in the molecule before generating a .csv file that can be appended to the DFT descriptors.

3.3.3 Reaction matrix generation

With descriptors for both substrates and catalysts obtained from either fingerprint, Mordred Descriptors, or DFT calculations, the final step is to generate a reaction matrix. For this, the user needs two different sets of descriptors and a list of reaction combinations. The ReactionMatrix.py script will generate the final matrix with the substrates and catalysts matched to the experimental outcomes which can then be used in any of the ML workflows.

3.3.4 ML Screening

To screen different ML algorithms, we designed a Python script to identify the most suitable ML algorithms/parameters (see Appendix 1 for full details of the 15 combinations tested). The RegressionScreening.py script is only for screening, further hyperparameter optimisation should be carried out to get the optimal model. Users only need to provide matrix which contains the descriptors and the experimental $\Delta\Delta G^{\ddagger}$. The number of cross-validation folds is a parameter set by the user, with the default being 10. A plot of the cross-validation error and overall accuracy is generated along with a .csv of the errors.

3.4 Results & Discussion

3.4.1 Model development

Reaction matrix generation

From a selection of literature[16, 38] and unpublished data from the Gouverneur group, 79 catalysts and 32 substrates were extracted which, when combined with experimentally determined enantioselectivity, resulted in a dataset of 258 reactions. Below an overview of the generation of each reaction matrix is provided, with a full detailed description available in the Section 3.5.3 **Model generation**. Five datasets were generated which differ in the type of descriptors used (Table 2). The simple ones are Reaction Matrix **1A** and **1B**, which use fingerprint and Mordred descriptors, respectively. Reaction Matrix **2A** uses fingerprints for substrate descriptors and DFT descriptors obtained using the FittoCore Method. Reaction Matrix **2B** uses fingerprint descriptors for substrates and catalyst descriptors obtained from conformational sampling using Path **A**. Reaction Matrix **3**, **4,** and **5** all use substrate DFT descriptors using Path **B**. Reaction Matrix **3** and **5** have catalyst descriptors obtained using Path **A** but with different levels of theory, while Reaction Matrix **4** uses descriptors obtained from Path **B.**

Table 2 Descriptors used in the generate of each Reaction Matrix described in this chapter. I, II, III refers to each level of DFT theory used, while Solvent model indicates if implicit solvent was used during DFT calculations. Complex generation specifies the methods by which complex geomertries were obtained. All DFT substrate calculations were performed using Path **B.**

Reaction Matrix	*Substrates*	*Catalyst-F*	*Solvent model*	*Complex generation*	*Descriptors*
1a	Fingerprints	Fingerprints	N/A	N/A	Fingerprints
1b	Mordred	Mordred	N/A	N/A	Mordred
2a	Fingerprints	**I**	Yes	Fit to core	Fingerprints / DFT + Sterimol
2b	Fingerprints	**I**	Yes	Conformational search Path **A**	Fingerprints / DFT + Sterimol
3	**II**	**I**	Yes	Conformational search Path **A**	DFT + Sterimol
4	**II**	**II**	No	Conformational search Path **B**	DFT + Sterimol
5	**II**	**III**	Yes	Conformational search Path **A**	DFT + Sterimol
I	PBE-D3BJ/def-TZVP//PBE-D3BJ/def2-SVP				
II	PBE-D3BJ/def2-SVP				
III	PBE-D3BJ/ma-def2-TZVP//PBE-D3BJ/ma-def2-SVP				

Visualisation of Sterimol descriptors.

Sterimol parameters represent steric effects through three parameters B_1, B_5, and L.[236] For a given bond X-Y (Figure 22A) B_1 represents the shortest distance perpendicular from the primary axis of attachment (minimum width of a substituent). B_5 represents the longest distance (maximum width of a substituent). L is the total distance following the primary axis of attachment (length).

Despite Sterimol descriptions being widely used in a range of MLR models, as introduced in Chapter 2, we saw important differences in their practical calculations, for instance, Sigman[236] creates tables of values for a range of isolated R groups using an in-house code. While Paton *et al,*[272] compute Sterimol values on a conformational ensemble of the isolated R group capped with an H or CH_3 meaning that directionality is not needed.

However, when the same R group is present on a catalyst or molecule the remaining molecule will influence the conformation of the R group, therefore a method that includes these effects could lead to a more accurate description of steric bulk.

To visualise which atoms, give rise to each of the Sterimol descriptors, we developed a Python script that can be used within Pymol to visualise the vectors and atoms. For the L vector, the direction and length of the descriptor are shown (red arrow Figure 22B, while for the B1 and B5 (yellow and blue arrow Figure 22B) descriptor the vector points to the atom from which the B1 or B5 value is calculated.

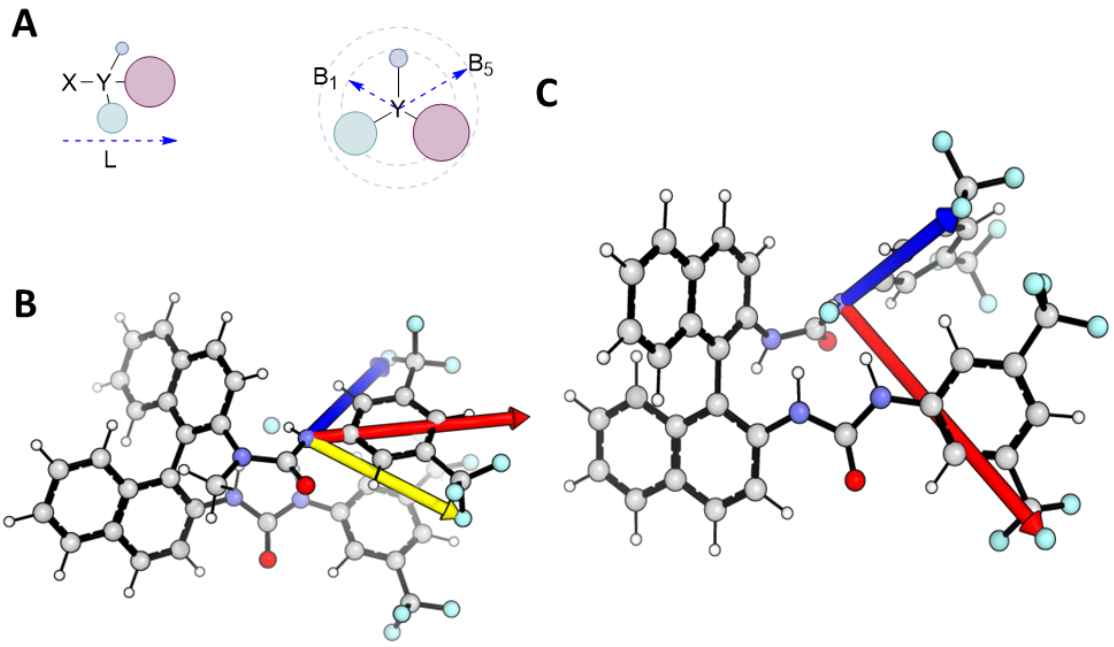


Figure 22 A) Illustration of the Sterimol parameters. B1 and B5 are the minimum and maximum widths of the group when viewed in profile looking down the primary axis and L is the total length along the same axis illustrated. B) Visualisation of Sterimol values, red arrow shows L value, blue arrow points to the atom responsible for B5, and the yellow arrow points to the atom responsible for B1. C) Visualisation of the atom from which the Sterimol B5 values are calculated from. The red arrow shows the atom identified by the unmodified Sterimol code while the blue arrow shows the atom expected to be identified.

However, during testing of our visualisation method, we observed some unusually large values, 4 or 5 Å larger than expected. When we inspected the atoms identified by the Paton code for catalyst **1** (Figure 22C), the expected B5 descriptor N−C bond would be the CF_3 groups (blue arrow) however we found it was giving the distance to the CF_3 on the other side of the complex (red arrow).

After studying the code that calculates the Sterimol values it was noticed that the method was calculating the values between the nitrogen of interest and all the atoms in the molecule before identifying the longest and shortest values to generate the Sterimol descriptors. This results in

the directionality of the descriptor not being preserved, we therefore needed to both identify a way to enforce directionality and evaluate if it was needed.

Does molecular conformation affect Sterimol values?

To answer this question, we first needed to develop a method that would calculate Sterimol values in an intact molecule and compare those to ones obtained from Sigman *et al.*[236] To enforce directionality, graph networks were utilised to determine which atoms should be considered for calculation. Fortuitously, the Python script developed by Paton already extracts the atoms and bonds from the input file, enabling the generation of a graph where atoms are represented as nodes and bonds edges. As the bond of interest from where the Sterimol values are calculated (atoms X−Y) is known, deleting the edge between them results in two subgraphs. The subgraph which containing atom Y, all atoms are extracted as a list and the calculations for L, B1, and B5 are performed between atom X and those atoms (Figure 23A). This method is referred to as *graphSterimol*. The original, unmodified code which does not preserve directionality is referred to a *unSterimol*.

To test the implementation and effectiveness of *unSterimol* and *graphSterimol* against the work reported by Sigman *et al*[236] two case studies were chosen: the desymmetrisation of bisphenols and the Nozaki–Hiyama–Kishi (NHK) allylation of carbonyls.

The values obtained from graphSterimol vs those from Sigman are mostly similar, with differences < 0.1 Å. However, values > 0.3 Å are obtained in some cases. For example, in the first case study, the L value changes from 5.1 Å to 6.1 Å for the $CHPh_2$ group (Figure 23B), when bisphenol is present, the two phenyl groups are closer together, which results in the L vector being significantly longer than in the free $RCHPh_2$.

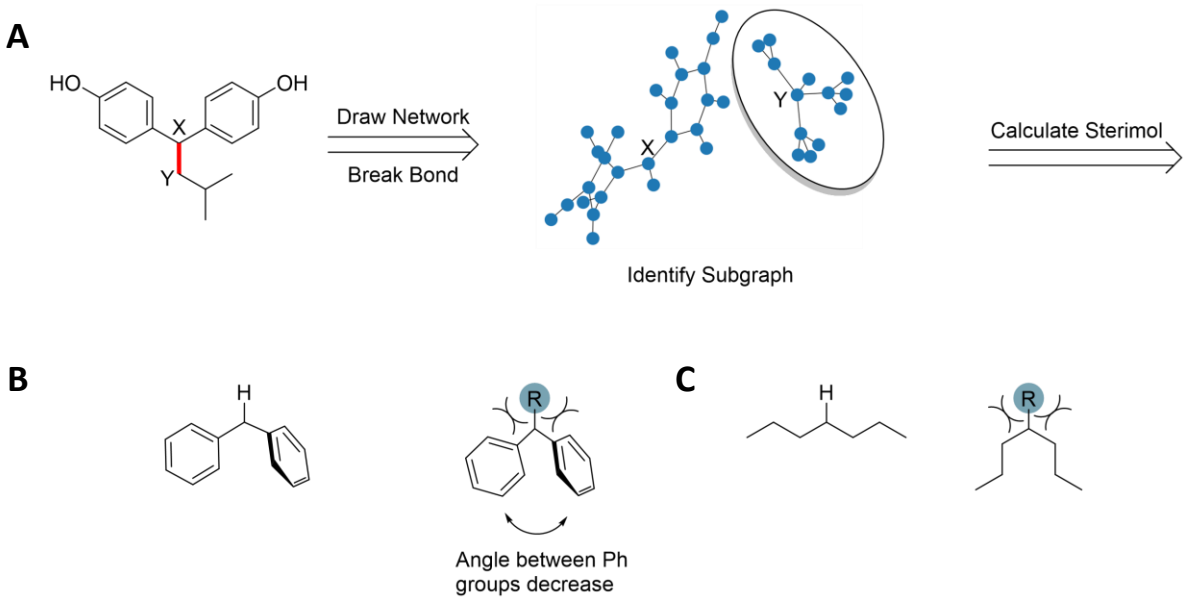


Figure 23 A) Method for identifying valid atoms. A graph is created from the 3D coordinates and the X-Y bond, highlighted in red, is broken. A subgraph search is used to identify which subgraph contains the Y atom. Atoms in the same subgraph as atom Y are then passed to the wSterimol code to calculate the Sterimol values. B) Effect of substitution R on the $RCHPh_2$ group (R= bisphenol) C) Comparing the conformation for the 4-heptane group, when R = H the conformation is linear, but when R is the NHK catalyst the conformer is syn. This significantly changes the Sterimol values.

The second example is for the 4-heptane group in the NHK reaction (Figure 23C). When optimised in the absence of the catalyst the 4-heptane group is a linear chain, however when the catalyst is present the two Pr groups move to be *syn* to each other. This results in L increasing by 1.97 Å, B1 by 0.16 Å and B5 by 1.14 Å. These examples indicate that the nature of the R group leads to subtle differences in the obtained descriptors. How this affects the model is discussed in the following section.

Does graphSterimol give a better representation of steric bulk

To understand if graphSterimol can give a better representation of steric effects in the prediction of enantioselectivity, we compared models generated with descriptors obtained from Sigman *et al.*,[236] *unSterimol* and *graphSterimol*.

For the desymmetrisation of bisphenols (Figure 24) *unSterimol* has a poor correlation with the enantioselectivity (R^2 = 0.59, RMSE = 0.31 kcal/mol), however when the values obtained by Sigman *et al.*[236] are used for excellent correlation is obtained (R^2 = 0.95, RMSE = 0.11 kcal/mol). When *graphSterimol* is utilised a similar level of accuracy as Sigman was achieved (R^2 = 0.94, RMSE = 0.125 kcal/mol). This confirms that the directionality of the Sterimol values is crucial in the development of an accurate model.

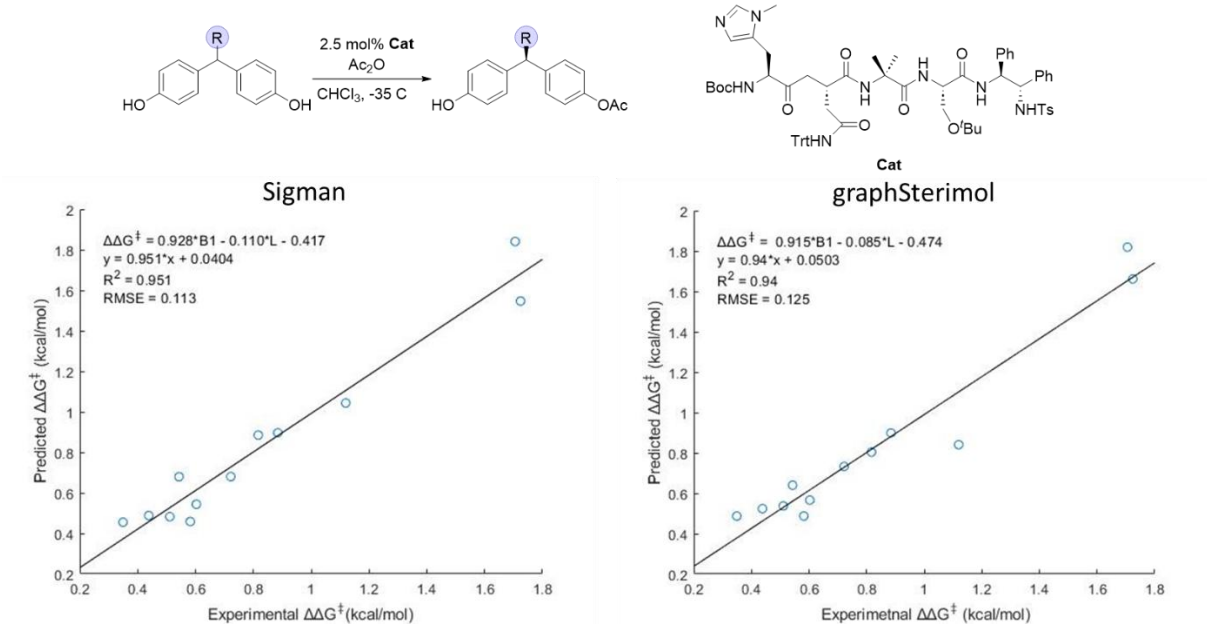


Figure 24 Comparison between different methods to calculate Sterimol values in the model for desymmetrisation of bisphenols. The Sigman model uses data reported in the original paper, while graphSterimol uses our method introduced above. Groups which are changed on the bisphenol is shown in blue.

In the NHK reaction, including directionality is also important, as not doing so results in MATLAB being unable to generate a model due to with no correlation with $\Delta\Delta G^{\ddagger}$. Comparing *graphSterimol* to that of Sigman results in a small improvement over Sigman's model in terms

of error 0.05 kcal/mol and R^2 (0.94 vs 0.882) (Figure 25). This highlights that including the whole molecule during conformational sampling is important, as both steric and electronic factors can impact the lowest energy conformation of each R group's position. These changes in position due to the rest of the molecule are something not considered in the models developed by Sigman or Paton, furthermore, this allows for the calculation of Sterimol values from whole molecules without the need for splitting up the molecule. For this reason, we utilise *graphSterimol* for the calculation of all Sterimol values.

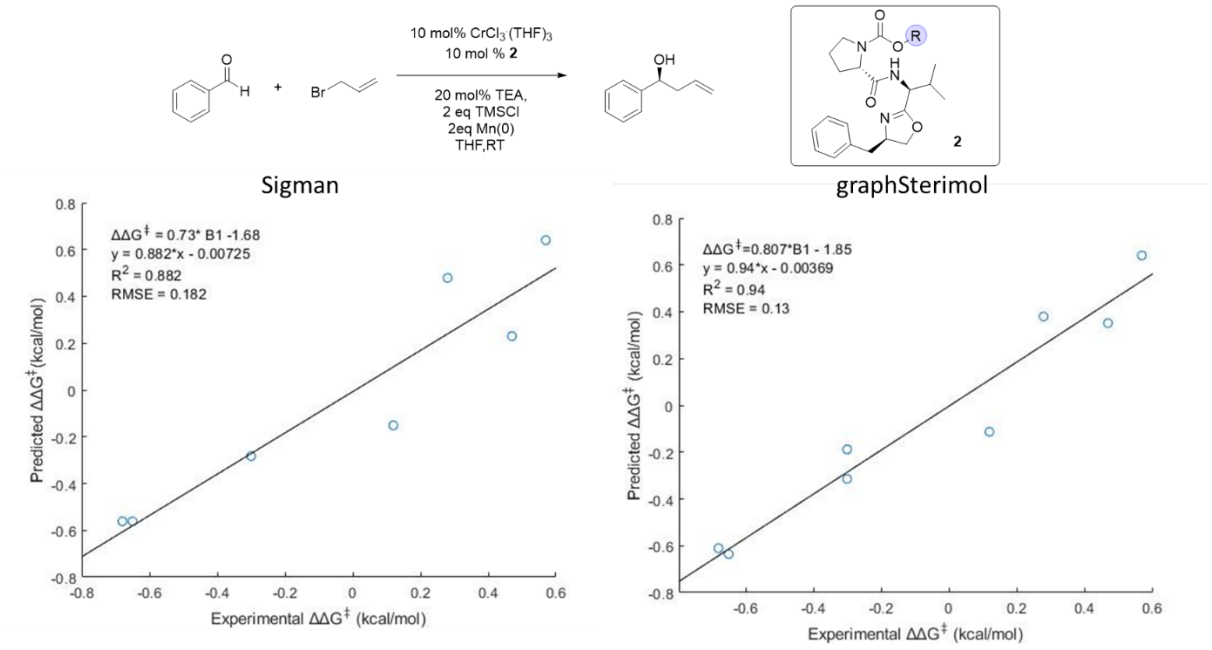


Figure 25 Comparison between different methods to calculate Sterimol values in the model for the NHK reaction. The Sigman model uses data reported in the original paper, while graphSterimol uses our method introduced above. R group which is changed is highlighted in blue on the catalyst

3.4.2 Screening machine learning algorithms

Identifying the best-performing ML algorithm is key to creating an accurate and usable ML model. In this work, different ML algorithms and their respective set hyperparameters were screened using Reaction Matrix **1A**, including Linear, Support Vector Machines (SVM), Decision Trees, K Nearest Neighbours, and Gaussian Processes.

Errors in the models varied from 0.8 kJ/mol to 1.7 kJ/mol (Figure 26), with LASSO, Linear SVMs, and Ensemble Decision Trees, resulting in the lowest errors, between 0.8 and 0.9 kJ/mol.

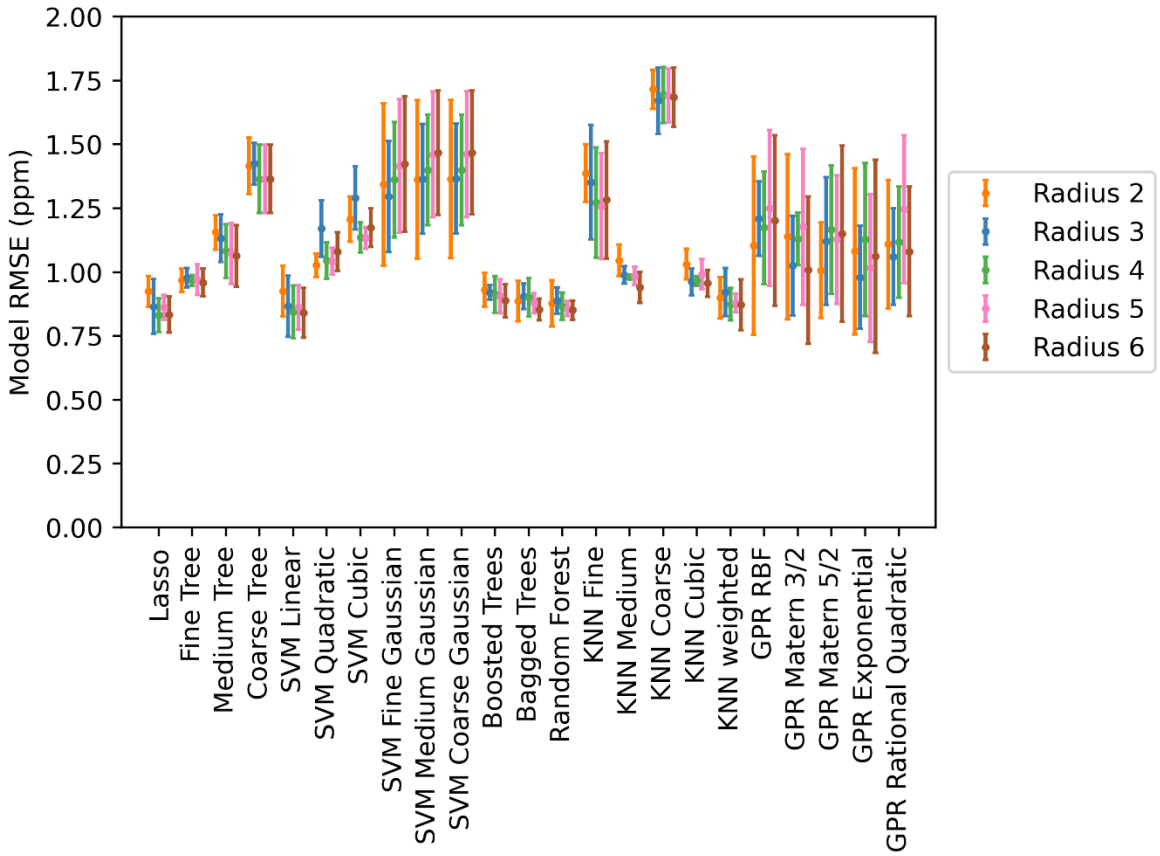


Figure 26 Screening of a variety of ML models using Morgan Fingerprints while varying the size of the fingerprints. Each model is trained with a 10-fold Cross-Validation. The error of the ensemble model is shown, while the error bars show the Cross-validation error.

As expected, increasing the radius of the fingerprints reduced the errors, for a Linear Model changing from radius from 2 to 6, decreased the error from 0.92 to 0.83 kJ/mol (Figure 26); however, changes from radius 4 to 6 only lead to a 0.03 kJ/mol decrease, suggesting radius of 4 being optimal. Increasing the length of each fingerprint or changing to RDKIT fingerprints did not improve the model.

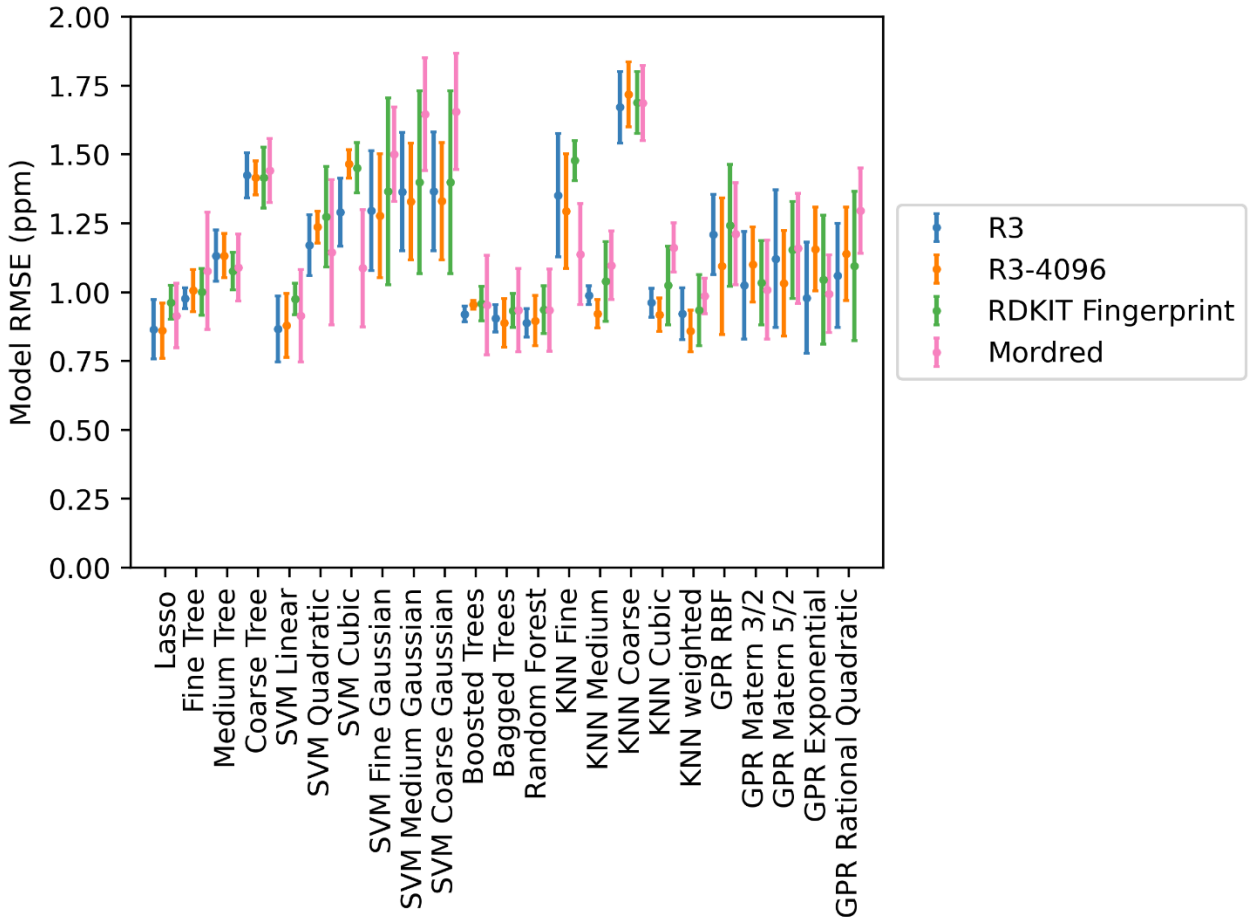


Figure 27 Comparison between model errors (RMSE) while varying the type of 1D descriptor used. R3 indicates a Morgan fingerprint with a radius of 3 and a length of 2048 bits, while R3-4096 is the same radius just a 4096-bit vector. RDKITFingerprints and Mordred, indicate the respective descriptors. Each model is trained with a 10-fold cross-validation. The error of the ensemble model is shown, while the error bars show the cross-validation error.

Molecular descriptors perform worse than fingerprints, with the best method obtaining an error of 0.93 kJ/mol. Again, Linear, Linear SVM, and Ensemble Decision Trees are the best performing architectures.

Considering the large number of descriptors selected by LASSO regression (> 50), we focused on Linear models as they are the least likely to overfit in the case of a large descriptor: dataset ratio.

Comparing model performance

With seven different reaction matrices in hand, as described in Section 3.4.1 , LASSO regression was used to provide a model for each dataset. Overall, all models perform equally well, with differences in $R^2 > 0.6$ and RMSE < 1.1 kJ/mol across all models on external test sets (Table 3).

Surprisingly Model **1A**, which only used fingerprints for both catalyst and substrate descriptors, performed as well as models based on electronic descriptors, such as Model **3** and **4**. However, while fingerprints can provide fast and accurate predictions, they lack interpretability. Furthermore, when predictions are made on unseen test data, this model performs in a bimodal manner, with either small errors of < 0.2 kJ/mol or large errors of > 1.3 kJ/mol, suggesting the model can interpolate but it is unable to extrapolate to new chemical groups being tested.

Table 3 Model accuracies for ML models using LASSO Regression. Train and Test errors are repeated 100 times and averaged. All RMSE is the Cross-Validation error over a 10-fold CV with no external test set and given in kJ/mol

	RMSEtrain (kJ/mol)	*R2train*	*RMSEtest (kJ/mol)*	*R2test*	*AllRMSE (kJ/mol)*	*AllR2*
*Model **1A***	0.78	0.83	0.96	0.72	0.77	0.84
*Model **1B***	0.81	0.82	1.02	0.67	0.78	0.83
*Model **2A***	0.89	0.78	1.00	0.68	0.82	0.81
*Model **2B***	0.87	0.79	1.04	0.65	0.85	0.80
*Model **3***	0.90	0.78	1.03	0.65	1.00	0.72
*Model **4***	0.95	0.75	1.05	0.67	1.00	0.72
*Model **5***	0.90	0.77	1.14	0.53	0.99	0.73

Models containing DFT descriptors **(**Model **3**,**4** and **5**) all show similar levels of error with RMSE's between 0.90 and 0.95 kJ/mol, however given Model **4** is less computationally intensive, as it uses a lower level of theory to obtain descriptors it was therefore chosen as the optimum model.

Using Model 4 for descriptors and LASSO regression results in both the training and test providing a good correlation and low error (Figure 28) The outliers observed for the experimental $\Delta\Delta G^{\ddagger} = 0$ are data points that represent reactions that either give low *ee* or have been tested only once and with a variety of temperatures. Outliers observed at higher $\Delta\Delta G^{\ddagger}$ values, are probably due to the lack of experimental data in that region (*ee* > 70%), for aziridinium substrates.

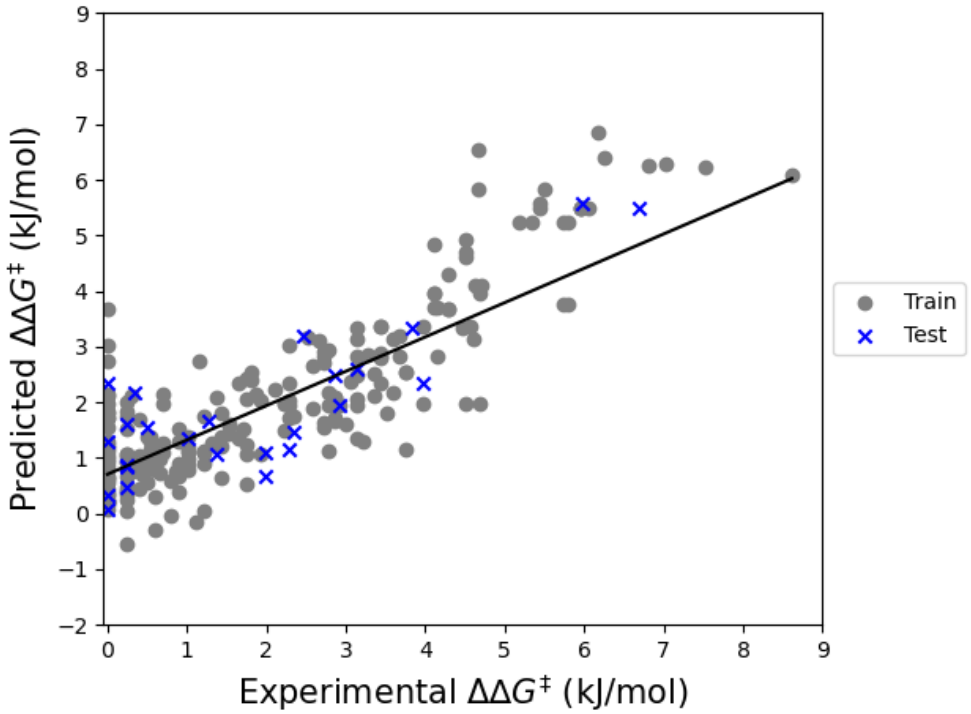


Figure 28 Multivariable linear regression model for one of the 100 runs for Model 4. In grey the data points used as a training set and in blue the data points used as a test set.

Can DFT descriptors and LASSO regression predict on unseen dataset

To validate the model's predicting power, both internal *10-fold-cross validation*, and external validation using 10% of the data selected at random were used. Additionally, data for a further eight reactions using new catalysts were obtained as a further external test set (Figure 29A). Predicting these new catalysts resulted in an $R^2 = 0.59$ and RMSE = 2.1 kJ/mol, with two data (Catalyst **80** and **81,** Figure 29B), yielding a low correlation. This is likely due to the fact that these structures have alkylation patterns not seen in the training data. The model is, however, able to extrapolate into areas where other modifications are taken on the catalyst backbone. When those two data points are disregarded the accuracy of the model becomes excellent with $R^2 = 0.96$ and RMSE = 0.80 kJ/mol.

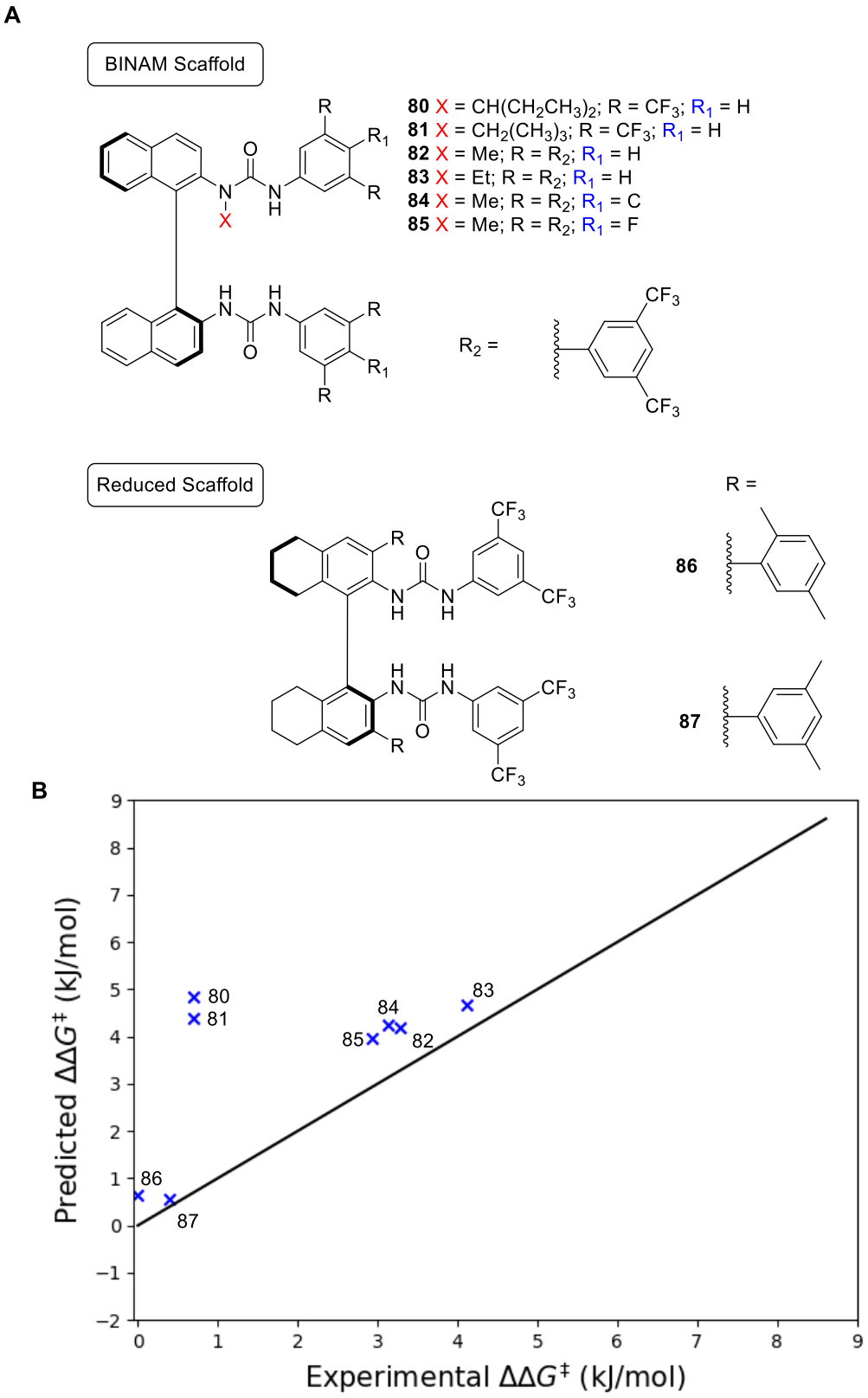


Figure 29 A) The eight extra catalysts which were utilised as an external test set due to their synthesis being carried out after the generation of the initial model B) Prediction of enantioselectivity using the MLR model. Predictions are labelled with the catalyst number.

3.4.3 Model interpretation

An advantage to using a MLR model is that the unitless coefficients can be analysed directly to understand which descriptors result in high selectivity. These can then be used to aid future experimental design.

$$\begin{aligned}\Delta\Delta G^{\ddagger} = -0.11 + 3.22BO_{HF1} + 1.25\text{Solvent} + 0.63Cat_{NH2(L)} \\ +0.55Sub_{YC1(B1)} + 0.36Cat_{NH2(B1)} + 0.32Cat_{NC1(B1)} + 0.12Cat_{NC3(B5)} \\ +0.04Sub_{YC1(L)} + 0.03Sub_{YC2(L)} + 0.008NMR_{C5} - 0.004NMR_{C3} \\ -0.008NMR_{C5} - 0.05Cat_{NC3(B1)} - 0.06Cat_{NC2(B1)} - 0.14Sub_{LUMO} \\ -0.28Sub_{HOMO} - 0.97BO_{\text{NH2}} - 4.96BO_{NC4}\end{aligned} \quad (3.1)$$

Table 4 Overview of descriptors selected by the MLR model, where X specifies the bond, the descriptor is from

Descriptor	*Meaning*
BO	Mayer bond order between two atoms
$Solvent$	Binary value representing solvent type
NMR	Calculated isotropic shielding constant
Sub_{HOMO}	Substrate HOMO energy
Sub_{LUMO}	Substrate LUMO energy
Cat_X	Catalyst Stermoil Descriptors for bond X
Sub_X	Substrate Stermoil Descriptors for bond X

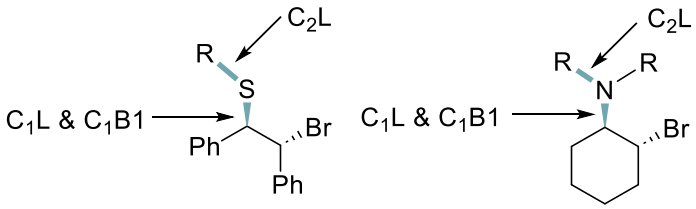
Equation 3.1 correspond to the prediction of reaction enantioselectivity obtained using descriptors from Model **4**. Given the linear nature of the ML model, the equation can be separated into the descriptors that are obtained from substrates or catalysts and analysed to identify their contributions to the selectivity. To identify the contribution from each term we then multiply the descriptor by the coefficient and then compare the contributions between similar molecules (Figure 30).

Firstly, the coefficient for the $Solvent$ agrees with the optimised experimental conditions reported by the Gouverneur Group which showed an improvement in the selectivity of 1 kJ/mol for switching solvents from CH_2Cl_2 to 1,2-difluorobenzene

In this model, the HOMO, LUMO and the steric descriptors L and B1 around the C1 carbons of the substrate are the most important. The HOMO enables to differentiate between the aziridinium and epi-sulfonium species (Figure 32A), the latter is on average lower by 0.5 eV, between -5.5 eV to -4.9 eV, than the aziridinium substrates (HOMO is between -4.8 eV to -4.5 eV). According to the model the difference in HOMO alone between the best-performing of epi-sulfonium and aziridinium substrates can contribute up to an extra 0.2 kJ/mol to the $\Delta\Delta G^{\ddagger}$.

.

Substrate selectivity



Substrate Homo & Lumo

$$\Delta\Delta G^{\ddagger}_{Sub} = 0.55Sub_{YC1(B1)} + 0.04Sub_{YC1(L)} + 0.03Sub_{YC2(L)} - 0.14Sub_{LUMO} - 0.28Sub_{HOMO} \quad (3.2)$$

Figure 30 Descriptors selected for the substrate are both shown with their coefficients in the MLR, Y is either N or S

The substrate LUMO relates to whether the backbone of the substrate is either aliphatic or aromatic. Stilbene substrates have a lower LUMO that their cyclohexane counterparts. With a negative coefficient, more negative LUMOs give higher contributions to the predicted $\Delta\Delta G^{\ddagger}$ (Figure 31A). The substrate steric descriptors L and B1 along the Y- C1 (where Y is N or S) carbon are considered important for selectivity (Figure 31B).Analysis of the L Sterimol values

show that naphthyl and phenyl groups have larger values than cyclohexyl or cyclopentyl groups. In the latter, ring flexibility results in a shorter L vector; the loss of the two linking carbons further reduces the L value.

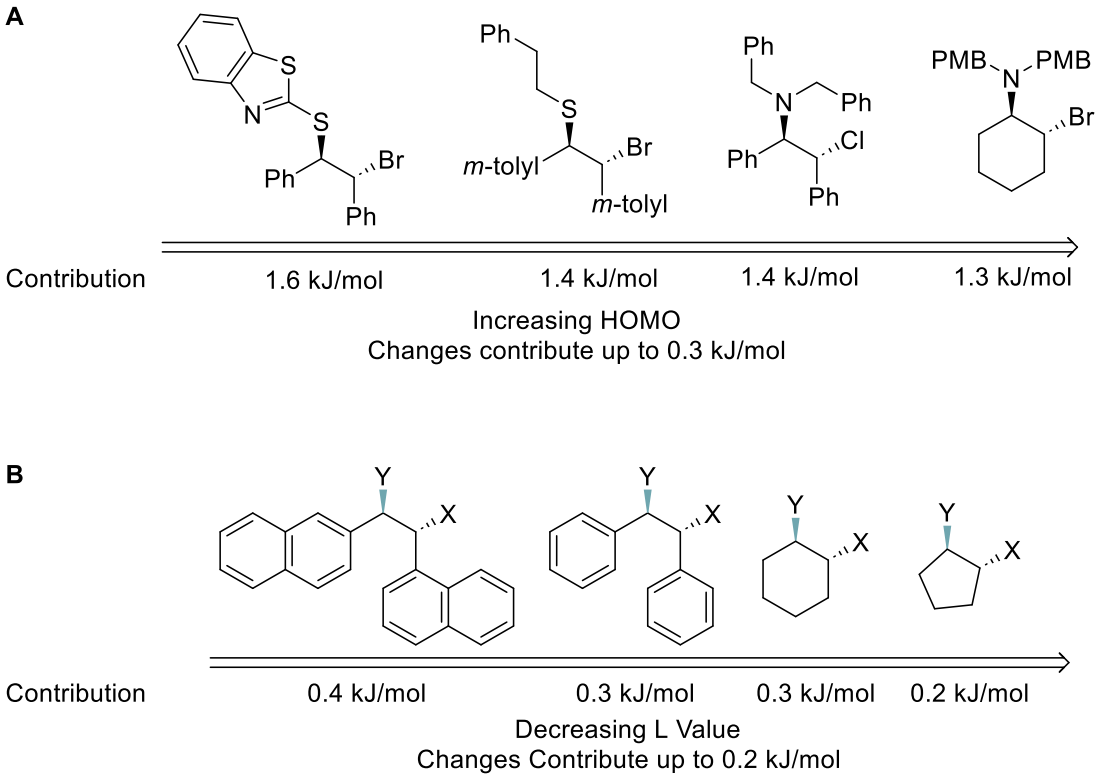


Figure 31 Contributions from the substrates as groups are varied. Y indicates either NR_2 or SR, while X is Cl or Br. Arrows show decreasing contributions A) The effect that varying the HOMO energy contributes to the predicted selectivity B) The impact that changing the substrate backbone on the size of the L YC_1 Sterimol values and its contribution to the predicted selectivity

For the B1 values, meta-substituted stilbene groups have the largest B1 value, while alkyl rings have the lowest (Figure 32A). When the meta positions are substituted, the two aromatic rings of the stilbene group adopt a *syn* conformation, which increases the B1 value. In contrast, when unsubstituted or *para* substituents, the two rings move away and adopt an *anti* conformation, reducing the B1 value (Figure 32B). For the alkyl ring, the minimum substitution is the C-H bond adjacent to the nucleophilic atom, resulting in the smallest possible B1 value, much smaller than any of the stilbene groups. Finally, the model favours longer L values, such as those of *para*-substituted benzene and benzyl-derived protecting groups (Figure 32C). Smaller

groups such as thiophene and methyl are less selective. However, the difference between para-substituted benzene and methyl groups is minimal, only 0.1 kJ/mol.

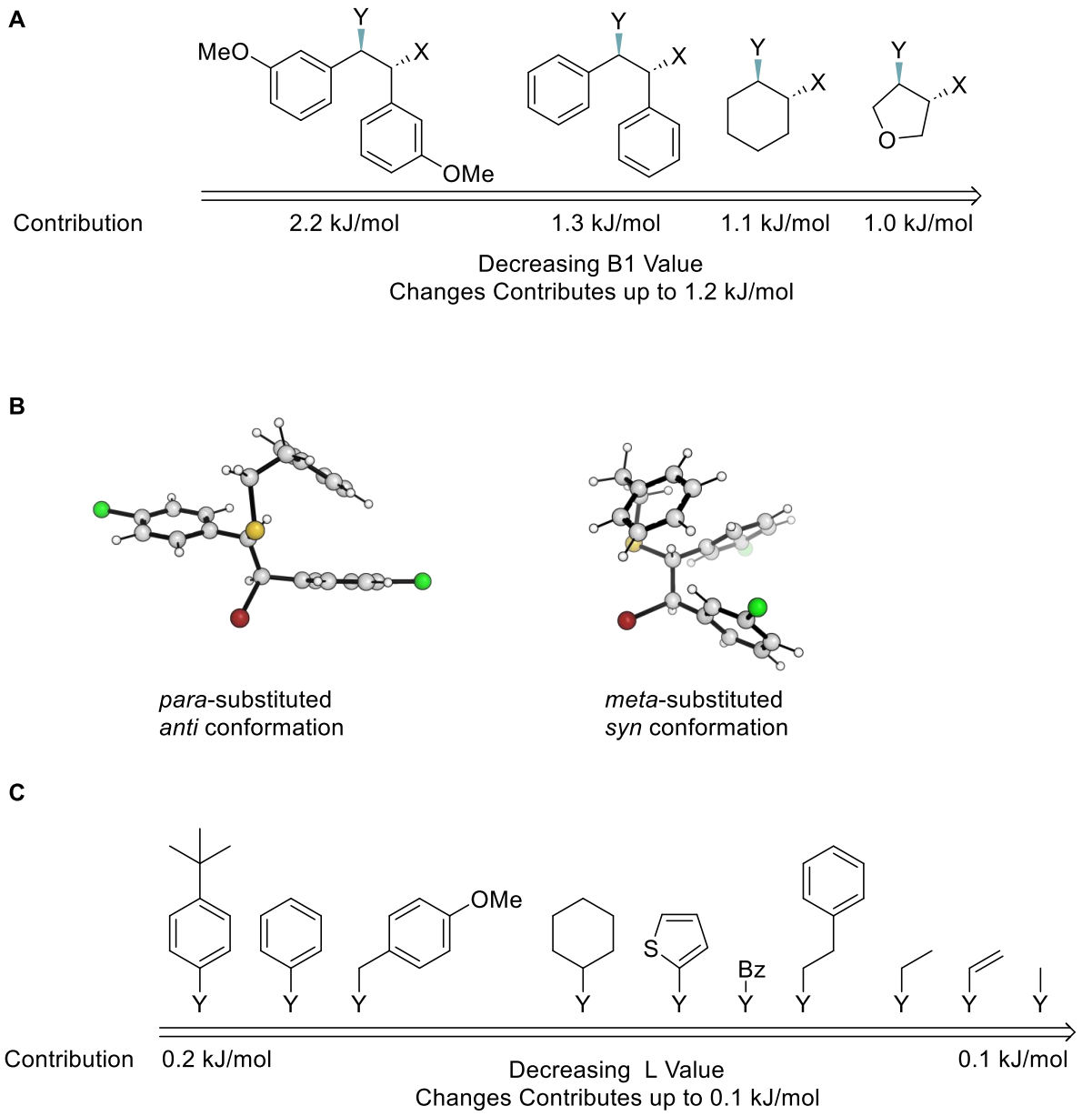
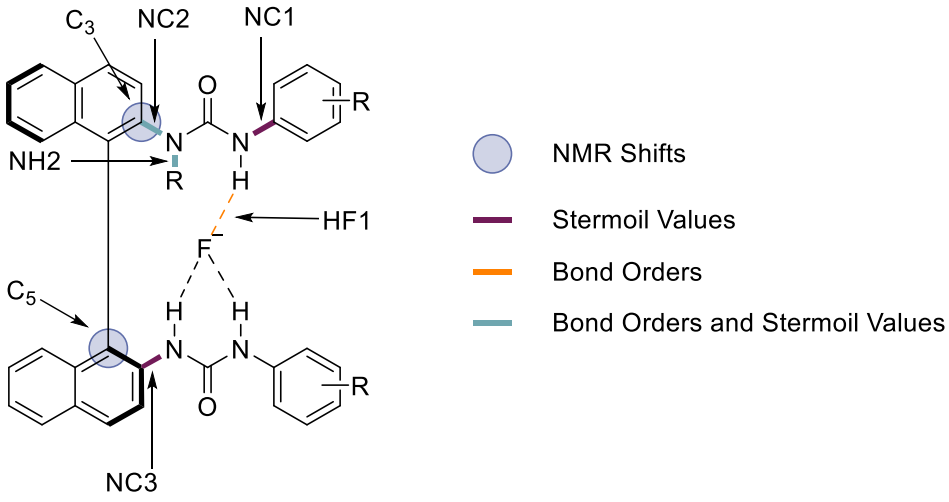


Figure 32 Contributions from the substrates as groups are varied. Y indicates either NR_2 or SR, while X is Cl or Br. Arrows show decreasing contributions A) How the variation of the backbone affects the B1 value for the YC_1 value B) Conformations for *meta* and *para* substituted substrates C) As the R groups are varied on the substrate the YC_2 L changes, however this contribution is small.

Catalyst descriptors

For the catalysts, 11 key descriptors are identified: five electronic and six steric. The electronic descriptors, BO_{HF1} measures the strength of the hydrogen bond between the H1 and the fluoride anion. This effect can be broken down into two types; changes that affect the coordination number, and changes that impact the electronic structure of the urea's. Increasing coordination number by adding more hydrogens bonds, decreases the HF1 bond order, leading to a decrease in selectivity of up to 0.4 kJ/mol (Figure 34A). As the number of hydrogen bonds to fluoride increases, the fluoride anion is stabilised over more hydrogen bonds, resulting in weaker H−F hydrogen bonds.

Changing the electronics is a more subtle effect; electron-rich R groups weaken the H−F1 bond order. This occurs because the electron-rich urea has stronger N−H bonds thereby reducing the charge on each hydrogen, making the protons less available for hydrogen bonding, therefore reducing the H−F bond strengths.



$$\begin{aligned}\Delta\Delta G^{\ddagger}_{Cat} = {} & 3.22BO_{HF1} + 0.63Cat_{NH2(L)} + 0.36Cat_{NH2(B1)} \\ & +0.32Cat_{NC1(B1)} + 0.12Cat_{NC3(B5)} + 0.008NMR_{C5} \\ & -0.004NMR_{C3} - 0.05Cat_{NC3(B1)} - 0.06Cat_{NC2(B1)} \\ & -0.97BO_{NH2} - 4.96BO_{NC4}\end{aligned} \quad (3.3)$$

Figure 33 Descriptors from the catalyst complex obtained from the MLR model and their respective coefficients.

The second electronic descriptor is the N−H bond at the site of alkylation (BO_{NH2}). Note that the N−H bond order is zero in this position when alkylated. Since the coefficient is negative, any non-alkylated catalyst is penalised, which aligns with experimental findings that alkylation is needed for high enantioselectivities. The N−H bond order is also affected by the nature of the aromatic backbone of the catalysts. A BINAM backbone has the strongest N−H bond, while the reduced BINAM backbone decreases the strength of the N−H bond. Finally, catalysts with a single aromatic group, such as Schreiner's urea, have an even weaker N−H bond (Figure 34B).

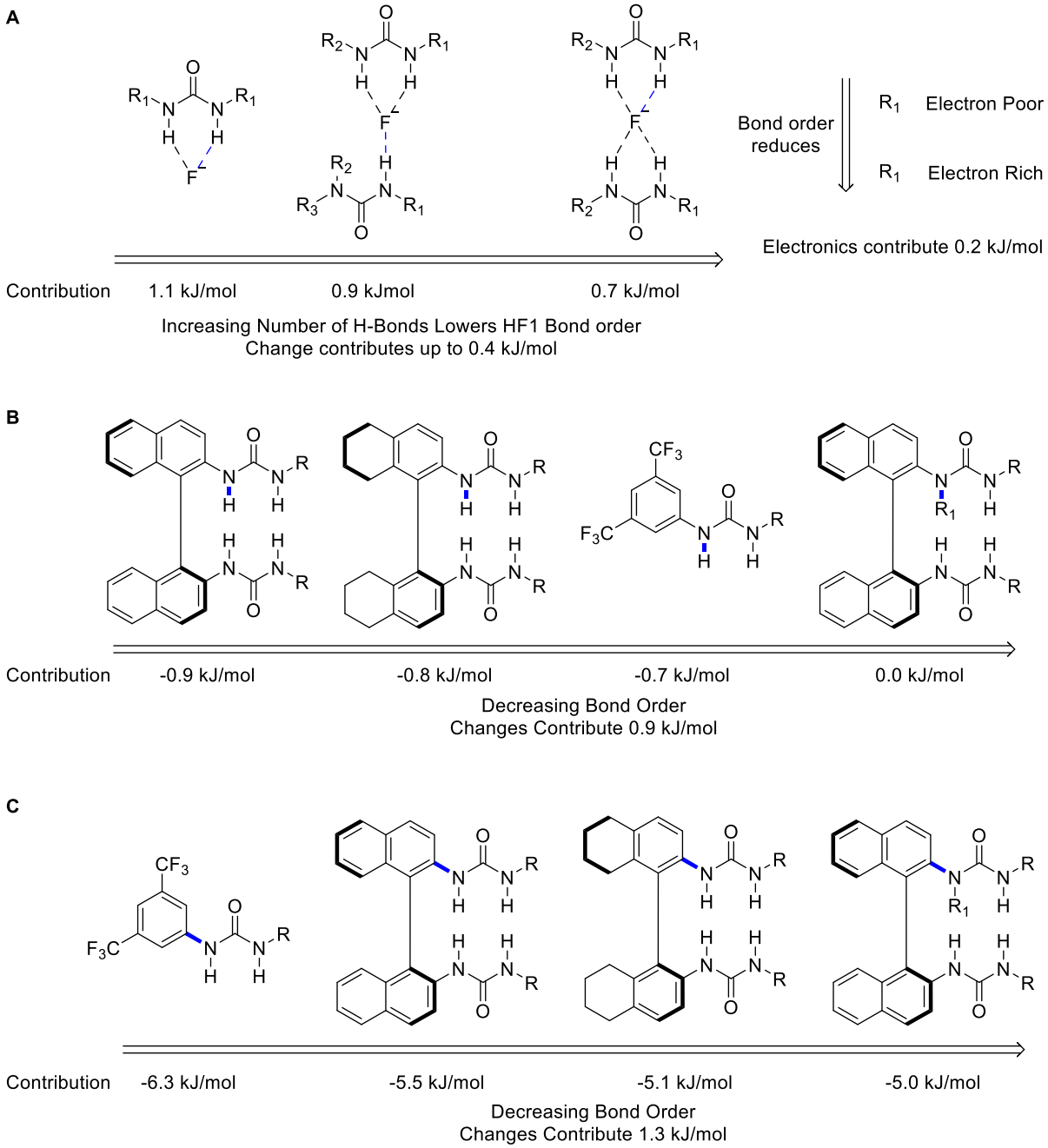


Figure 34 Breakdown of the contributions from catalyst-anion complexes. Bonds in Blue show the origin of the descriptor, arrows show decreasing absolute contributions. R = 3,5-bis(trifluoromethyl)benzene A) The impact of changing the number of hydrogen bonds on the predicted selectivity. B) How the variation in the Catalyst backbone affects the NH2 Bond order and its contribution to selectivity. R_1 = Me C) Impact of varying the Backbone on the Bond order for the NC2 Carbon. R_1 = Me

The third electronic descriptor is the bond order of the urea to the BINAM ring (BO_{NC4}). This descriptor provides information on the oxidation state of the backbone ring with the reduced

backbone having a lower bond order than a comparable aromatic backbone, this shows a lack of conjugation between the urea and aromatic system. This reduced bond order is further seen with the alkylation of the catalyst. This decrease in bond order suggests that conjugation across the BINAM backbone, which causes the N−C bond to have a higher bond order, correlates with a reduction in the selectivity of the catalyst as it can indicate the binding mode of the catalyst. Similarly, alkylation reduces the electron density on the nitrogen further reducing the bond order across the N−C bond (Figure 34C).

The final set of electronic catalyst descriptors are obtained from the ^{13}C NMR shifts the C3 and C5 carbons of the catalyst backbone (Figure 35). The C3 position is a good identifier of the type of backbone, aromatic vs alkyl (Figure 35A). There is little difference in the C3 NMR shift across BINAM catalysts and therefore this descriptor acts as a binary choice between the catalyst structures. Having a BINAM backbone leads to an increase in the $\Delta\Delta G^{\ddagger}$ of 0.4 kJ/mol. The C5 NMR shift is a good identifier of the oxidation state of the aromatic backbone, with high NMR shift for aromatic backbones, therefore this descriptor is selecting for BINAM based catalysts. Mono urea's have no C5 NMR at all, so are set as zero. Electronic changes on the backbone do not have significant effects on either the contributions from the C3 and C5 NMR shifts as the small changes of ppm result in less than 0.01 kJ/mol difference in the predicted selectivity.

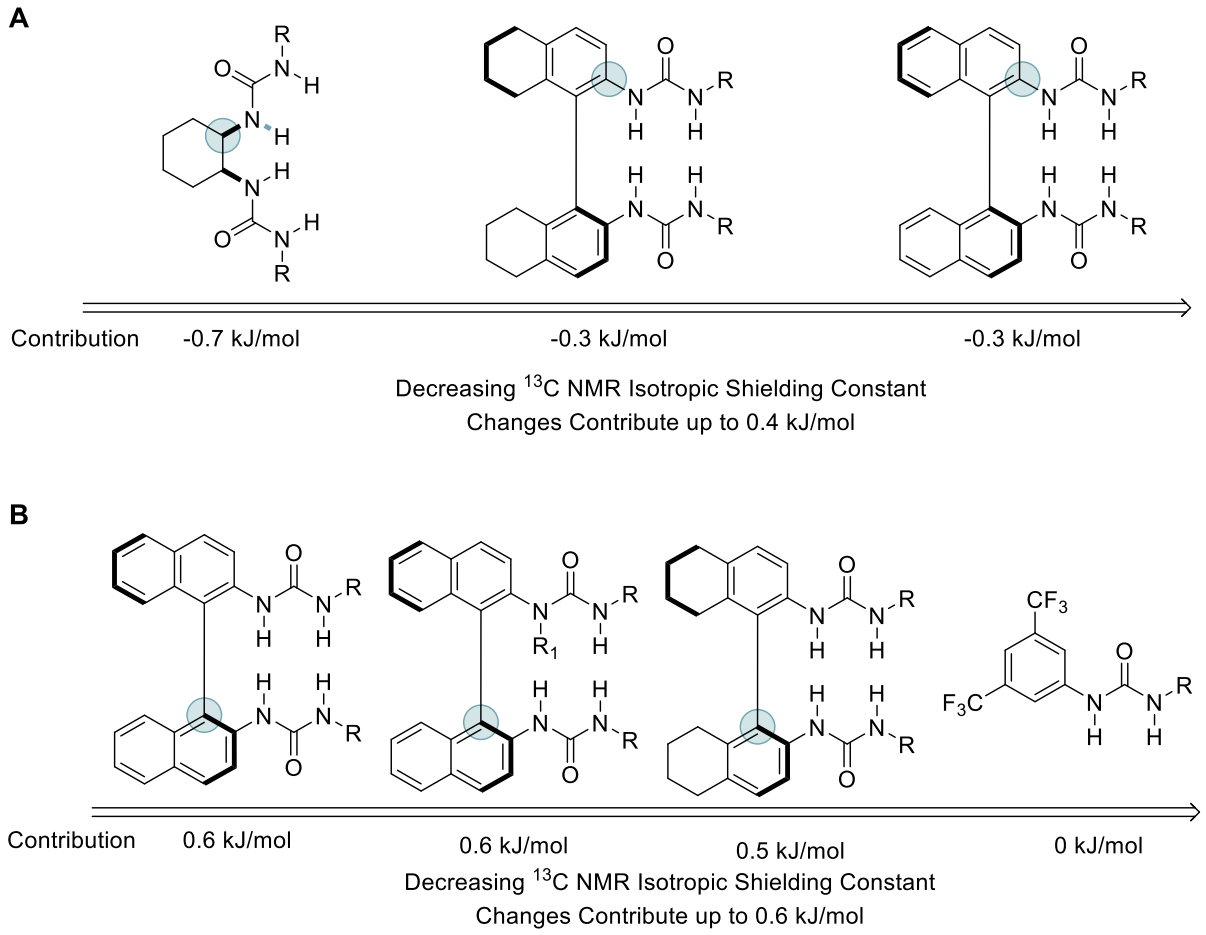


Figure 35 Final set of electronic descriptors selected by the MLR model. Atoms from which electronic descriptors are obtained are highlighted in blue, arrows show decreasing absolute contributions. R = 3,5-bis(trifluoromethyl)benzene R_1 = Me A) The C3 carbon NMR gives information on the catalyst backbone B) The C5 carbon NMR indicates information on the catalyst backbone oxidation state.

Important steric descriptors include $Cat_{NC1(B1)}$, which describes the sterics of the alkyl or aromatic groups on the right hand side of the catalysts (Figure 36A) and the steric descriptors associated with the urea backbone bond, $Cat_{NC3(B1)}$, $Cat_{NC3(B5)}$ and $Cat_{NC2(B1)}$.As expected, the largest groups, such as the substituted *m*-terphenyl groups, have the largest steric bulk.

One unexpected observation when analysing the $Cat_{NC1(B1)}$ values is the difference between bis-3,5 substituted benzene and unsubstituted m-terphenyl, which have similar contributions of 0.8 kJ/mol, even though the latter is significantly larger. However, on inspection, in the case of the *m*-terphenyl group, the atom from which the B1 is derived is a Carbon atom on the

terminal ring, rather than a hydrogen. Therefore, due to the overall geometry of the *m*-terphenyl group, the B1 value is similar to that of the 3,5-bis(trifluoromethyl)benzene groups, 2.37 Å and 2.38 Å respectively.

Analysis of the descriptors associated with the urea backbone bond include the B1 descriptors, which is directly affected by the substitution at the 3,3' position on the BINAM backbone. Given it is a negative coefficient, the smallest B1 value will have the least detrimental effect on selectivity. This corresponds to smaller groups such as H or Br, while the large naphthyl or 2,5 substituted phenyl rings significantly increase the B1 value and a negative effect on selectivity (Figure 36B). The B5 descriptor is more difficult to interpret as it represents the total steric bulk of the catalyst, which can be influenced by any substitutions either at the 3,3' position or on the area. However, in either case, large groups, substituted benzenes, or naphthyls are preferred on the catalyst structure, but in general, the more sterically bulky catalysts are more selective.

The final set of descriptors all describe the sterics interactions arising from the alkylation of the urea. Two descriptors were selected by the model with a positive coefficient, showing that increasing the size of this group will always increase the predicted selectivity. The L descriptor is the most significant giving up to 1.3 kJ/mol in selectivity over the unfunctionalized catalysts for the n-propyl group, while smaller groups provide reduced benefits. The B1 descriptor shows a smaller range of contributions of 0.3 kJ/mol between *i*Pr and H (Figure 36C).

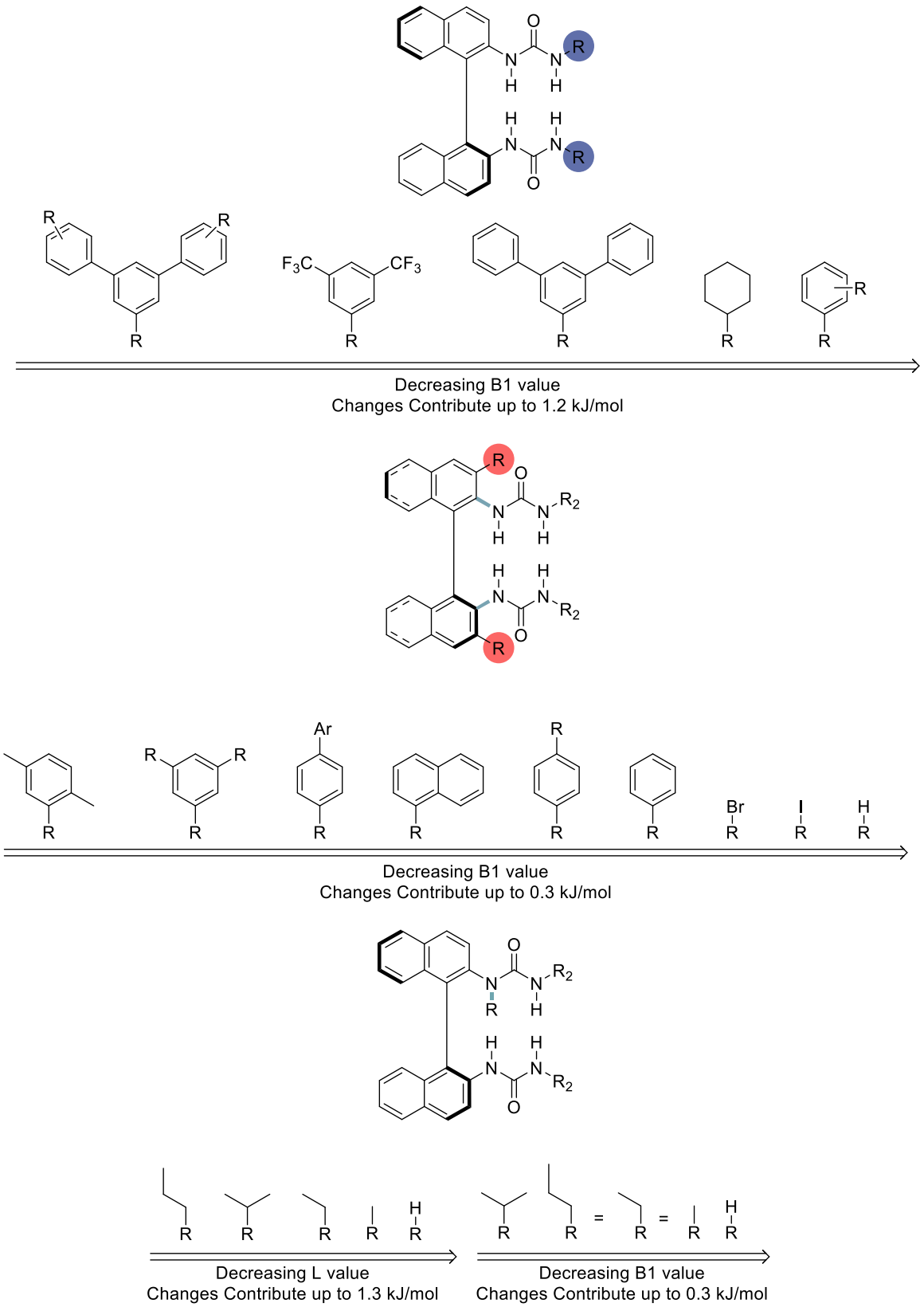


Figure 36 Steric catalyst descriptors selected by the MLR model. Bonds from which the descriptor is derived from are highlighted in blue. Arrows show decreasing absolute contributions R_2 = 3,5-bis(trifluoromethyl)benzene A) The impact that varying the R groups on the urea has on its B1 value, and therefore its selectivity. B) Variations at the 3-3' positions can give small changes to the selectivity C) Variations on alkylation has a significant impact on the overall predicted selectivity

3.4.4 Catalyst design: Database and generation

Computational substrate screening

Using our computational workflow set out in 3.2.2 Conformational sampling, 17 new substrates with a cyclohexane core were proposed for the HBPTC reaction (**Error! Reference source not found.**A), we also compared them again to the two common experimentally tested substrates (**N1** and **N2**). Computational descriptors were obtained, using Model **4** for all substrates and computationally screened against the 79 catalysts already synthesised by the Gouverneur Group to identify those combinations which led to the greatest selectivity.

Overall, the best performing catalysts from the training data are predicted to be the most selective on the new substrates. The worst substrate (**N4)** and best (**N15**) are consistently the best and worst, independent of which catalyst is used suggesting that catalyst design is significantly more challenging than substrate engineering, validating further catalyst design over substrate testing.

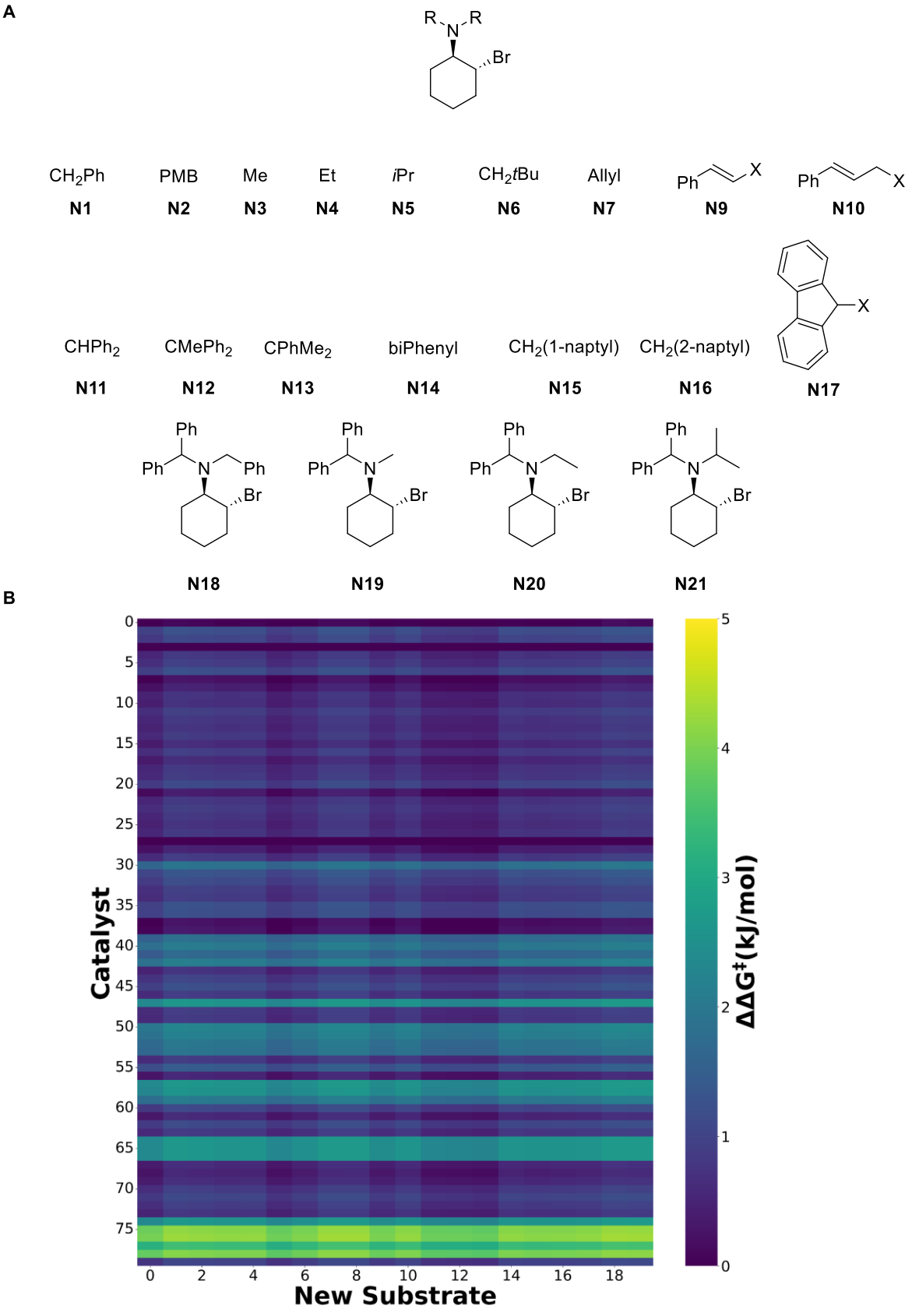


Figure 37 A) Proposed substrates for computational screening. B) Predicted selectivities for substrates against all training catalysts

From these results, the following trends emerge (**Error! Reference source not found.**). F irstly, purely aliphatic R groups perform less well (**N3**-**N5**, ~0.5 kJ/mol less selective) compared to aromatic R groups (**N14**-**N16**). This difference can be explained by analysing the LUMO of these substrates, a key descriptor in the model (See Substrate selectivity). For the aliphatic substrates, the LUMO is significantly higher (> 1 eV), than the best best-performing group(**N14**). The mixed protecting groups (**N19**-**N21**), as expected sit between the di aliphatic, and the di-$CHPh_2$ (**N11** -1.45 eV). In these cases, the presence of the $CHPh_2$ group has the effect of lowering the LUMO further compared to that of the demethylated amine (**N19** -1.24 eV vs **N3** -0.43 eV). Bulky protecting groups close to the nitrogen, such as $CMePh_2$ (**N12**), CMe_2Ph (**N13**), and CMe_2Ph (**N14**) are well tolerated and give some of the highest selectivity. However, bulky aliphatic groups, such as CH2tBu (**N7**), do not follow this pattern, showing that the steric bulk is less important than the electronic effects of the R groups.

Asymmetric substrates perform poorly (**N18**-**N21)** and are less selective than those protected with two CMe_2Ph groups (**N11**). This lower selectivity, as mentioned earlier, is due, in the model, to the raising of the LUMO caused by the alkyl groups, which reduces the predicted selectivity for this catalyst. Therefore, this mismatch of protecting groups results in less effective substrate. One important addition if that by using two different protecting groups, the aziridinium intermediate is no longer meso as two possible diastereomers are formed. This could lead to matched and mismatched TS's which are not considered in the model and could cause further issues. Of the tested substrates, the benzhydryl (**N14**) and naphthyl protecting groups (**N15**, **N16**) showed the highest selectivity.

Identifying key catalyst scaffolds

Our computational catalyst screening was informed by the previous results and the experience from our experimental collaborators. Firstly, except for the alkylation of the urea, only symmetrical catalyst functionalisation was considered, which is synthetically more

accessible. This resulted in only 10 possible positions for further functionalisation (Figure 38A). Considering this would result in a combinatorial explosion (10! = 3,628,800), we further considered functionalisation at positions 3,5 and 4, on the aromatic urea section of the catalyst as the others were synthetically challenging, along with , the alkylation of the urea, and modifications at the 3,3’ and 6,6’ positions on the BINAM backbone. This gives 5! permutations which results in a more manageable number of 120 scaffolds (Figure 38B). As a final constraint, each functionalisation should involve no more than 3 synthetic steps, which further reduces the number of core scaffolds to 4 (Figure 38C). These scaffolds are referred to as Core 1, Core 2, Core 3 and Core 4.

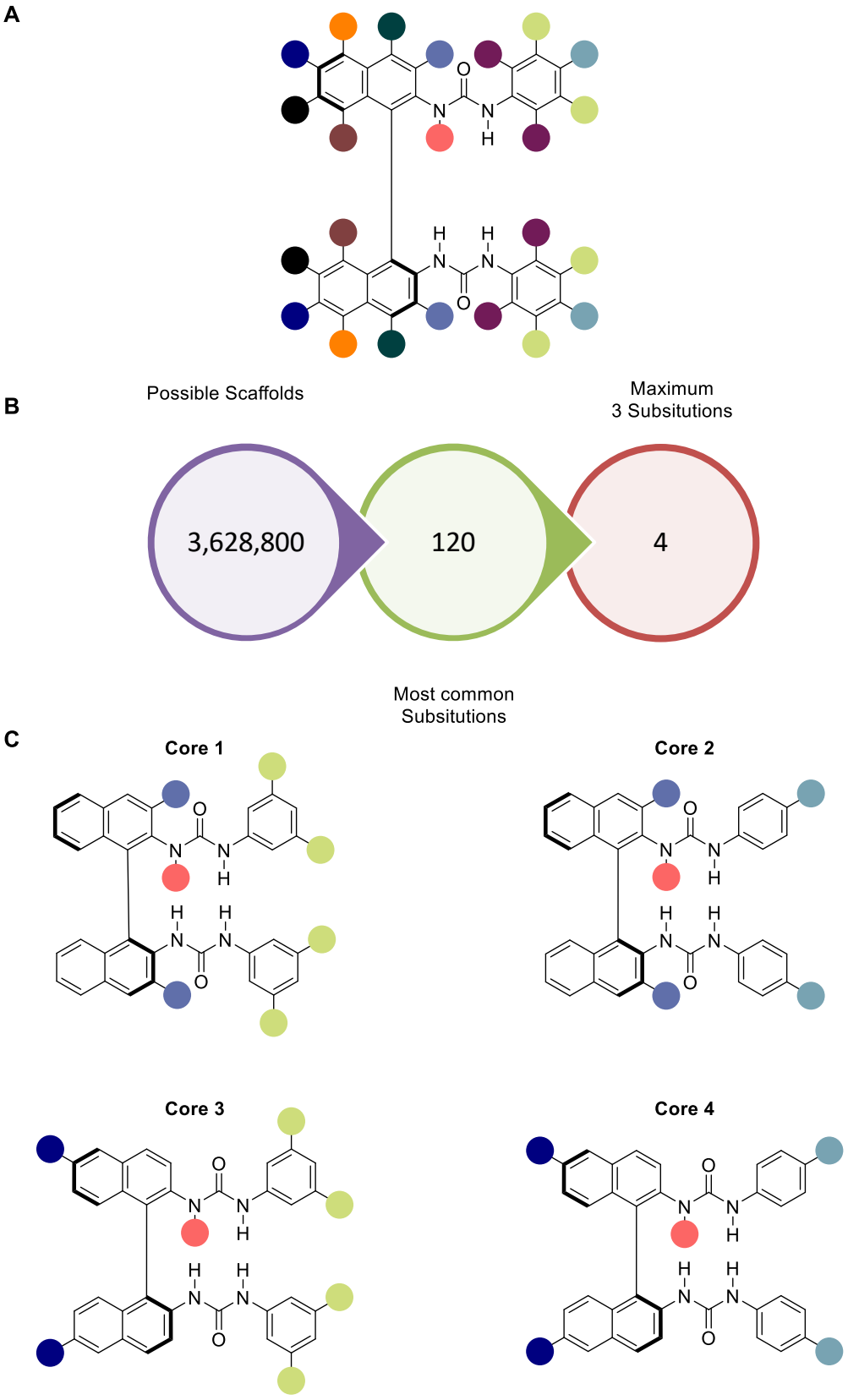


Figure 38 A) Number of possible scaffolds for all modifications and then when constraints are applied. B) Shows the final 4 scaffolds with modifications points highlighted

In collaboration with members of the Gouverneur Group, 53 possible R groups were hand-selected based on aryl and alkyl groups either previously used within the group or considered interesting targets (See 3.5.4 Catalyst generation for a complete list of all functional groups). When placed on the four possible cores, this resulted in 562,224 possible catalysts. Using our computational workflow, we generated 2000 of these possible catalysts; before using DataWarrior's clustering tool to cluster all the generated catalysts. From each cluster the most representative catalyst was selected for further calculations. This finally gave 138 catalysts which were subjected to our DFT workflow and predicted against the standard cyclohexyl aziridinium substrate (**2**).

Screening of computationally generated catalysts

To understand the effect that difference substituted groups have on the predicted enantioselectivity, we separated our analysis into effects on each type of core before looking at how the R groups affect selectivity.

Core 1

Catalysts with this core are predicted by the MLR to perform well (max 6.2 kJ/mol, min 2.7 kJ/mol, median 4.6 kJ/mol, mean 4.4 kJ/mol). Key features include a conformationally flexible group at the 3-3' position which is preferred rather than rigid groups, as seen in the predicted best-performing catalyst of this type (**H89,** 6.3 kJ/mol). The top five catalysts also follow this pattern, and all have a predicted selectivity above 5 kJ/mol (**H89**, **H35**, **H106**, **H86**, **H59**), these catalysts have their steric bulk at least one carbon away from the 3-3' position. Modifications on the urea side are more flexible with both rigid and flexible groups being well tolerated. These predicted worst performing catalysts also have a rigid steric bulk directly attached to the 3-3' position, either cyclohexane or CF_3 groups (**H101** 2.9 kJ/mol, **H79,** 2.7 kJ/mol) further showing that steric bulk close to the reaction centre could have a significant penalty to the reaction enantioselectivity.

Given the proximity of the 3,3' position to the hydrogen-bonded fluoride it is possible that the steric bulk close to the transition state does allow for a larger differentiation between the two enantiomers. However, where the steric bulk at the 3-3' position is placed directly above the fluoride anion, this changes the coordination around the anion, which in general leads to a lower predicted selectivity.

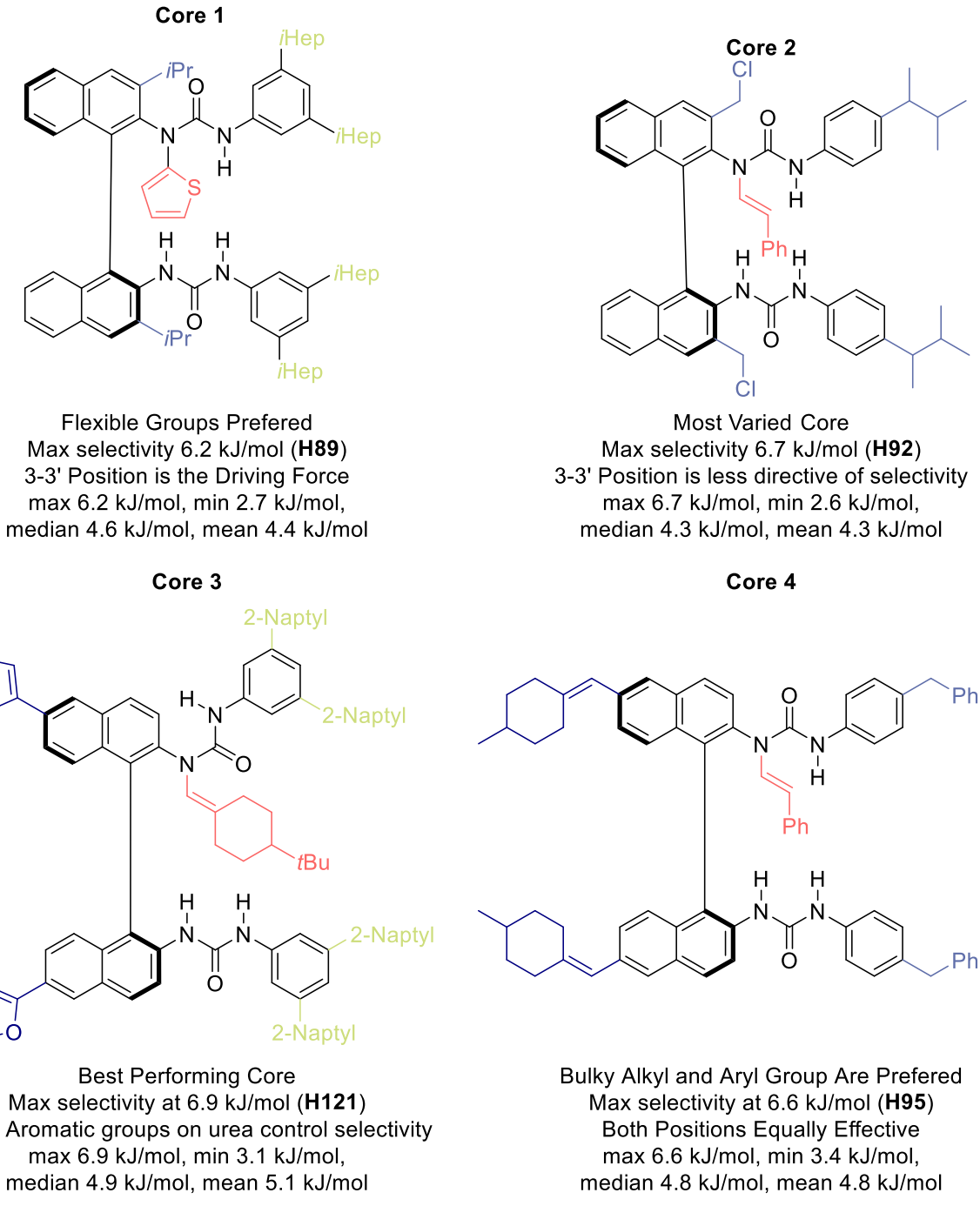


Figure 39 Summary for each core for the catalysts in the Handcrafted Dataset. Values in brackets indicate catalyst

Core 2

This set is predicted to be the worst out of all the 4 cores (max 6.7 kJ/mol, min 3.7 kJ/mol, median 4.7 kJ/mol, mean 4.8 kJ/mol). The steric demands of the 3-3' position, as seen in Core 2, are less important here with both small alkyl chains (**H92**, 6.7 kJ/mol), aromatic groups (**H22** and **H102**, 6.5 kJ/mol and 6.4 kJ/mol) and large CPh_3 groups (**H131**, 6.0 kJ/mol) all being well tolerated and in the top five predicted catalysts of this group.

We rationalise these predictions as the substitution on the urea is now at the para position and therefore less steric clash between the functionalisation on the backbone and those on the urea. This results in both positions being more flexible in the variety of functional groups tolerated. An interesting example is found with **H107** (4.1 kJ/mol), which has 3-propylseptane at the 4 position, benzyl groups at the 3,3' position, and a CF_3 group on the urea. The impact of electronic changes on the urea is something that has not been tested experimentally and is beyond the scope of this work and could be an interesting example for further study.

Core 3

This set of catalysts is the best predicted of all four cores (max 6.9 kJ/mol, min 3.1 kJ/mol, median 4.9 kJ/mol, mean 5.1 kJ/mol). In particular, large bulky groups such as extended aromatics and benzhydryls are all highly effective (**H121** 6.9 kJ/mol, **H112** 6.9 kJ/mol, and **H41** 6.8 kJ/mol), and rigid or bulky alkyl groups are well tolerated. Linear alkyl chains perform the worst when compared to other groups. Substitution at the 6-6' position seems to have a limited effect with both sterically bulky and small groups well tolerated. Given the proximity of this site to the reaction centre, the further experimental study focused on this position could provide an extra dimension for future catalyst design.

Core 4

This set of catalysts overall are predicted to be the second best and the most consistent of all 4 cores (max 6.6 kJ/mol, min 3.5 kJ/mol, median 4.8 kJ/mol, mean 4.8 kJ/mol). Overall, this core is tolerant of most functional groups, however aryl groups or bulky alkyl groups are those that give the most selectivity as demonstrated by the top five catalysts (**H95** 6.6 kJ/mol, **H116** 6.3 kJ/mol, **H20** 5.5 kJ/mol, **H119** 5.5 kJ/mol). When small groups are placed on the 6-6' position of the BIMAM ring the corresponding 4 positions of the urea must be a large bulky aromatic group such as benzhydryl (**H53**, 5.3 kJ/mol) or large aryl groups such as naphthalene (**H90** 5.0 kJ/mol) or CPh_3 (**H128**, 4.7 kJ/mol) otherwise the predicted selectivity will decrease. This core seems the most tolerant to large groups, since the substitutions at the 4 and the 6-6' position are far enough away in space to stop steric clash changing the geometries around the fluoride anion.

N Alkylation

As expected in the predicted best-performing catalysts, the alkylation of the catalyst has a significant impact on the predicted selectivity. Catalysts which have been alkylated by sterically bulky groups show excellent selectivity which matches our breakdown of the linear model (See 3.3.3 Catalyst descriptors). We also find an interesting subset of catalysts that have direct arylation on the nitrogen. This is not something that has been carried out experimentally to date, however, could provide a new method to modulate the electronics of the urea hydrogen bonds. We see both furans and thiophenes have a predicted level of selectivity that is higher than other catalysts (**H112** 6.8 kJ/mol, **H89** 6.2 kJ/mol, **H76** 6.0 kJ/mol). Another set of interesting functionalisation's are those that place the CF_3 group onto the urea (**H55** 4.8 kJ/mol, **H128** 4.7 kJ/mol, **H107** 4.1 kJ/mol). The electron-withdrawing nature of the CF_3 group has the potential to increase the hydrogen bond strength to the structures, which according to our model results in higher selectivity.

3.5 Conclusions

In this work, we present a computational workflow that can be utilised to generate ML models for hydrogen bonding catalysis. The protocol is automated and easy to follow, and it allows for the rapid identification of the important parameters (electronic and steric) that govern and regulate the stereochemical output of the reaction, helping chemists in the guided design of new effective catalysts. We aim that the workflows developed within this project can be applied to a range of other hydrogen-bonded catalysts allowing for ML models to be built on the fly alongside future experimentation.

We also introduce *graphSterimol*, a method which allows for the inclusion of whole molecules in the calculation of Sterimol values building on the work of Paton and Sigman. This allows for the bulk of a molecule to affect the conformations of the group which Sterimol values are calculated upon. This inclusion of the whole molecule does result in models with a higher R^2 and a lower RMSE when compared to work by Sigman. One area which would be a useful extension would be to include a conformational ensemble similar to the work of the Paton group. This would further allow for the inclusion of bulk conformations on the Sterimol values and could result in a more accurate description of steric bulk.

With a workflow for the generation of both descriptors and Sterimol values, we then applied this to the study of the HBPTC reaction developed by the Gouverneur Group. From the workflow, we found that while both fingerprints and DFT descriptors perform with similar levels of error, fingerprints perform worse on out-of-distribution test data and therefore DFT derived descriptors were utilised going forward. This model can predict enantioselectivities with errors of less than 1.05 kJ/mol for a range of catalysts and substrates. One limitation however, as with many MLR models is the prediction on catalysts, where there is an unexpected change in selectivity. We see that in these cases the model predicts higher

enantioselectivity, as would be expected from previous data, however, experimentally the enantioselectivity drops, therefore resulting in a larger error in our model.

With a trained model, we then interpret the model's coefficients to gain insight into the functionalities and motifs that are predicted to result in high enantioselectivities, allowing for quantification of the contributions from different parts of each catalyst or substrate. We found that catalyst contributions are significantly more important than substrate descriptors, and there are significant penalties for using aziridinium substrates over epi-sulfonium. This seems to be an interpretation of the data, where for the same catalysts, aziridinium substates perform worse than their epi-sulfonium counterparts. Therefore, further changes to catalyst design, over substrate engineering, are needed for higher enantioselectivities. The models' coefficients, however, do not directly match up with the previous DFT calculations on the important cation-π interactions and have to find a correlation between different descriptors to predict the selectivity.

With an understanding of both the requirements for optimal substrate and catalyst descriptors, in collaboration with members of the Gouverneur group we designed and computationally screened a range of catalysts and substrates to identify which could be possible future synthetic targets. Unfortunately, the results of our computational screening were finished in January 2020 and therefore the onset of COVID-19 meant that our experimental collaborators were unable to perform any experimentation on our suggested substrates. However, further experimental collaboration could result in the testing of some of these catalysts and substrates.

3.6 Methods

3.6.1 Data curation

To generate a dataset of experimental reactions and enantioselectivity, we collated data from both literature[16, 38] or internal unpublished data. Catalysts and substrates were stored as SMILES strings in the Cats.smi and Subs.smi files. See Appendix 2 for all substrates and catalysts. For each reaction, the enantioselectivity was converted to $\Delta\Delta G^{\ddagger}$ using the following equation:

$$\Delta\Delta G^{\ddagger} = -RTln(e.r) \tag{3.4}$$

where *e.r.* is the enantiomeric ratio, T is the temperature (K) at which the reaction was performed, and R is the gas constant (8.3145 J/K mol).

Finally, the reaction list was assembled which contained the index of each catalyst and substrate in every reaction along with the $\Delta\Delta G^{\ddagger}$, reaction temperature, and yield. To model the two experimentally tested solvents we utilised binary values: dichloromethane (CH_2Cl_2), is represented by 0, and 1,2-difluorobenzene (DFB) by 1.

3.6.2 Model generation

For all models ORCA 4.2[270] and xTB version 6.2[50] was used. The Aldrich's def2 basis sets was used throughout for DFT calulations.[105, 273] Dispersion corrections were taken into account using Grimme's D3 empirical method with Becke-Johnson damping (D3BJ) [102-104] For solvation effects, the SMD[123] implicit solvation method was employed with CH_2Cl_2 as the solvent.

3.6.3 Model generation

Model 1A - Molecular Fingerprints.

For the generation of the fingerprint reaction matrix, the Fingerprints.py script was utilised, along with the reaction, catalyst and substrate lists. The fingerprint radius was varied between two and six with the fingerprint length remaining at 2,048 bits. For one example, r = 3, the length was extended to 4,096 bits. The RDKit fingerprint was also tested using its default settings using a modified variation of the Fingerprints.py script.

Model 1B - Molecular descriptors

For the calculation of molecular descriptors the Mordred Python package (v1.2.0)[250] was used in the Mordred.py script. As with the generation of the molecular fingerprints, both the reaction, catalyst, and substrate list were given to the script with the overall reaction matrix being generated at the end of the script.

Model 2A – Fitting to core

For a core, a crystal structure of catalyst **78** was obtained from the Gouverneur Group,[16] and modified to only contain the main backbone of the catalyst. This along with the FittoCore.py script and the catalyst list was used to generate the 3D coordinates for each complex. For catalysts that failed to fit to the core, a second core which only contained one urea was used to generate the 3D coordinates of the complexes. For those catalysts that failed to generate coordinates using this second core, 3D coordinates were generated in the RDKit before the fluoride anion was added manually. Original testing to compare DFT functionals was performed by Stamatia Zavitsanou.

Each complex was then optimised at the PBE-D3BJ/def2-SVP level of theory before a final single point energy was calculated at the PBE-D3BJ/def2-TZVP level of theory along with NMR calculations. All optimisations and single points made use of the SMD[123] solvation model with CH_2Cl_2 used as the solvent. Descriptors were extracted using the Descriptors.py script.

Substrates were represented as Morgan fingerprints calculated in RDKit with a radius of 3 and a length of 2048 bits. The reaction matrix was generated using the substrate fingerprints, DFT descriptors, and the reaction list.

Model 2B – Conformational sampling path A

Complexes obtained from the FittoCore.py script were subjected to simulated annealing in xTB. Single point energies for each conformer were evaluated at the PBE-D3BJ/def2-SVP level of theory and a Boltzmann weighting was performed on conformers that summed to 90% were then optimised at the PBE-D3BJ/def2-SVP level of theory using the SMD[123] implicit solvation method. Complexes that failed to converge during optimisation were subjected to optimisation at the PBE-D3BJ/def2-SVP level of theory before being optimised with implicit solvent.

Finally, the LowestEnergy.py script identifies the lowest energy conformer for each catalyst and transfers that to a central folder that contains the lowest energy conformers for all catalysts. The electronic and structural descriptors were then calculated at the SMD(CH_2CL_2)-PBE(D3BJ)/def2-TZVP level of theory.

Model 3

For *Model 3* only electronic and structural descriptors are used. The model uses the lowest energy conformation for both the catalyst-anion complex and each substrate. The catalyst descriptors are the same as in Model **2B**, substrates are subjected to the same workflow just with the optimisation and descriptor calculations carried out at the PBE(D3BJ)/def2-SVP level of theory using Path **B**.

Model 4

Model 4 relies on electronic and structural descriptors with modifications to the level of theory and to the descriptors used, this utilised Path **B** in our computational workflow.

For each complex, the lowest energy conformer after the conformational sampling is extracted. In this model, the descriptors for the catalysts were calculated at the PBE(D3BJ)/def2-SVP level of theory with no geometry optimizations performed. Additionally, the average values of the NMR shifts (per atom type), the average values of charges (per atom type), and the summation of the values of bond orders (per bond type) are also included. Extraction and alignment are performed in the same manner as in Model **3**. The substrate descriptors were not changed.

Model 5

In this model, the catalyst descriptors are calculated using Path **A** as in Model **2B** and Model **3** however with optimisations performed using the SMD(CH_2Cl_2)-PBE(D3BJ)/ma-def2-SVP level of theory and the final single points at the SMD(CH_2Cl_2)-PBE(D3BJ)/ma-def-TZVP level of theory. The rest of the workflow and descriptor extraction were unchanged. The substrate descriptors were also not changed.

Table 5 Descriptors selected to represent the substrates under study.

Catalyst Descriptors	*Method*	*Units*	*Substrate Descriptors*	*Method*	*Units*
NMR shifts	GIAO	ppm	*Atomic Charges*	Hirshfeld	e
Atomic charges	Hirshfeld	e	*Dipole Moment*	-	Debye
Bond Order (BO)	Mayer	-	*HOMO/LUMO*	-	eV
Dipole Moment	-	Debye	*Sterimol parameters*	*Sterimol*	Å
HOMO/LUMO	-	eV			
Sterimol parameters	*Sterimol*	Å			

3.6.4 Catalyst generation

In collaboration with members of the Gouverneur Group, we selected a range of R groups with a variety of steric and electronic groups which would either be chemically plausible to add or lead to new groups not previously tested. This combined list contains 53 different groups in total.

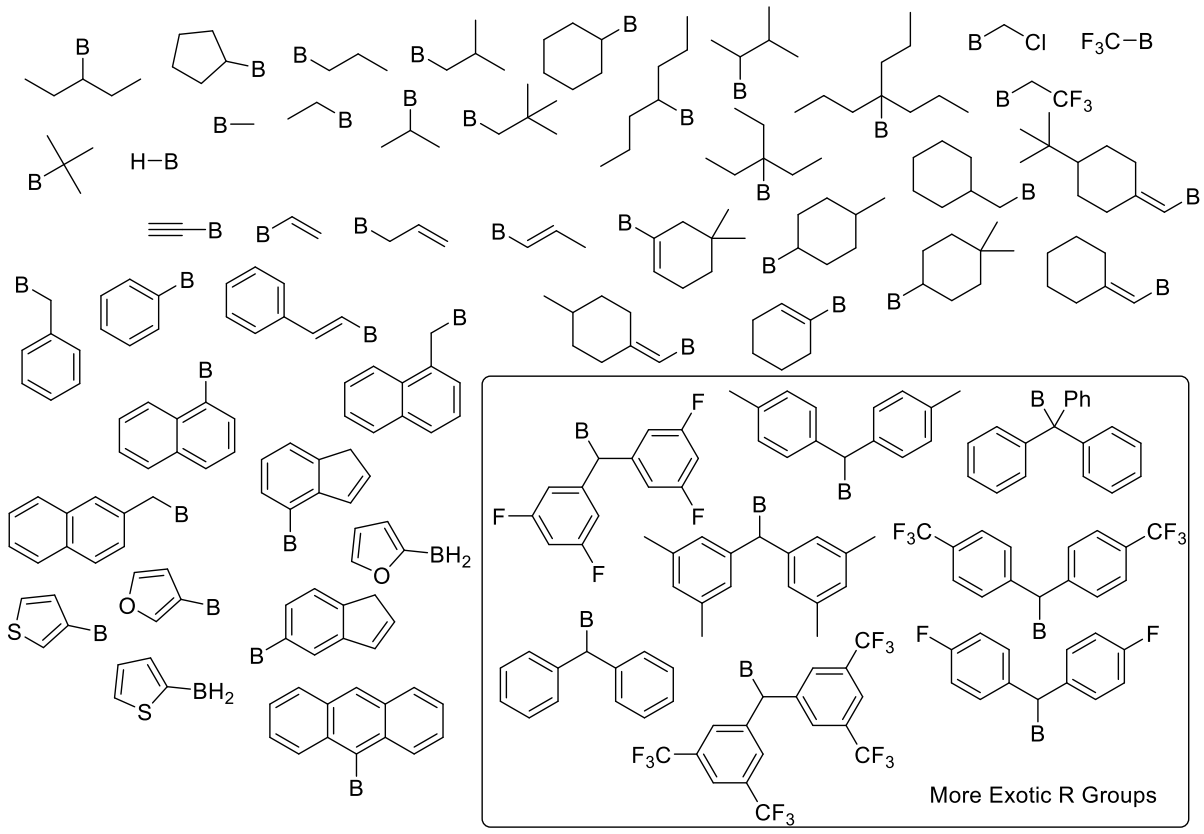


Figure 40 Handcrafted R Groups selected in collaboration with the Gouverneur group

Each catalyst was generated with the Database.py script using random generation to sample from the list of R groups this generated 2000 different catalysts. Catalysts that had a deviation of > 0.1 Å RMSD from the core were excluded. The remaining catalysts were then clustered in DataWarrior[274] and the representative example of each cluster was then taken forward for sampling using Path **B**. Catalysts which failed to converge in xTB (seven catalysts) or those which fragmented during optimisation were removed (nine catalysts).

3.6.5 Sterimol testing

Sterimol parameters and substrates from the work of Sigman *et al*,[236] were tabulated. To generate coordinates for *Sterimol* and *graphSterimol*, substrates, and catalysts for the desymmetrisation of bisphenols and NHK reaction respectively were generated. Conformational sampling was carried out using CREST (CREST v 2.1, and xTB 6.2).[50, 275] The lowest energy conformer was used to generate the Sterimol parameters either via Sterimol or graphSterimol. Linear models were generated in MATLAB 2020a in line with those reported by Sigman *et al*.[236]

Chapter 4 Calculation of ^{19}F NMR and $^{1}J_{HF}$ Coupling Constants in Hydrogen-Bonded Fluoride Complexes

4.1 Abstract

Experimentally determined NMR spectra are key pieces of data that are regularly utilised to identify and confirm the overall structure of organic molecules. Recently ^{19}F NMR shifts and $^{1}J_{HF}$ coupling constants were used to probe the conformational structure of Hydrogen Bonding Phase Transfer Catalysis (HBPTC) complexes and identify the positioning of the fluoride anion relative to the catalyst's hydrogen bonds. While there has been much progress over the last two decades in the DFT-based calculation of NMR properties, these methods have yet to be evaluated on these more unusual systems.

Herein we outline an approach for the calculation of both ^{19}F NMR shifts along with $^{1}J_{HF}$ coupling constants over the H-F hydrogen bond. When testing on small solvent fluoride clusters and Mono Urea Fluoride (MUF) complexes, we observe that explicit solvation of the fluoride anion results in significant differences for both properties with differences of up to 50 ppm and 20 Hz respectively. The developed approach is then applied to the prediction of $^{1}J_{HF}$ coupling constants for a range of HBPTC complexes, which achieves a Root Mean Square Error (RMSE) of 3.4 Hz over six different catalysts. Overall, this approach can replicate relative trends in a range of different hydrogen bonded fluoride complexes, however, absolute value calculation remains challenging.

4.2 Results and Discussion

In HBPTC catalysis, the strength and position of the hydrogen bonds around the fluoride anion can have a profound impact on the selectivity of the catalysts, this was observed both from the experimental work laid out in Chapter 1 and seen in our own MLR models described in Chapter 3. To further understand the positioning of the fluoride anion in the catalyst, we wished to develop an approach for the calculation of both ^{19}F NMR shifts and $^{1}J_{HF}$ coupling constants for hydrogen-bonded fluoride complexes. While the $^{1}J_{HF}$ coupling constants and ^{19}F NMR shifts have been measured experimentally,[39] we were curious if computational calculation of these measurements could lead to a better understanding of the location where the fluoride anion sits within the catalyst. This chapter is broken down into four sections: firstly, a DFT method is identified which can reproduce $^{1}J_{HF}$ and ^{19}F NMR on small clusters, before the influence of explicit solvation on the ^{19}F NMR is investigated. The computational workflow is then applied to the calculation of ^{19}F NMR and $^{1}J_{HF}$ on MUF complexes, before being applied to the HBPTC catalysts synthesised and characterised by the Gouverneur Group.

4.2.1 Benchmarking

To evaluate different DFT methods for the calculation of ^{19}F NMR shielding constants (σ) and $^{1}J_{HF}$ coupling constants, we compared a range of DFT functionals belonging to different rungs of the Jacob's ladder introduced in Chapter 1 along with a range of basis sets. The geometries, along with $^{1}J_{HF}$ and σ values, are then compared to the geometries and values obtained via *ab inito* calculations.

For the comparison of geometries, both the overall Root Mean Square Error (RMSE) of the H-F distance and the Root Mean Square Deviation (RMSD) of overall complexes were used to identify good structural agreement between different computational methods. For the prediction of $^{1}J_{HF}$ the RMSE between experimental and computed values are compared, while

for σ the RMSE between *ab inito* and DFT are compared along with correlation (R^2) to experimental data.

Datasets

The first dataset from Limbach *et al*.[67] contains coupling constants for $F^-(HF)_n$ (n=1-4) clusters in solution. These measurements were obtained using low-temperature NMR at temperatures between 110 K − 150 K with the complexes solvated in a 2:1 mixture of CDF_3 and CDF_2Cl. Alongside their experimental work, Limbach *et al.* also carried out DFT and CAS calculations to predict the $^1J_{HF}$ coupling constants in all four complexes. This experimental dataset has been previously used in theoretical studies by Bartlett *et al.*[71] and Pecul *et al.*[73] who used EOM-CCSD and MCSCF calculations respectively to predict the $^1J_{HF}$ coupling constants.

The second dataset reported by Christie *et al.*[60] contains NMR shielding constants for the tetramethylammonium fluoride (TMAF) complex in nine solvents (DCM, MeCN, H_2O, MeOH, CHF_3, EtOH, acetone, DMSO, and $CHCl_3$). Christie *et al.*[60] discovered that to predict the ^{19}F NMR you could approximate the full solvation of the fluoride anion by only using one solvent molecule in a gas phase and then extrapolating this to the full solvation through a linear fit.

Limbach dataset – H–F Clusters

To identify an appropriate basis set for CCSD(T) calculations we first compared a variety of basis sets to obtain a method that would balance computational cost and accuracy.

With the CCSD(T)/aug-cc-VDZ method as reported by Bartlett,[71] our H-F Bond length is 1.1519 Å almost identical to the one reported by Bartlett (1.1517 Å). However, with larger basis sets such as aug-cc-V5Z we obtain a slightly shorter H−F bond (1.1405 Å) at the basis set limit. The smaller def2-TZVP basis set reproduces the aug-cc-V5Z bond lengths whereas diffuse basis sets (ma) perform slightly worse when compared to the aug-cc-V5Z (Figure 41). Therefore, def2-TZVP was used to obtain geometries for all four H−F clusters.

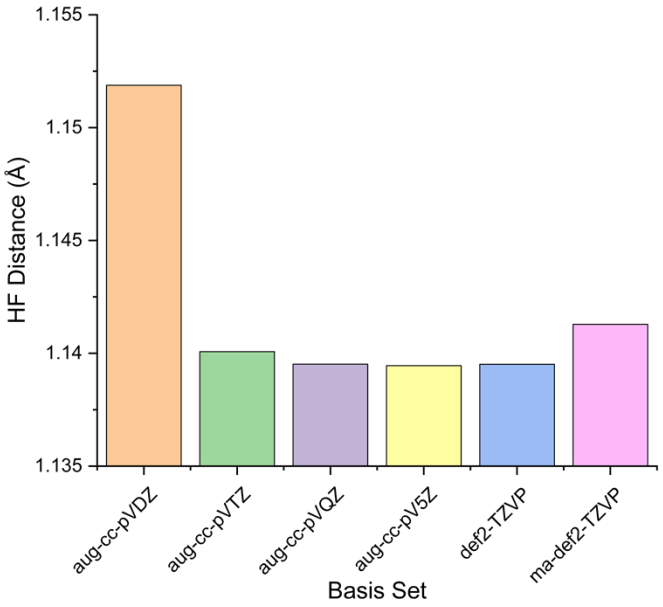


Figure 41 Effect of varying basis set size and type on the FHF complex. The average H−F distance is compared between the different basis sets.

Each F-$(HF)_n$ cluster structure was optimised at CCSD(T)/def2-TZVP and then used as reference structures for the evaluation of DFT methods. Among the DFT functionals tested, PBE0/D3BJ has the smallest RMSE errors (Figure 42, RMSE = 0.075 Å). Surprisingly, adding diffuse functions did not lead to improvement. The best compromise of accuracy and basis set size was PBE0(D3BJ)/def2-TZVPP. Selected data points were retested in ORCA 5.0 and showed negligible differences in the obtained structures.

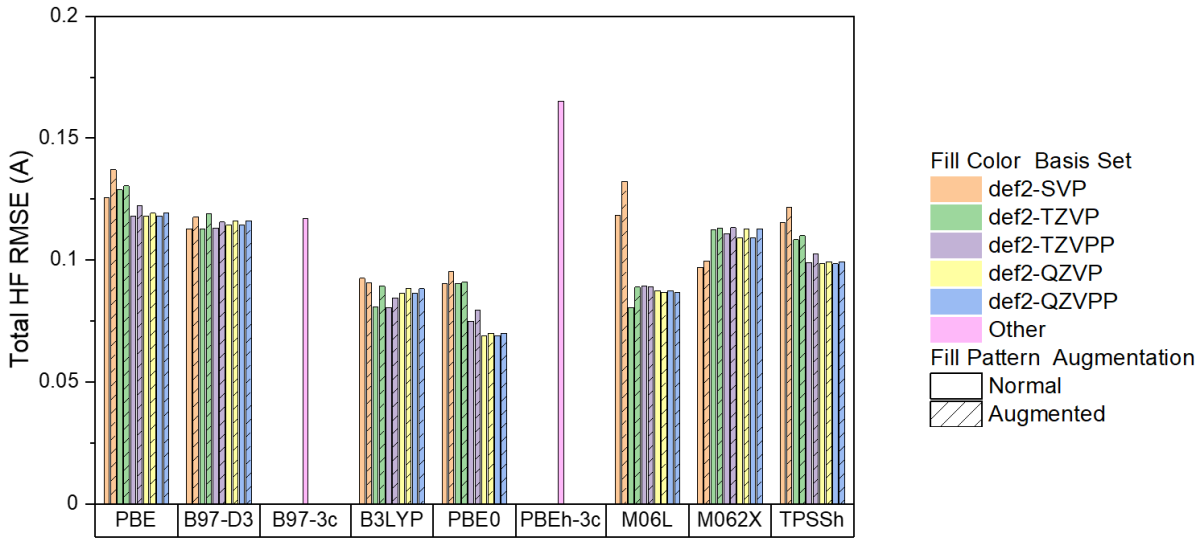
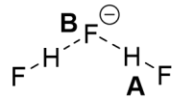


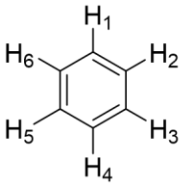
Figure 42 Overall RMSE for DFT optimisations over all $F(HF)_n$ clusters for n = 1-4 compared to CCSD(T)/def2-TZVP.

Employing the geometries obtained at the CCSD(T)/def2-TZVP level of theory, a range of DFT functionals including GGA, meta-GGA, and hybrid-GGA functionals were screened to compute $^1J_{HF}$ coupling constants in ORCA 4.2. The best performing of each class was retested in ORCA 5.0. In this case, significant changes were observed. Differences in up to 30 Hz for $^1J_{HF}$ coupling constants calculated using M062X were observed, while $^3J_{HH}$ coupling constants had much smaller differences of < 0.3 Hz (Figure 43). Given there was a significant change to

how the integration grids were implemented in ORCA 5.0 vs ORCA 4.2 we believe that this is the source of the observed changes.



	H-F A (Hz)	H-F B (Hz)	F-F (Hz)
Orca 4.2	224.550	-77.325	-80.734
Orca 5.0	188.408	-81.542	-142.540



	H1-H2 (Hz)	H2-H3 (Hz)	H3-H4 (Hz)	H4-H5 (Hz)	H5-H6 (Hz)	H6-H1 (Hz)
Orca 4.2	8.394	8.389	8.402	8.396	8.387	8.398
Orca 5.0	8.658	8.445	8.448	8.659	8.443	8.453

Figure 43 Comparing differences between coupling constants calculations using M062X/def2-TZVP in ORCA 4.2 and ORCA 5.0 for $F(HF)_2$ and benzene. Identical geometries are used for both calculations.

Each functional was analysed considering the NMR parameters associated to either the covalent H−F bond (type **A**) or the hydrogen bond between the H and the fluoride anion (type **B**, Figure 43). Overall, all the tested DFT functionals using Aldrichs def2-XVP basis sets underestimate the coupling constants for bond **A** when compared to the experimentally observed values. Extra polarisation or diffuse basis sets do not improve the calculations. The use of Jensen's pcJ-X basis sets, which were developed for the calculation of coupling constants, were found to slightly improve the results by 50 - 100 Hz. PBE0 and B3LYP show a similar performance with errors of 75 Hz when using the pcJ-3 or pcJ-4 basis set (Figure 44). M06-2X shows convergence problems with the Aldrichs basis sets; however, when using the pcJ-4 basis set the errors are 50 Hz or lower. Among the functional tested, B97-D3 perform the worst with errors of 100 to 200 Hz, even with the large optimised pcJ-4 basis sets.

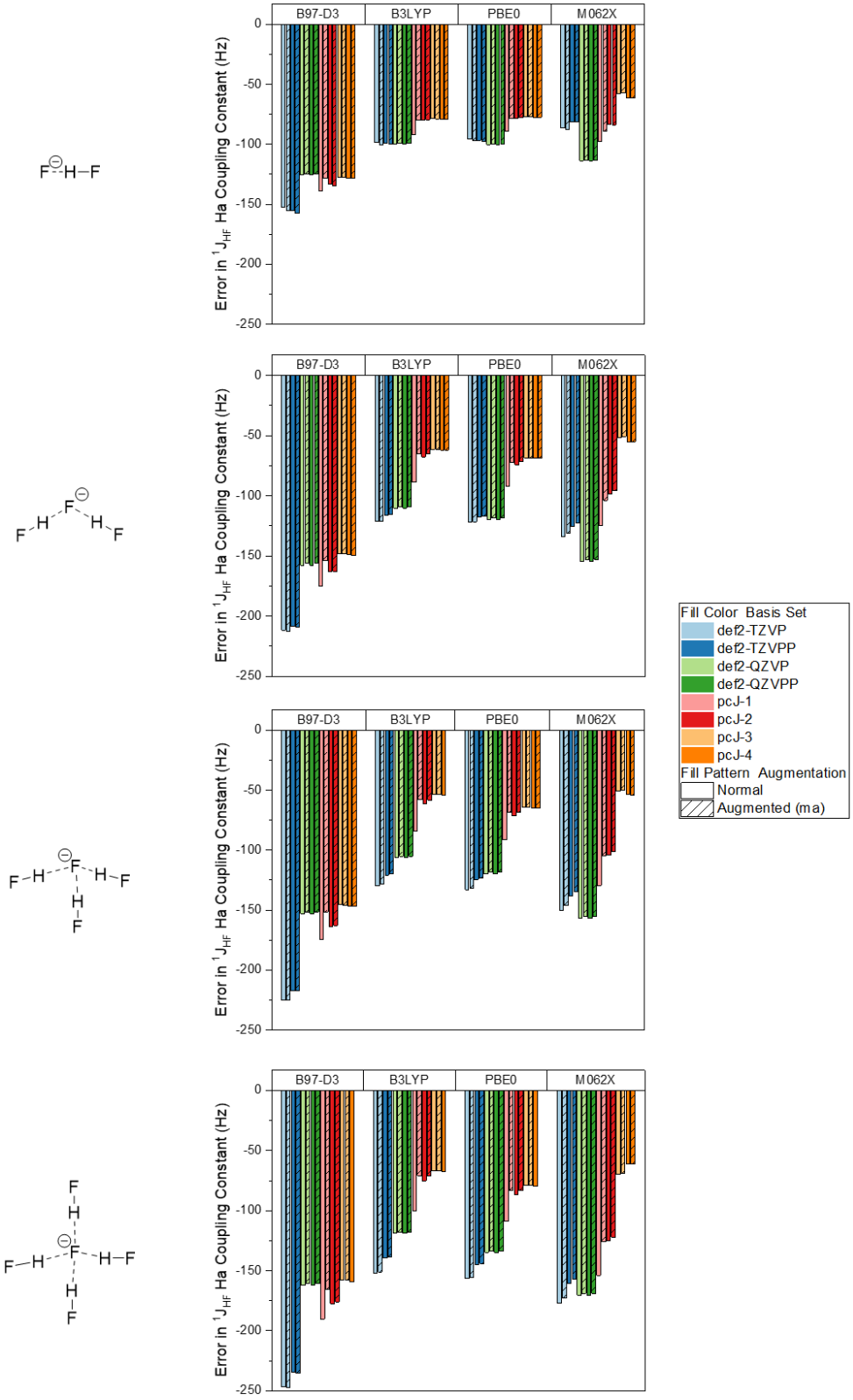


Figure 44 Error in $^1J_{HF}$ coupling constants for Ha in $F(HF)_n$ for n = 1-4 compared to experimentally determined coupling constants.

For the hydrogen bond **B**, comparing experimental and computed values is more challenging as reported in previous works.[71, 73] This is due to the fact that the assignment of the $F(HF)_4$ experimental values is tentative due to signal overlap with the $F(HF)_3$ cluster in the NMR spectrum.[67] The calculated coupling constants are consistent across different levels of theory. For example, both Aldrichs and pcJ-X basis sets behave similarly, with differences of 1-2 Hz (Figure 45). Similarly, B97-D3, PBE0, and B3LYP give similar errors when using def2-QZVP and pcJ-4 basis sets. Once again, M062X suffers from convergence issues, showing larger differences when using Aldrichs or Jensen basis sets.

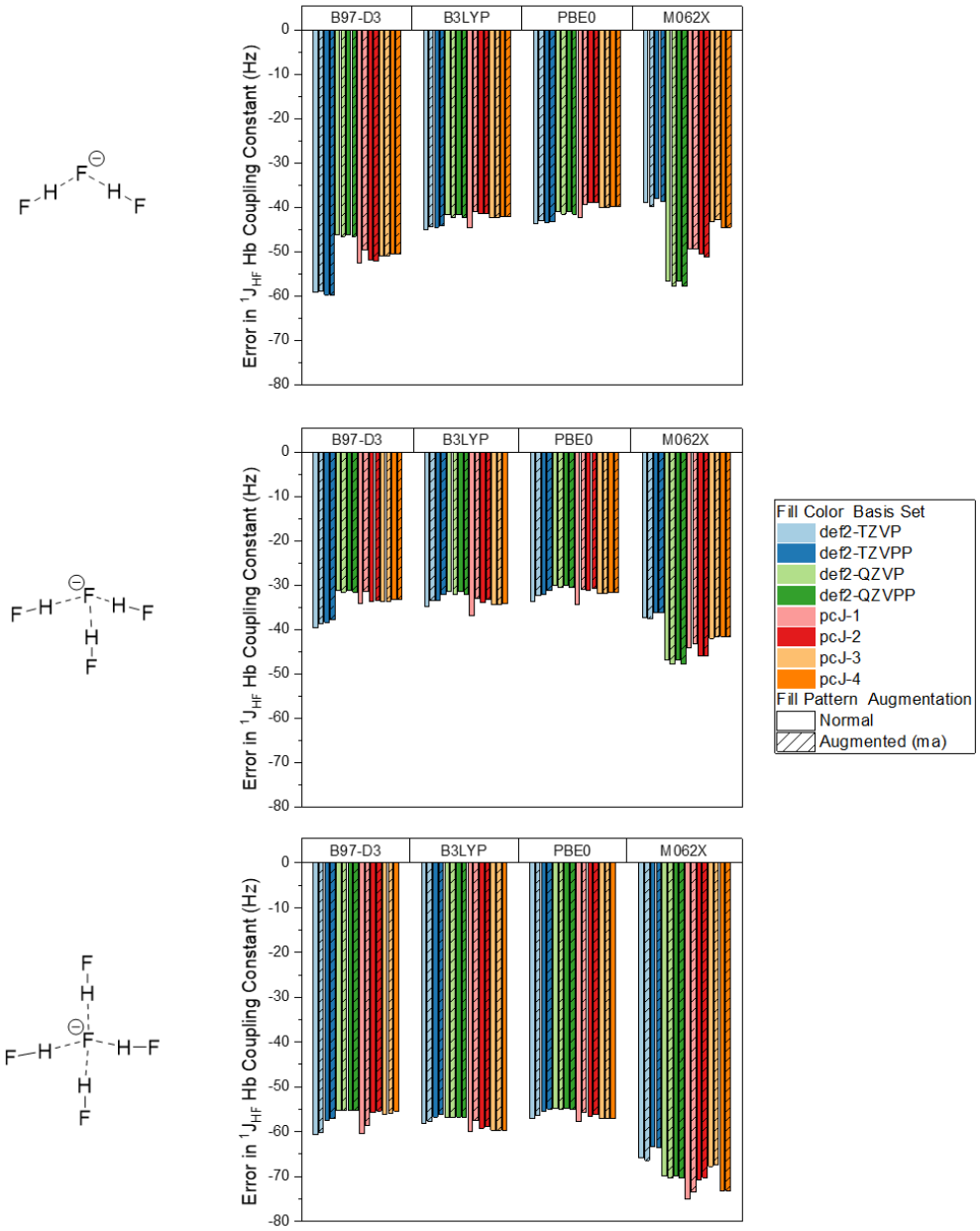


Figure 45 Error in $^1J_{HF}$ coupling constants for H-F hydrogen bonds in $F(HF)_n$ for n = 2-4 compared to experimentally determined coupling constants.

Christie dataset – Solvent-F Clusters

As with the Limbach dataset, to obtain the optimal basis set for benchmarking DFT functionals we began by testing a range of the basis sets on the F−H_2O cluster, using CCSD(T) for optimisations and MP2 for NMR calculations.

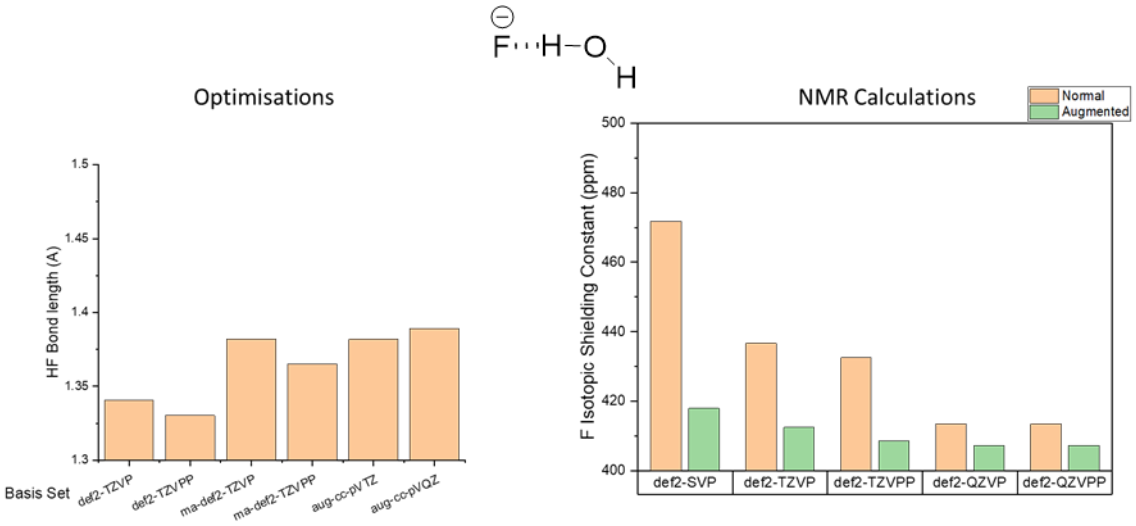


Figure 46 Basis set limit for CCSDS(T) Optimisation and MP2 NMR calculation on a F−H_2O cluster.

With geometry optimisations, the CCSD(T)/ma-def2-TZVP was able to reproduce geometries of the more computationally expensive CCSD(T)/aug-cc-pVQZ reference (Figure 46). For the calculation of ^{19}F isotropic shielding constant (σ_{iso}), as we increase the basis set size we converge once we reach the quadruple-zeta sets. We therefore selected RI-MP2/ma-def2-QZVP for the calculation of ^{19}F σ_{iso}.

When calculated over all of the nine clusters this method reproduces previously calculated results by Christe *et al.*[60] (Figure 47, Method **A**, $R^2 = 0.88$). Surprisingly, the use of implicit solvation during optimisation with either SMD or CPCM, leads to no correlation in the predicted NMR shifts (Figure 47, Method **B,** $R^2 = 0.00$). This unexpected behaviour is due to the changes in H−F hydrogen bond length. When including implicit solvent for the MeCN monomer the bond length increases by 44 % (0.608 Å) between the gas phase (1.382 Å) and

the implicit solvent geometry (1.990 Å). The difference for the water monomer is only 0.116 Å and an increase in the bond length of only 8 %.

To investigate if implicit solvation had a significant impact on the calculation of NMR shifts, the geometries of each complex were optimised in the gas phase, and implicit solvent was only included in the NMR calculations. This gives a better correlation to the experimental data (Figure 47, Method **C**, $R^2 = 0.86$). Therefore, the observed deviation from experimental results when using implicit solvent methods is due to changes in the geometry of the complex, rather than due to the NMR calculation.

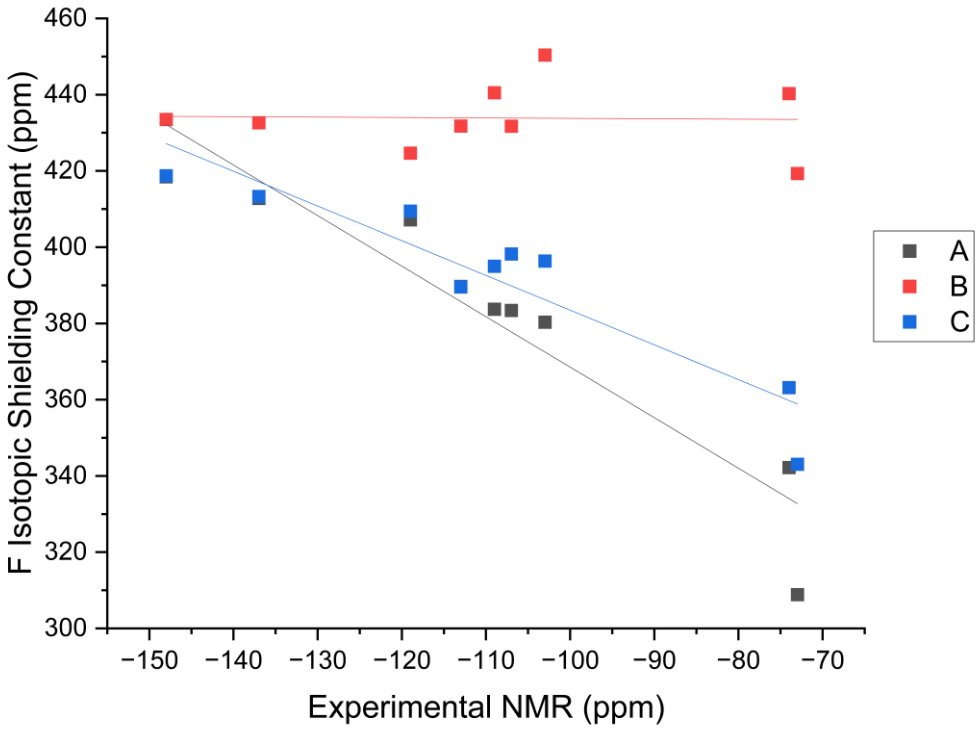


Figure 47 Prediction of ^{19}F NMR isotropic shielding constants again experimental NMR comparing the effects of implicit solvation on NMR prediction. Tested computational methods A: RI-MP2/ma-def2-QZVP//CCSD(T)/ma-def2-TZVP B: CPCM(Solvent)-RI-MP2/ma-def2-QZVP//CPCM(Solvent)-CCSD(T)/ma-def2-TZVP C: CPCM(Solvent)-RI-MP2/ma-def2-QZVP//CCSD(T)/ma-def2-TZVP.

As with our investigation into the H−F clusters, the initial benchmarking for geometries was carried out in ORCA 4.2, with the four best-performing functionals at different levels of theory for the H−F clusters tested in ORCA 5.0.

Overall, the use of augmented basis sets is important to reproduce the CCSD(T)/ma-def2-TZVP geometries. When comparing increasingly larger basis sets, convergence is observed with the (ma)-def2-TZVPP. B97-D3, B3LYP, and PBE0 with a ma-def2-TZVPP basis set all perform well with global RMSD < 0.01 Å and a H−F bond distance RMSD < 0.1 Å (Figure 48). M062X performs worse than all other functionals tested with an H-F bond distance RMSD >0.1 Å for the H−F bonds even when a diffuse basis set is used.

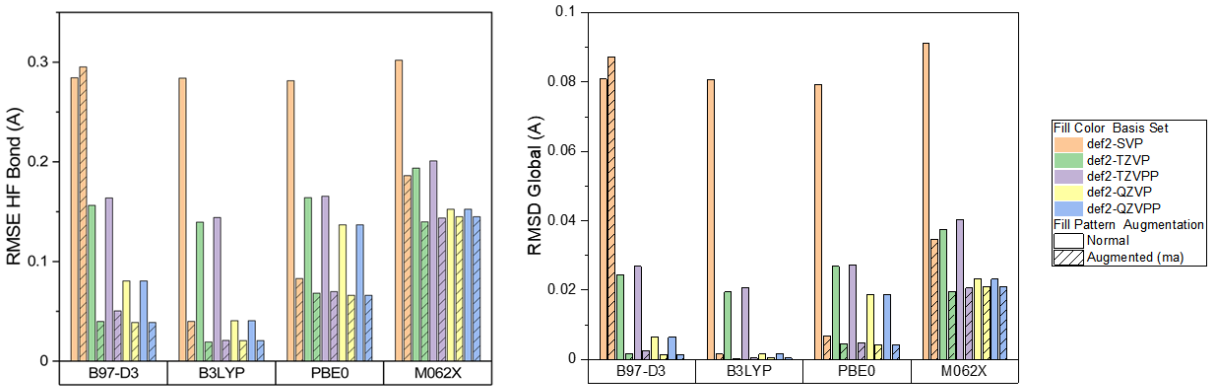


Figure 48 Comparison between CCSD(T) geometries and DFT geometries. HF RMSE is for the distance between solvent-H and fluoride while RMSD Global is the RMSD over all atoms in each cluster.

In 16/30 cases and 2/30 cases, both the DMSO and Ethanol clusters respectively had a large RMSD when compared to the reference structure. This is due to the different conformation that the complex adopts during optimisation; however, reoptimising from the CCSD(T)/ma-def2-TZVP reference structure removed these outliers.

For single-point NMR calculations, the same four functionals as tested for $^{1}J_{HF}$ coupling constants were tested on the CCSD(T) geometries. Surprisingly the best performance is obtained for the ma-def-SVP small basis set, independent of functional, compared to ma-def2-

QZVP and ma-def2-QZVPP, unfortunately, this is likely due to error cancelation rather than a better description of the electron density.

All the tested DFT methods correlate well with the experimental NMR signals ($R^2 > 0.6$), with M062X/def2-QZVP having the best correlation at $R^2 = 0.88$. Overall, M062X/def2-QZVP has the lowest errors compared to RI-MP2/ma-def2-QZVP calculations, RMSE = 20 ppm (Figure 49).

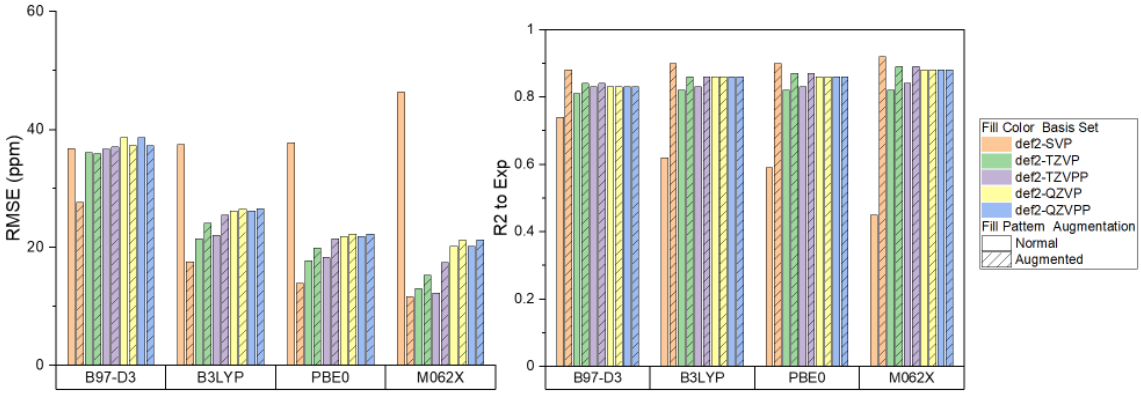


Figure 49 Error between DFT calculations and MP2 calculated shielding constants. R^2 of fit is between computed results and experimental results.

Benchmarking summary

Overall, PBE0/def2-TZVPP is the optimal level of theory for geometry optimisation for both datasets. The only difference between the datasets is the inclusion of diffuse functions; the first dataset of Limbach's H−F clusters does not require diffuse functions, but Christie's solvent-fluoride clusters do. However, due to an increase in computational cost with the use of diffuse basis sets, we only include them when needed.

For $^1J_{HF}$ calculations PBE0/def2-QZVP and PBE0/pcJ-4 are the best-performing methods, displaying errors of 30-80 Hz to experiment.[67] When compared to the best-performing computational method, EOM-CCSD[71], our DFT approach has errors of 10-30 Hz while being significantly less memory intensive and faster.

If we only consider the calculation of $^{1}J_{HF}$ across the hydrogen bond (**B)**, B97-D3/def2-QZVP also shows similar accuracies to that of PBE0/def2-QZVP with a significant speed up and a lower memory requirement. For the calculation of ^{19}F NMR shifts of the fluoride anion, M062X/def2-QZVP is the best-performing functional but shows poor performance for $^{1}J_{HF}$ coupling constants. Considering performance and efficiency, our method for geometry optimizations is PBE0/def2-TZVPP (with diffuse functions added for solvent clusters) and M062X/def2-QZVP for ^{19}F NMR shielding constants, while B97-D3/def2-QZVP is used for $^{1}J_{HF}$ coupling constants.

In conclusion, while our DFT workflow described above can accurately predict ^{19}F NMR shielding constants for anionic fluoride complexes, $^{1}J_{HF}$ are less well predicted, and none of them can match *ab initio* results of covalent H-F bond. Keeping in mind that our target is to obtain relative trends rather than absolute values, we believe that this method should capture these trends.

4.2.2 Influence of explicit solvation on ^{19}F NMR values

To evaluate the effect of explicit solvation upon ^{19}F NMR values we first studied the effect of varying the distance and angle between two solvent molecules solvating fluoride, before designing complexes with between one and six explicit solvent molecules for each of our nine solvents. We hypothesised that the sequential addition of an explicit solvent for both gas phase and implicit solvent calculations would result in the ^{19}F NMR shift converging.

Geometry scans

To understand the effect of solvent position on ^{19}F NMR σ_{iso} we performed a geometry scan on dimer complexes, where two solvent molecules are coordinated to the fluoride anion. We vary the H-F distance (r_1 and r_2, Figure 50) while keeping the angle between the solvent molecules constant (θ, Figure 50).

The resulting potential energy surface (PES) shows clear differences between the solvents with hydrogen bond donor groups (Water, MeOH and EtOH), those with halogen-containing atoms (DCM, $CHCl_3$ and CHF_3) and those that are polar aprotic (MeCN, Figure 50). The energy minimum is found with a H-F distance of 1.53 Å for water, 1.76 Å for DCM, and 1.86 Å for MeCN.

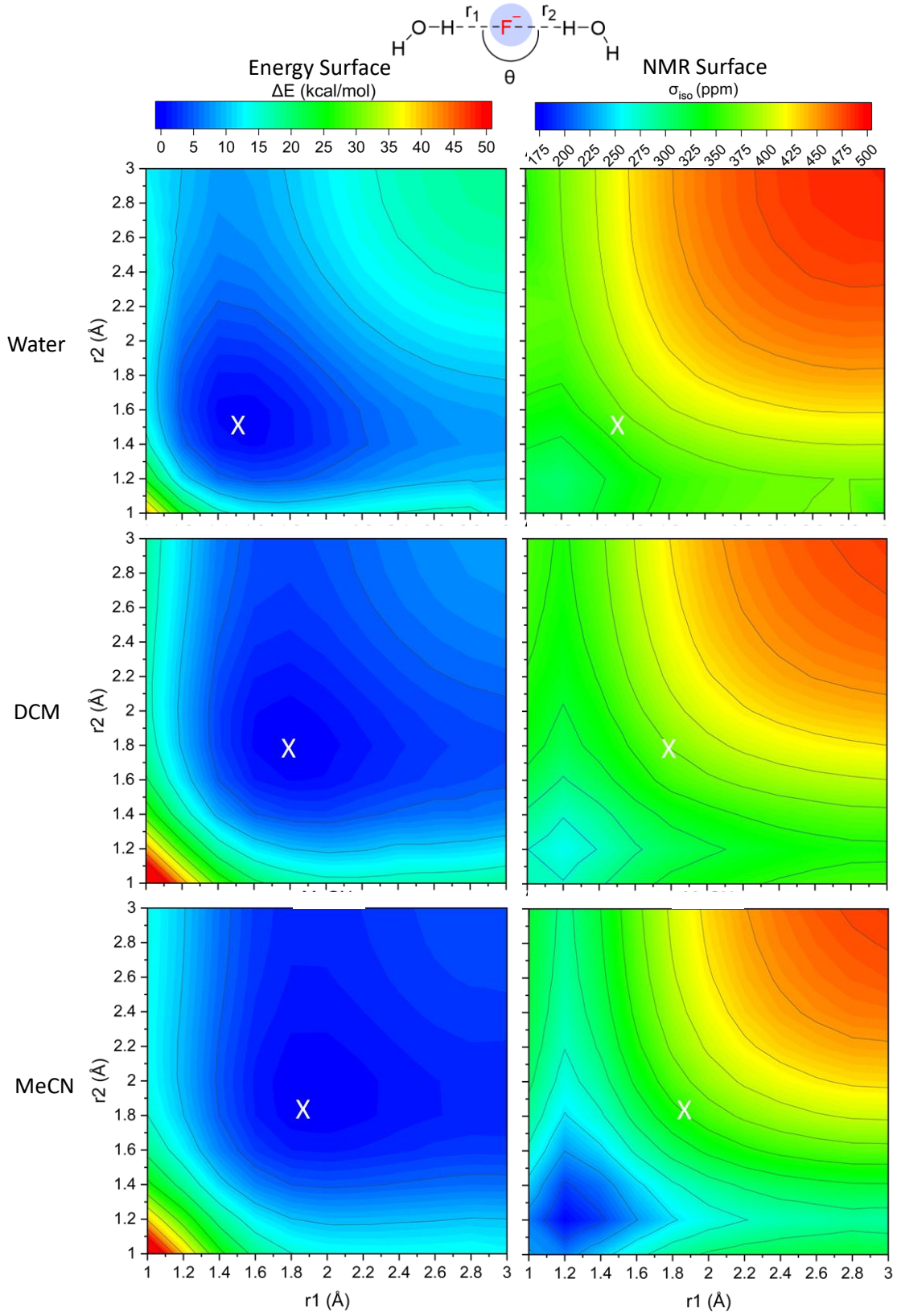


Figure 50 2D geometry scan for Water, DCM, and MeCN dimer across the r1 and r2 H-F bonds as shown on the example for the Water dimer. Calculations performed at the CPCM(Solvent)-M062X/def2-QZVP// CPCM(Solvent)-PBE0(D3BJ)/ma-def2-TZVPP level of theory. X signifies the PES minima. Contour lines are shown for every 5 kcal/mol and 25 ppm.

Analysis of the ^{19}F NMR σ_{iso} surface shows a similar pattern independent of solvent type, with the minimum in the σ_{iso} occurring at an H-F bond distance of 1.2 Å. However, there is a significant difference in the σ_{iso} at this point: MeCN is at 177 ppm while MeOH is 322 ppm. Inspection of the structures at this point explains this result; for MeCN the C-H bond has significantly elongated to 1.26 Å, from 1.09 Å, which causes a deshielding effect as electron density is moved away from the fluoride anion onto the approaching hydrogen. In the case of the MeOH the OH bond elongates less only from 0.95 Å to 1.09 Å. Therefore, less electron density is moved from the fluoride anion onto the proton.

As expected, as solvent molecules move further away from the fluoride anion the shielding on the anion increases until we approach a maximum of 480 ppm at 3 Å. At this point there is no removal of electron density from the anion anymore and therefore the NMR shift is purely the fluoride anion in the CPCM solvent, which as seen in our initial WFT calculations does not affect the ^{19}F NMR σ_{iso}.

When scanning the angle between the two solvent molecules and the fluoride anion, if the θ is between 90° and 170° there are minimal changes in the ^{19}F NMR σ_{iso} (< 10 ppm change for all solvents), and as the angle decreases further there is a significant decrease. This is an effect of the total geometry, as at small angles the solvent molecules begin to clash resulting in distorted solvent structures.

Sequential addition of solvent molecules to F-(Solvent)$_n$ complexes

For each of the nine solvents we placed solvent molecules around the fluoride anion in positions similar to those reported by Gerken *et al.*[60] While this would not give a full representation of all possible confirmations, it would allow us to evaluate the impact of changing the coordination on the ^{19}F σ_{iso}.

For each complex, both gas-phase and implicit solvent calculations were performed and plotted against the experimental ^{19}F NMR shift. As seen in our wave function work in the previous

section, gas-phase calculations with one explicit solvent molecule correlate well with experimental results ($R^2 = 0.90$), while the same calculations performed with implicit solvent do not correlate well ($R^2 = 0.13$) (Figure 51A, light vs dark blue lines). The more pronounced difference can be seen for polar aprotic solvents, such as MeCN and DMSO (Figure 51A, -70 to -80 ppm).

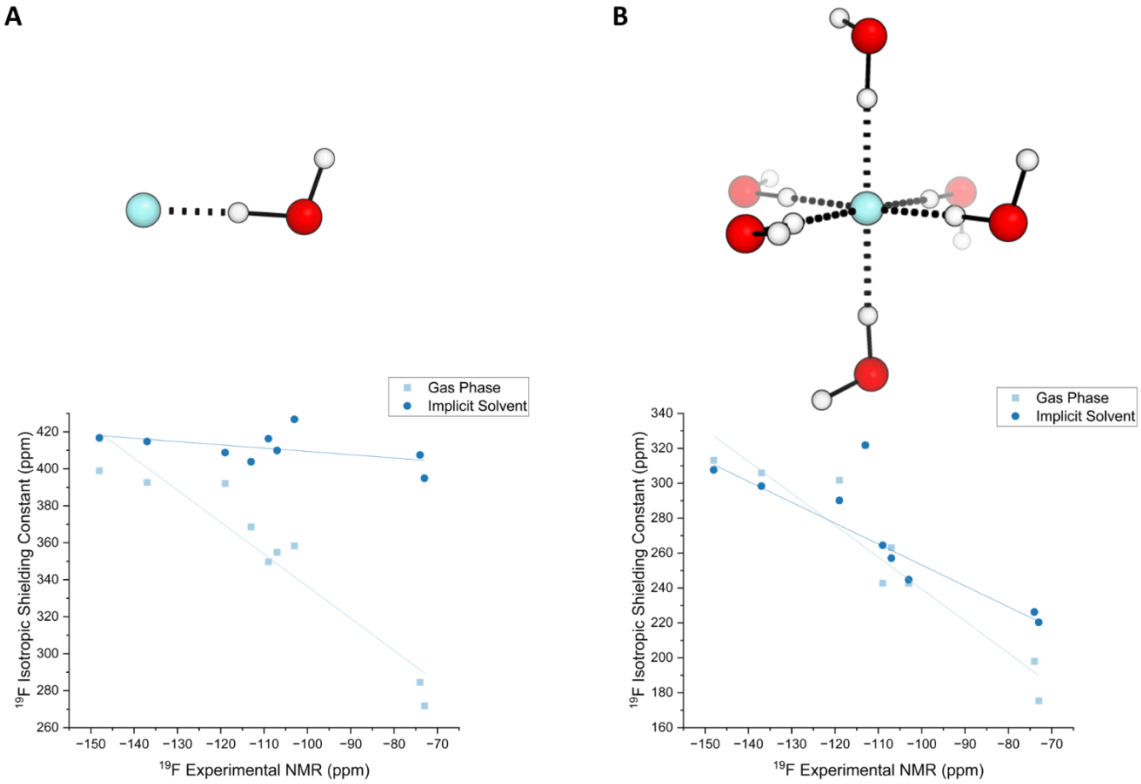


Figure 51 The effect of addition of extra explicit solvent molecules on ^{19}F σ_{iso} using the M062X/def2-QZVP//PBE0(D3BJ)/ma-def2-TZVPP methodology, each datapoint represents a different solvent. CPCM solvation is used for implicit solvation calculations. A) Mono complexes, with one solvent molecule and the fluoride anion. $R^2 = 0.90$ for gas phase and $R^2 = 0.13$ for implicit solvent calculations. B) Hexacoordinate fluoride complexes, $R^2 = 0.95$ for gas phase calculations and 0.91 for implicit solvent.

Our results show that with the six explicit solvent molecules, the differences between the gas phase and implicit solvation converge to within 10 ppm of each other, apart from DMSO and MeCN which remain more than 30 ppm apart (Figure 51B). At this point, both sets of

calculations correlate well with the experimentally determined ^{19}F NMR shift, $R^2 = 0.95$ for gas phase calculations while implicit solvation has $R^2 = 0.91$.

During the sequential addition, it became noticeable that $CHCl_3$ was consistently an outlier in the data, both with gas phase and implicit solvation methods. We initially suspected that this could be due to a lack of spin-orbit coupling in the DFT functionals, which results in NMR shifts on atoms adjacent to a Cl being highly inaccurate when compared to experimental results.[276-281] However, in our case, the ^{19}F anion is not directly adjacent to the Cl but 3 bonds away, and these changes would be very subtle, and are unlikely to be passed across the hydrogen bond. Furthermore, while $CHCl_3$ is an outlier, the closely related CH_2Cl_2 is not an outlier in any of our calculations. The more likely explanation therefore is that the experimental data either is misreported, or the solvent is partially contaminated, for example with water and therefore the coordination around the anion is not purely $CHCl_3$. This can be a significant problem where solvents such as $CHCl_3$ which are hydroscopic and so making a sample free to water totally is practically impossible, and therefore even a single molecule of water could bind to fluoride, could lead to significant differences in the observed shifts

Conformational sampling

As seen in Figure 51, even with six explicit solvent molecules solvating fluoride (i.e. when the F^- is fully solvated) we had not seen convergence to a single value, especially in the case of MeCN and DMSO. Initial conformational sampling was attempted in CREST; however, these calculations gave unreliable results using implicit solvent where most clusters broke apart. This led us to pursue an MD-based strategy, for detailed information on the implementation of the MD simulations see section 4.4.4 .

This would allow for the simulation of fluoride in a range of solvents to gain information about the coordination environment around fluoride. Using the radial distribution function (RDF) we could extract clusters and identify if MD solvation would result in a more accurate prediction

of the ^{19}F NMR. After MD simulation the RDF showed between 5-8 solvent molecules in the first solvation shell, so we clustered the first eight solvent molecules around the fluoride and extracted the most populated cluster before performing DFT calculations.

We calculated the σ_{iso} before and after DFT optimisation, and as to be expected the values before optimisation do not correlate with either the results from our 6-coordinate complexes or experimental results.

After optimisation, the computed values correlate well with experimental values with an R^2 = 0.75 (Figure 52A). The values of protic solvents match very well with those obtained from our 6-coordinate complexes, however, the only differences which are greater than 4 ppm are those of MeCN, DMSO and Acetone.

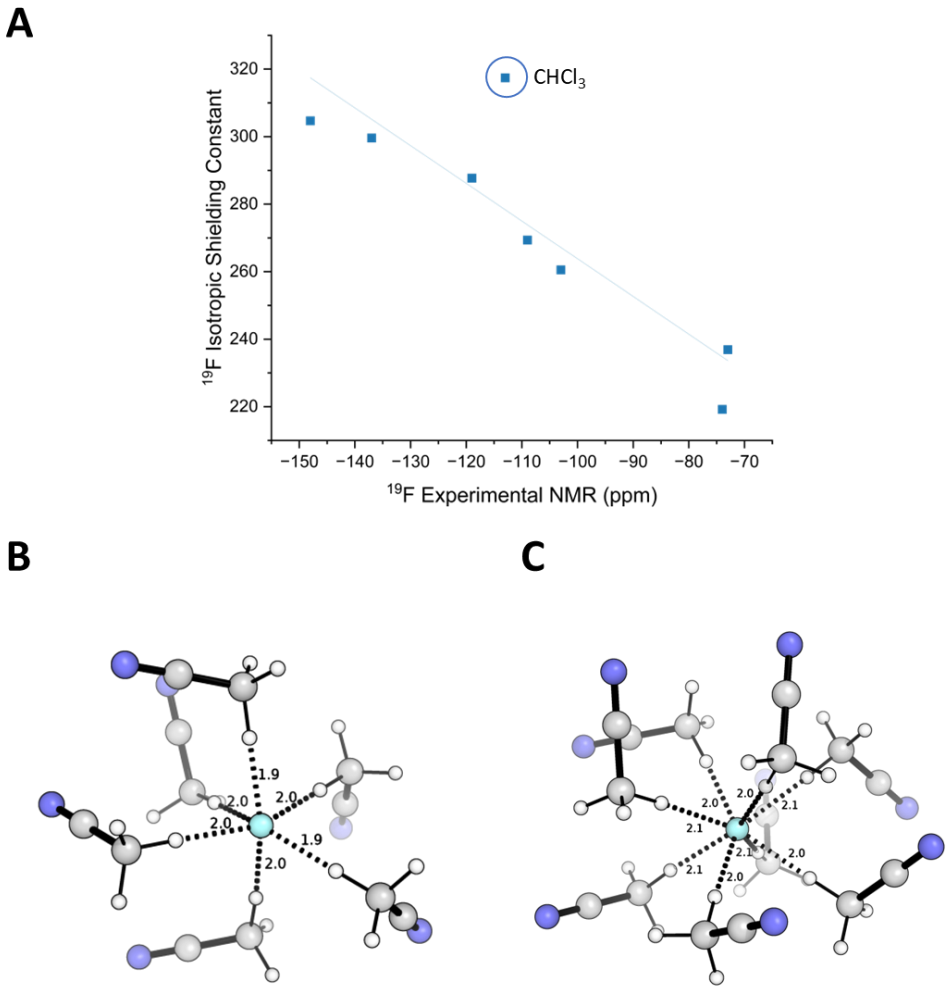


Figure 52 A) Comparison of calculated ^{19}F σ_{iso} to experimental data $R^2 = 0.75$. B) Hexacoordinated fluoride by MeCN, distances in Å C) Octacoordinated fluoride by MeCN obtained from MD simulations. All calculations performed using CPCM(Solvent)-M062X/def2-QZVP// CPCM(Solvent)-PBE0(D3BJ)/ma-def2-TZVPP.

In the case of MeCN, the reason for this large discrepancy in ppm is obvious. After optimisation, we have an 8-coordinate complex rather than a 6-coordinate one (Figure 52B vs Figure 52C). Furthermore, the average H-F bond distance is 2.05 Å compared to 1.96 Å for the 6-coordinate complex. These slight changes in the bond distances but an increase in the number of hydrogen bonds result in the fluoride becoming deshielded, so the σ_{iso} drops by a further 10 ppm.

In the cases of DMSO and Acetone, the major difference is due to different conformations and methods of coordination. In our built complexes both DMSO and Acetone acted as bi-dentate

ligands, however from the MD simulations these solvent molecules can also act as mono-dentate ligands. This change in conformation is therefore responsible for the observed differences in ^{19}F σ_{iso}.

While MD simulations do generate more diverse conformations, and consequently a more accurate description of the solvation around the anion, the obtained values for most solvents are similar to our empirical work of modelling solvents around the fluoride anion. Therefore, MD simulations are not required to predict the NMR shift computationally.

4.2.3 Micro-solvation

Given the importance of explicit solvation in accurate ^{19}F NMR calculations of fluoride in solution, we decided to automate the procedure of adding solvent molecules. We achieved this using the protocol developed by Joseph Silcock, a Part II in the Duarte group. Here a sphere of explicit solvent molecules is constructed around a molecule. This code was extended to use the CPCM surface of the molecule generated from PCMSolver.[282] This calculates the CPCM surface of a molecule and passes the centre of each tesserae to our Python code. Solvent molecules are then placed on this surface until the CPCM surface is covered. A UFF[136] forcefield optimisation is then performed before DFT optimizations can be performed. This code also includes the functionality to solvate part of the molecules as demonstrated in Figure 53A. This micro solvation allows users to specify an atom to solvate along with the radius from that atom to be considered. Therefore, only tesserae that fit within this criterion are solvated.

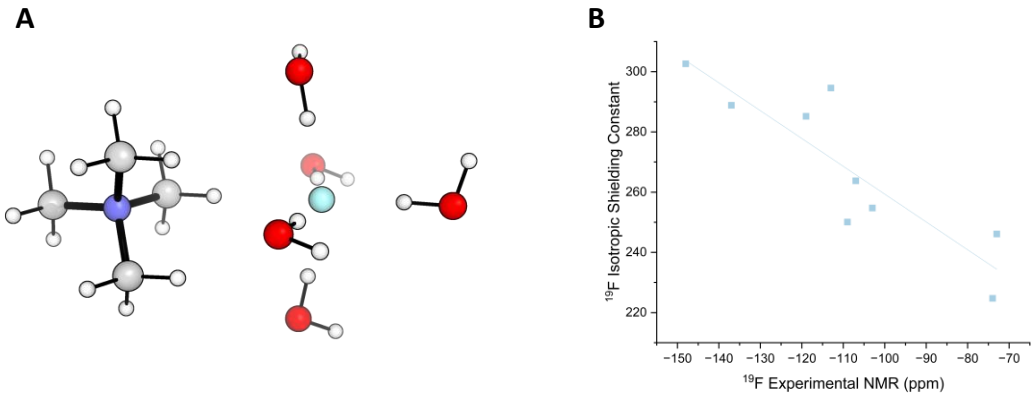


Figure 53 A) Output of Microsolvation code on TMAF using water. B) Comparison of calculated ^{19}F σ_{iso} against the experimental ^{19}F NMR for each of the solvated TMAF complexes. $R^2 = 0.77$.

With the ability to microsolvate part of a TMAF complex, we then generated seven micro-solvated TMAF complexes with varying solvents. The NMR shift was calculated after optimisation. Overall, the model works reasonably well, with an $R^2 = 0.77$ and $CHCl_3$ remaining as a significant outlier in the data (Figure 53B). Given that TMA is usually a non-coordinating cation, it is not surprising that its inclusion in the complex does not lead to an improvement in the R^2 compared to that of the MD solvation work, or the empirical solvation work.

Overall, in this section we have shown that while the approximation of one solvent molecule and a fluoride anion to predict the ^{19}F NMR of the anion does work but produces unrealistic geometries. While MD solvation and micro-solvation correlate with experimentally measured values, the most effective and efficient method is explicit solvation with six solvent molecules, which correlates best with experimental results while being less computationally expensive.

4.2.4 Mono urea fluoride complexes

With a method in hand to achieve reliable ^{19}F NMR values for fluoride in different solvent environments, we were interested in identifying if this approach could be extended further to complexes where the solvent did not change between complexes. We chose the Mono Urea Fluoride (MUF) complexes reported by the Gouverneur group as a starting point. While no

experimental NMR values have been reported to date, we believe that this system will supply a qualitative example of the effects of varying hydrogen bond donors coordinated to the F^- anion (Figure 54A).

Using our previous method, we used our micro-solvation script to solvate 8 different MUF complexes with DCM. These complexes were then optimised at the PBE0(D3BJ)/def2-TZVPP level of theory before $^1J_{HF}$ coupling constants were calculated at the B97-D3/def2-QZVP level of theory and ^{19}F NMR σ_{iso} were calculated at the M062X/def2-QZVP level of theory. To understand how explicit and implicit solvation affect the σ_{iso} and $^1J_{HF}$ coupling constant, we compared two different sets of structures: MUF complexes with or without explicit solvation (Figure 54A and Figure 55A). For both sets of complexes we performed optimisations either in the gas phase or with implicit solvation. This therefore resulted in four different sets of results.

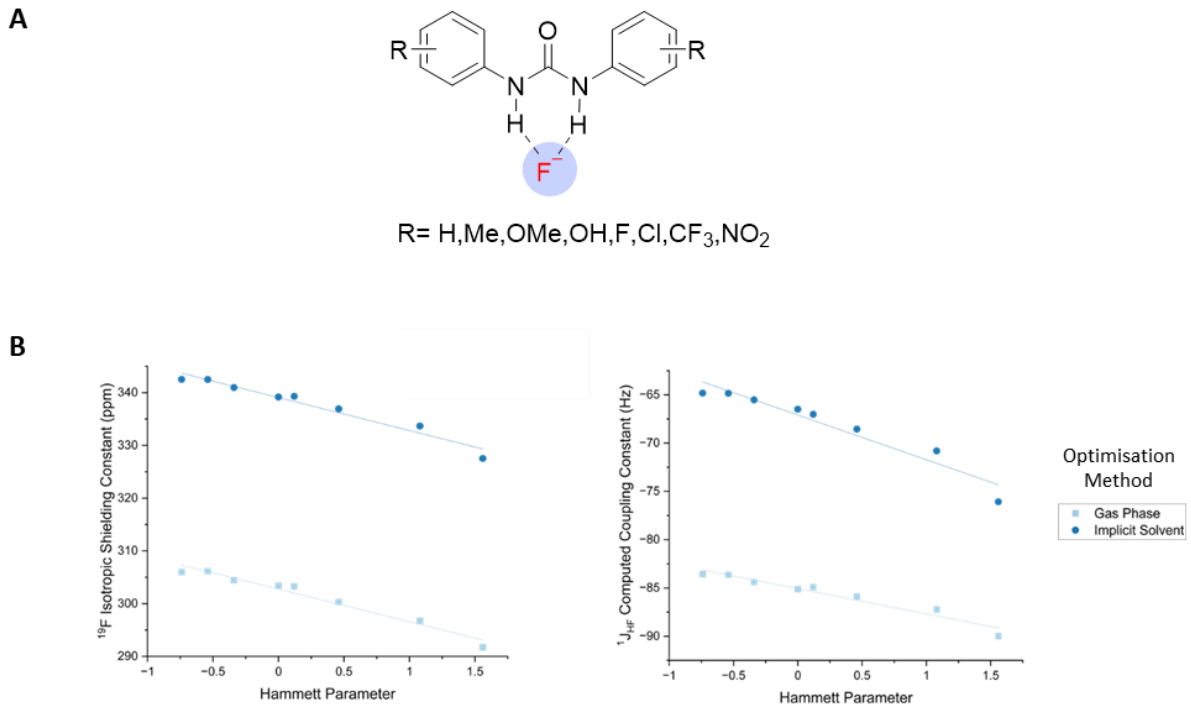


Figure 54 A) MUF complexes, without explicit solvation. B) The effect that including implicit solvation during optimisation has on the ^{19}F σ_{iso} and the $^1J_{HF}$ coupling constant on MUF Complexes when compared to the Hammett Parameters for each complex. ^{19}F σ_{iso} gas phase optimisation $R^2 = 0.96$ Implicit Solvent optimisation $R^2 = 0.96$. For $^1J_{HF}$ coupling constants gas phase $R^2 = 0.95$ and implicit solvent $R^2 = 0.93$

We plotted against the 2σpara Hammett Parameters for the non-solvated complexes for each complex against the ^{19}F σ_{iso}. We observe a very good correlation between the Hammett Parameters and both the ^{19}F σ_{iso} ($R^2 = 0.96$ for both gas phase and implicit solvent optimisations Figure 54B) and the $^{1}J_{HF}$ coupling constants ($R^2 = 0.95$ for gas phase optimisations and $R^2 = 0.93$ for implicit solvent optimisations). The major difference however is in the absolute values for each of the complexes (Figure 54B).

The difference between the gas phase and implicit solvation optimisations is significant for both the σ_{iso} and $^{1}J_{HF}$ coupling constant. The σ_{iso} are generally about 40 ppm more deshielded for gas phase calculations, while the coupling constants are usually 20 Hz more negative. These two patterns however should be expected. In the gas phase calculations, as seen with our work on solvent clusters in the previous section, the H-F bond is shorter usually by 0.1 Å which results in the electron density on the fluoride anion being reduced, thus deshielding the nucleus. This shorter H-F bond means that there is a much stronger coupling between the two atoms, which increases the magnitude of the $^{1}J_{HF}$ coupling constant.

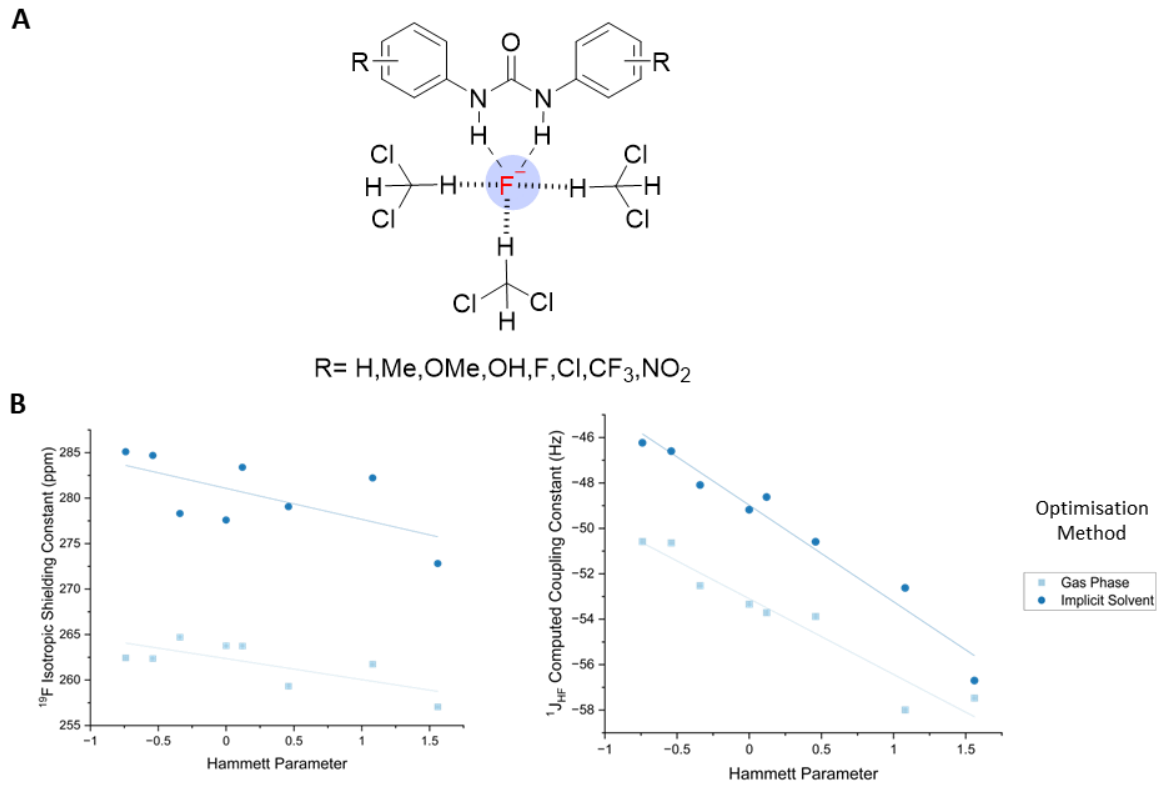


Figure 55 A)MUF complexes, with explicit solvation B)When compared to the Hammett Parameters for each complex, the effect that including explicit solvation has on the ^{19}F σ_{iso} constant and the $^{1}J_{HF}$ coupling constant on MUF complexes. ^{19}F σ_{iso} gas phase $R^2 = 0.52$ and 0.42 for gas phase and implicit solvation respectively. For $^{1}J_{HF}$ coupling constants $R^2 = 0.93$ for gas phase and $R^2 = 0.95$ for implicit solvation.

The first observed difference between the calculations performed with explicit solvent molecules is the range of calculated ^{19}F σ_{iso} and $^{1}J_{HF}$ (Figure 55B). These are 50 ppm and 40 Hz smaller respectively than compared to calculations performed in the absence of explicit solvent. This is consistent with our work on solvent-fluoride complexes where explicit solvation reduced the electron density of the fluoride through solvent-fluoride interactions. We see this when comparing the H-F bond lengths which are similar between optimisations in the gas phase and with implicit solvent, with only a difference of 0.03 Å. However, these are still longer by > 0.1 Å than complexes with no explicit solvation.

While the difference in hydrogen bond length between gas phase and implicit solvent optimisations is small, the difference in ^{19}F NMR σ_{iso} is still pronounced at 20 ppm. This difference is a result of the solvent molecules being closer to the fluoride anion in gas phase calculations, resulting in the extra deshielding of the ^{19}F σ_{iso}.

When the σ_{iso} is plotted against the $2\sigma_{para}$ Hammett Parameters there is a quite poor correlation (R^2 = 0.52 and 0.42 for gas phase and implicit optimisations), yet the correlation between $^1J_{HF}$ coupling constant and Hammett values remain with an R^2 = 0.93 and 0.96 for gas phase and implicit optimisations respectively.

To understand if the lack of correlation observed with the NMR shifts was due to a lack of conformational sampling of the MUF complex and the solvent, we decided to use CREST to generate conformers which could be then Boltzmann weighted. To Boltzmann weight conformers we used the CPCM(CH_2Cl_2)-ωB97X-D3/def2-QZVP, as this has previously been benchmarked for reproducing binding in urea fluoride systems.[38, 283]

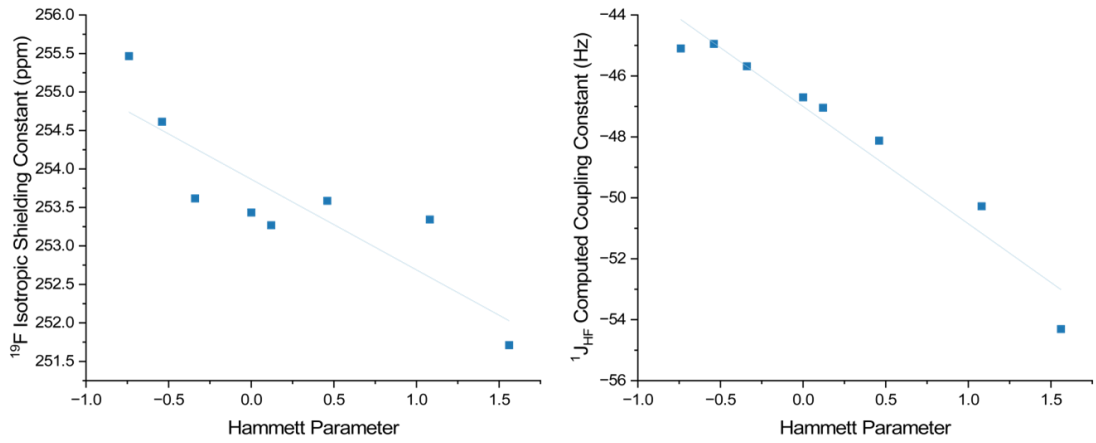


Figure 56 Boltzmann weighted ^{19}F σ_{iso} and $^1J_{HF}$ coupling constants for the MUF complexes. All optimisations were carried out with implicit solvation. For NMR calculations R^2 = 0.75 and for $^1J_{HF}$ calculations R^2 = 0.94

With Boltzmann weighted NMR and shifts and structures, we see that the R^2 between the ^{19}F σ_{iso} and the Hammett values improves to 0.75, however there is still a large spread in the data (Figure 56). We would expect that there would be some correlation between the Hammett Parameters and the NMR shift as previous work has shown that Hammett Parameters do correlate with the binding energy of the fluoride anion to the urea.[283] This stronger binding we would expect to see a difference in the NMR parameters of the complex.

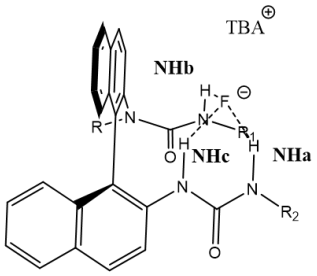
We observe that the range of the σ_{iso} are very small, with only 4 ppm between the most extreme data points. This suggests that the majority of the deshielding of the fluoride nucleus is coming from the DCM solvent molecules, and therefore there is little influence on the shielding of the fluoride anion from the urea. Therefore, future changes in the electronics on the urea backbone will not have a significantly discernible influence on the ^{19}F σ_{iso}.

In the case of the $^{1}J_{HF}$ coupling constants, these do remain and correlate even when we Boltzmann weight the conformers with $R^2 = 0.94$ (Figure 56). This shows that Boltzmann weighting of conformers has a small effect on the $^{1}J_{HF}$ coupling constants, with differences of less than 2 Hz between fixed geometries and conformational sampling.

In conclusion, we have shown that implicit solvation and explicit solvation around the fluoride anion in MUF complexes result in large differences of up to 100 ppm while the $^{1}J_{HF}$ coupling constants change by over 20 Hz. These differences are due to changes in the H−F bond length which vary based on the solvation method, inclusion of implicit solvent, and urea electronics. However, in all cases the $2\sigma_{para}$ Hammett Parameters do correlate with both the σ_{iso} and $^{1}J_{HF}$ coupling constant with $R^2 > 0.75$ showing that electronics will correlate with both NMR parameters within each computational method, which is supported by previous work on the calculation of MUF binding energies.[38, 283] The end result is that while the absolute values for both the coupling constants and σ_{iso} may have a larger error, the relative trends are still preserved.

4.2.5 BINAM urea complexes

With a method in hand for the prediction of ^{19}F NMR shifts and the $^{1}J_{HF}$ coupling constant, we wished to apply this method to that of the HBPTC catalysts. First, we wished to compare the impact that diffuse functions in geometry optimisations would have on the calculated $^{1}J_{HF}$ coupling constants. Optimisations were carried out with the CPCM(CH_2Cl_2)-PBE0(D3BJ)/def2-TZVPP or CPCM(CH_2Cl_2)-PBE0(D3BJ)/ma-def2-TZVPP level of theory. $^{1}J_{HF}$ coupling constants were then calculated using CPCM(CH_2Cl_2)-B97-D3/def2-QZVP and the values Boltzmann weighted according to their energies at the CPCM(CH_2Cl_2)-ωB97X-D3/def2-QZVP level of theory. Overall, there is only a maximum deviation of 2.3 Hz between the two methodologies and an MAE of 1.4 Hz (Figure 57). This implies that diffuse functions show no significant difference in the prediction of $^{1}J_{HF}$ coupling constants on these BINAM urea complexes. For that reason, we therefore performed all other calculations using the CPCM(CH_2Cl_2)-PBE0(D3BJ)/def2-TZVPP methodology to reduce computational cost.

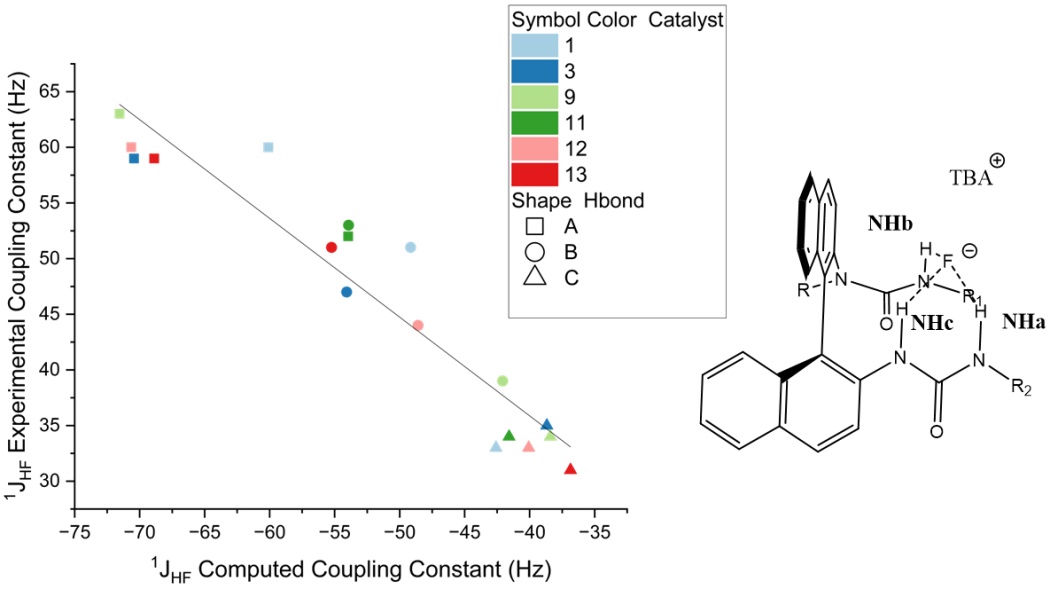


	NHa (Hz)	NHb (Hz)	NHc (Hz)
def2-TZVPP	-60.11	-49.15	-42.59
ma-def2-TZVPP	-62.41	-49.32	-40.80

Figure 57 Differences in $^{1}J_{HF}$ coupling constants when geometry optimisations include diffuse functions. Calculations are carried out at the same level of theory for NMR calculations and Boltzmann weighting. R = iPr R_1 = R_2 = 3,5-bis(trifluoromethyl)benzene.

Taking six catalysts from the work of Ibba *et al.*[39] with variable electronics on the R_1 and R_2 groups, we conformationally sampled each catalyst before Boltzmann weighting the $^{1}J_{HF}$ coupling constants. When plotted against the experimentally determined $^{1}J_{HF}$ our approach performs well with an R^2 = 0.90 to the experimental data (Figure 58). Furthermore, the sign of the coupling constant is negative, in line with previous computational and experimental studies

for coupling constants across a H-F hydrogen bond.[71, 284] The relative strengths of the $^{1}J_{HF}$ couplings for the NHa, NHb, and NHc couplings are also preserved with NHa > NHb > NHc.



Catalyst	R1	R2
1	3,5-bis(trifluoromethyl)benzene	3,5-bis(trifluoromethyl)benzene
3	3,5-difluorobenzene	3,5-difluorobenzene
9	Phenyl	3,5-bis(trifluoromethyl)benzene
11	3,5-bis(trifluoromethyl)benzene	3,5-dimethylbenzene
12	3,5-difluorobenzene	3,5-bis(trifluoromethyl)benzene
13	3,5-bis(trifluoromethyl)benzene	3,5-difluorobenzene

Figure 58 Calculated $^{1}J_{HF}$ coupling constants using our DFT methodology categorised by catalyst and hydrogen bond. R = iPr for all catalysts. R^2 = 0.90, RMSE = 3.4 Hz, MAE = 2.8 Hz.

However, the model underperforms when it comes to the prediction of the exact magnitude of the coupling constants. In general, both the NHa and NHc coupling constants are higher than those experimentally determined, with NHc being 5-10 Hz larger and NHa being > 10 Hz larger than reported. However, when we apply a linear fit to the data, we can obtain an RMSE of 3.4 Hz and an MAE 2.8 Hz across all H-F hydrogen bonds. To investigate the reason for this difference, we further studied catalyst 3, for which H-F bond distances had been determined using HOSEY NMR experiments.

Comparison of H-F bond distances to experimental data

To experimentally determine the coordination environment of the fluoride anion, Ibba *et al.*[39] used quantitative HOESY experiments. The nuclear overhauser effect (NOE) could then be converted into distances using the known H-F vinyl distances. They determined that the H-F distances were 1.83 Å, 1.91 Å, and 2.05 Å for the NHa-F, NHb-F, NHc-F distances respectively. In comparison, our calculations give distances of 1.58 Å, 1.67 Å, and 1.75 Å for the NHa-F, NHb-F, and NHc-F distances.

In all cases, the H-F bond distances are significantly shorter than expected, by at least 0.2 Å. Therefore, the fluoride is binding much more strongly to the complex than experimentally determined. This, coupled with the fact that B97-D3 $^{1}J_{HF}$ coupling constants underestimate the true experimental values, results in a cancellation of the errors, meaning that the values are closer than expected for the complexes with this short bond length.

Complex conformation effects

Inspecting the conformers that make up the Boltzmann weighted coupling constants we observed two different binding modes: an open and a closed conformation (Figure 59). In the closed conformation the TBA counter ion lines above the fluoride anion, shielding the anion from the solvent. The open confirmation is when the TBA counter ion sits below the urea oxygens. This results in the fluoride anion being fully exposed to the bulk solvent.

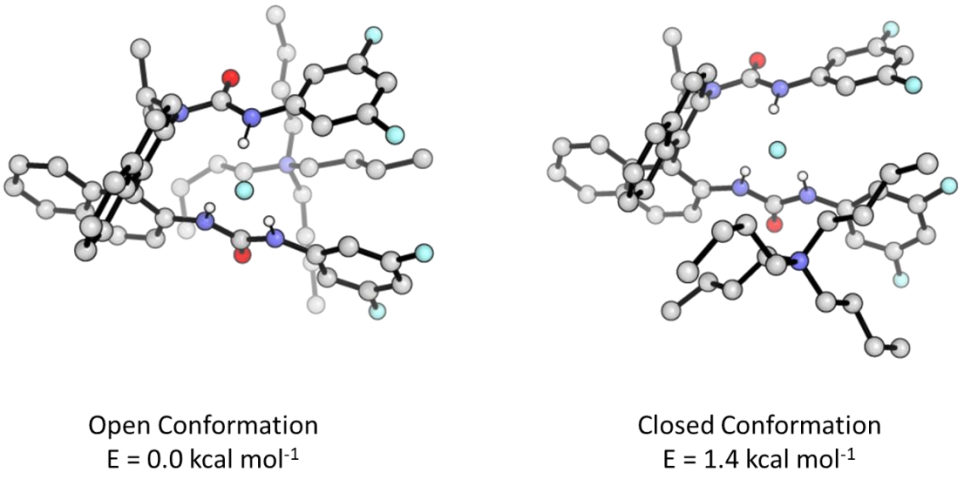


Figure 59 Examples of the open and closed conformation for Catalyst 3.

In the case of catalyst **3**, the lowest three conformers are all open conformations summing to 95% of the Boltzmann weighting, while the remaining 5% comes from the closed conformation. The major difference in the $^{1}J_{HF}$ coupling constants originates from the difference of the NHa H-F bond. In the case of the open conformation this sits slightly longer, 1.57 Å vs 1.61 Å, yet this difference results in a change of -71 Hz to -62 Hz in the coupling constant. Therefore, this slight variation in the conformation and the change in H-F bond distance can result in large changes in the calculated coupling.

Reasons for overbinding

As noted in the previous section, the computational workflow currently results in the overbinding of the fluoride anion to the urea protons, resulting in distances significantly shorter than those experimentally determined. We believe that one of the major problems comes from the lack of inclusion of explicit solvent when considering these structures.

This is very pronounced in open conformations, where the anion is exposed to the bulk solvent. As seen with both the solvent clusters and MUF complexes, not including explicit solvation leads to shorter H−F distances than when it is included. In the case of the MUF complexes, differences of more than 0.1 Å are observed for the H-F hydrogen bond with explicit solvation.

This importance of explicit solvation is noted in unpublished work by David Ascough,[283] which showed that to reproduce binding constants for BINAM urea an explicit water molecule must be included above the fluoride.

Therefore, while our model can reproduce the trends seen with experimental data, we cannot reproduce the exact structures in solution using these methods.

4.3 Conclusions

Herein we present an approach for the calculation of ^{19}F NMR for the fluoride anion in a range of solvents. The approach can calculate both the σ_{iso} along with relative $^{1}J_{HF}$ coupling constants across the H-F hydrogen bond. However, in all methods DFT calculations underestimated the $^{1}J_{HF}$ coupling constants, especially for the covalent H-F bond where errors are greater than 200 Hz. While absolute values are not easily calculable, the relative values of the $^{1}J_{HF}$ coupling constants can be calculated.

In the study on MUF complexes, we find that changes in $^{1}J_{HF}$ coupling constants do correlate with electronic changes on the urea backbone. Furthermore, we also investigated the impact that explicitly solvating the fluoride has on both the ^{19}F NMR shift and the $^{1}J_{HF}$ coupling constant. What we discovered is that addition of explicit solvents significantly changes the absolute value of the σ_{iso} and the $^{1}J_{HF}$ constant and preserves the correlation between the Hammett parameter and $^{1}J_{HF}$ coupling constants. Correlations remain for ^{19}F σ_{iso}, but the differences become very small, showing that the shielding from the solvent has a much larger contribution to the overall shielding than that from the urea.

When our approach is applied to the BINAM complexes we find that our approach can predict the $^{1}J_{HF}$ coupling constants well with a fitted RMSE of 3.4 Hz, meaning we can differentiate between all H-F hydrogen bonds accurately. We also show that two conformations can be found, an open and a closed form, where the fluoride is either exposed to the bulk solvent or covered by the TBA cation with the open conformer being the major conformer.

Given our results on the effect that including explicit solvent has on the $^{1}J_{HF}$ coupling constants, it is therefore not a surprise that our model shows significant overbinding to the fluoride compared to experimental data. This shows that explicit solvation is essential to accurately calculate the $^{1}J_{HF}$ coupling constant. Sampling in the presence of explicit solvent is an area which future work should be directed towards.

4.4 Methods

4.4.1 Quantum mechanical calculations

Quantum mechanical calculations were carried out using ORCA 5.0.3[285] unless otherwise stated. Minima were characterised using the TightOPT keyword, corresponding to tolerances of 10^{-8} Hartrees for the SCF energy change, and 10^{-6} Hartrees for the optimization step. To speed up calculations the resolution of identify chain of spheres exchange (RIJCOSX) approximation[286] was employed for all hybrid and meta-hybrid DFT functionals and the RIJ[287] approximation for all non-hybrid DFT functionals with the AutoAux[288] keyword used for the generation of the needed auxiliary orbitals for the RIJ and RIJCOSX approximations.

The integration grid (DEFGRID3) was used in all calculations in ORCA 5.0.3, and calculations in ORCA v 4.2 used integration grids Grid 7 corresponding to a Lebedev-770 angular grid, and a radial integral accuracy (IntAcc) of 5.67. Calculations employing a RJCOSX approximation utilized the "GridX7" procedure, corresponding to IntAcc = 4.34 and a Lebedev-302 angular grid.

MP2 calculations make use of the Resolution of Identity Approximation (RI) for the calculation of correlation integrals.[289] CCSD(T) optimisations made use of the numerical gradients (NUMGRAD) and the TightOpt criteria. The solvent was described using the implicit solvation methods CPCM[121, 290-292] and SMD,[123] CPCM solvation made use of the Gaussian charge scheme.

Dispersion corrections were considered using Grimme's D3 empirical method with Becke-Johnson damping (D3BJ)[103, 104, 293] with the exceptions of B97-D3, B97-3c, PBEh-3c and ωB97XD3 which apply their own dispersion correction. CCSD(T) optimizations were performed with Dunning's correlation consistent basis sets[103, 104, 293] using the numerical gradients (NUMGRAD) keyword and the TightOpt criteria.

The Aldrichs type def2 basis sets were used extensively throughout the DFT study and includes an effective core potential (ECP) for Cs.[105, 273] For $^{1}J_{HF}$ coupling constants, Jensen's optimised coupling constant pcJ-X basis sets[107] were also tested.

NMR calculations used the GIAO method[294] to ensure gauge invariance of the NMR shielding tensors. Unless otherwise stated distances are given in Angstroms (Å), angles in degrees, and NMR tensors (σ) in ppm.

4.4.2 Geometry scans

For geometry scans the optimised dimer xyz coordinates were used as starting coordinates. For the distance scans, the H-F bonds were scanned between 1 and 3 Å in steps of 0.2 Å, with the H-F-H angle constrained to its initial value. In the case of CH_2Cl_2 the C-H-F angle for each solvent molecule was also constrained to stop molecular rotation. The H-F bond length was constrained for the angle scans while the H-F-H angle was scanned between 10 and 170 degrees. In each case, a relaxed scan was performed at the CPCM(Solvent)-PBE0-D3BJ/ma-def2-TZVPP level of theory followed by NMR calculations using CPCM(Solvent)-M062X/def2-QZVP.

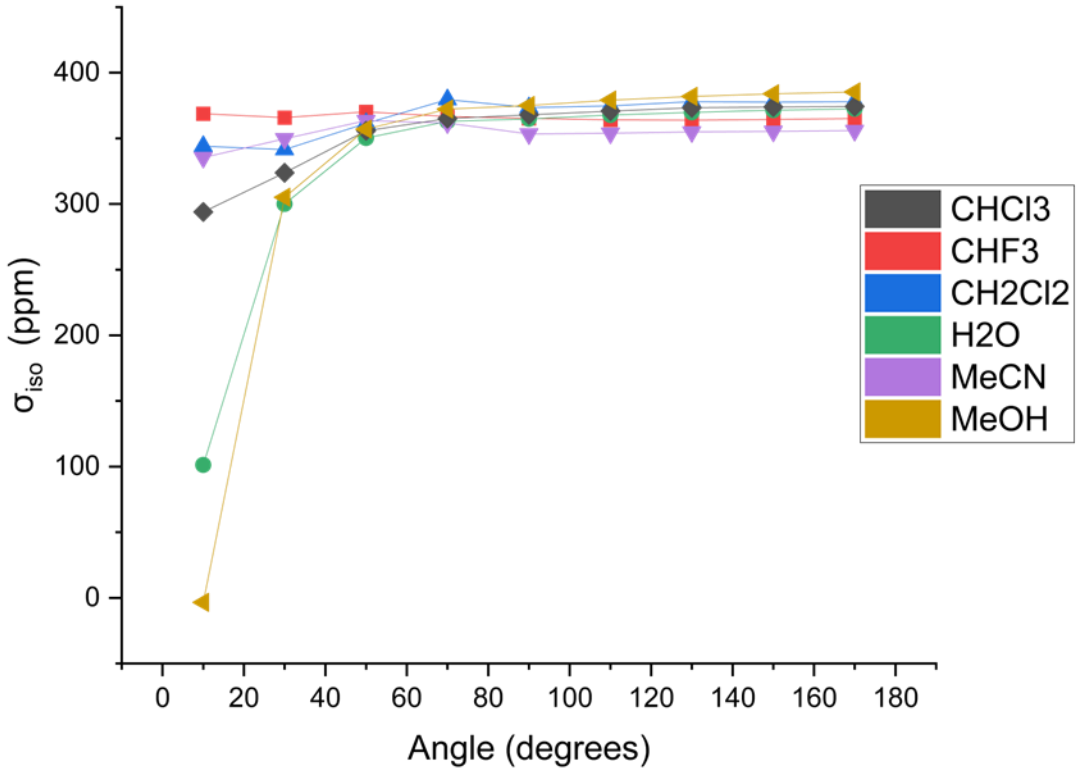


Figure 60 Geometry scans for the solvent fluoride dimers across the H-F-H angle. Calculations performed at the CPCM(Solvent)-M062X/def2-QZVP// CPCM(Solvent)-PBE0(D3BJ)/ma-def2-TZVPP level of theory.

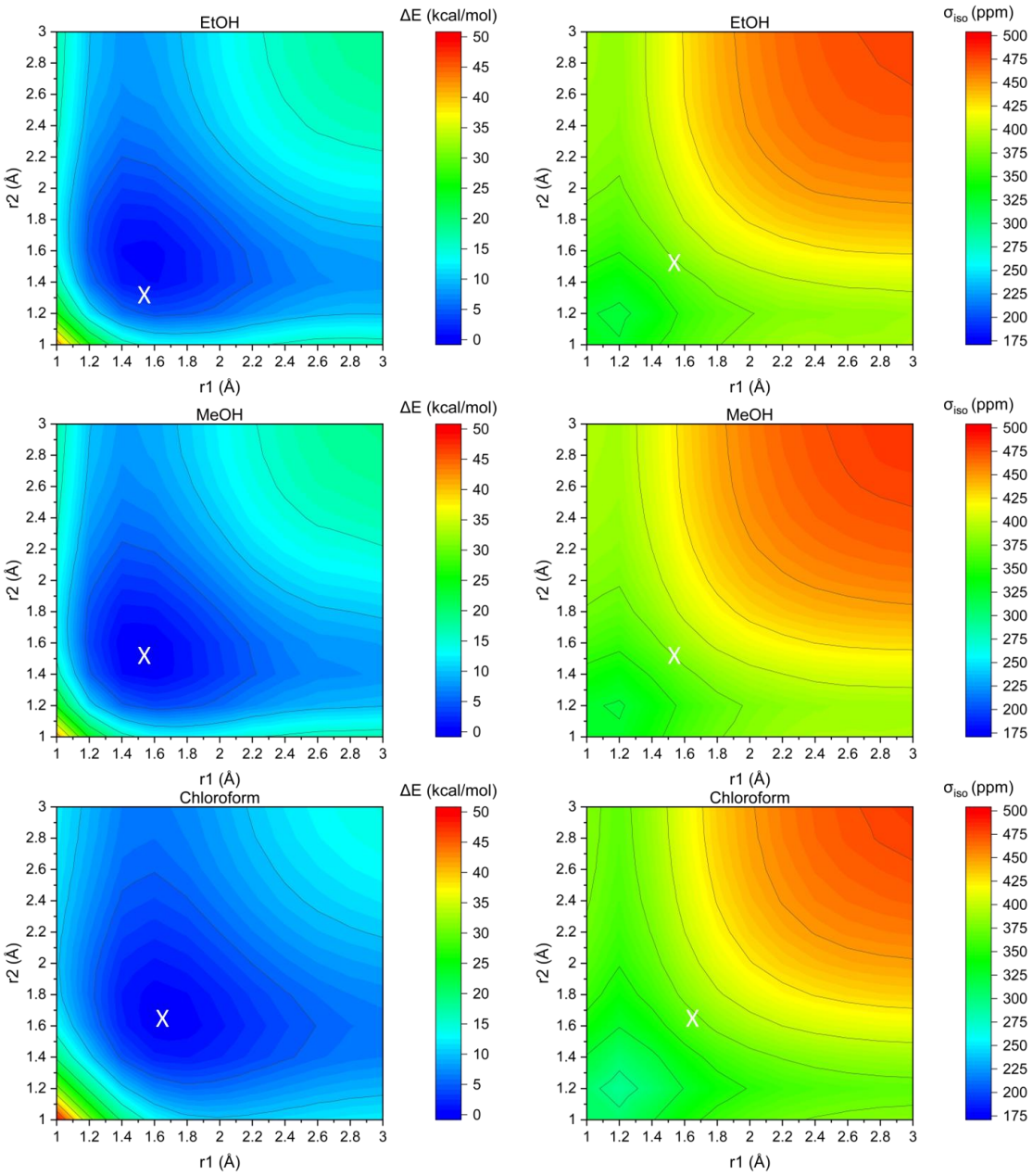


Figure 61 Extra 2D geometry scan for EtOH, MeOH, and DCM dimer across the r1 and r2 H-F bonds. Calculations performed at the CPCM(Solvent)-M062X/def2-QZVP// CPCM(Solvent)-PBE0(D3BJ)/ma-def2-TZVPP level of theory. X signifies the PES minima. Conor lines are shown for every 5 kcal/mol and 25 ppm.

4.4.3 Boltzmann weighted constants

For the generation of Boltzmann weighted ^{19}F σ_{iso} and $^{1}J_{HF}$ coupling constants the following workflow was carried out for MUF and HBPTC complexes.

CREST conformational sampling (CREST v 2.11[51, 295] and xTB v 6.4[50, 131, 296]) was carried out for each complex using the albp solvation method and the nci mode.

An RMSD cut-off for 1 Å on all atoms was performed before CPCM(CH_2Cl_2)-PBE0-D3BJ/def2-TZVP singe point was performed. A Boltzmann weighting was performed and those conformers which summed to 90% of the total energy were then optimized at the CPCM(CH_2Cl_2)-PBE0-D3BJ/def2-TZVPP. $^1J_{HF}$ coupling constants were then calculated using CPCM(CH_2Cl_2)-B97-D3/def2-QZVP and ^{19}F σ_{iso} were calculated at the CPCM(CH_2Cl_2)-M06-2X/def2-QZVP level of theory. The values from each conformer were then weighted according to their energy calculated at CPCM(CH_2Cl_2)-ωB97X-D3/def2-QZVP.

For MUF complexes we excluded any conformers where the anion was not solvated by each of the explicit CH_2Cl_2 molecules.

4.4.4 MD simulations

Sampling of the solvent anion clusters were performed using the GROMACS (version 2020.1, mixed precision)[297, 298] molecular dynamics package with the OPLS-AA forcefield.[138]

The complex of study was placed in a cubic box with a 3-dimensional periodic boundary condition (PBC), with a minimum boundary distance of 1.2 Å. A cut-off of 1 nm was used for both the Van der Walls interactions and long-range electrostatics were described with the Particle Mesh Ewald (PME) algorithm.[299] Simulations made use of the linear constraint solver algorithm (LINCS).[300] The temperature of the system was maintained at 298.15 K using the V-rescale thermostat[301] and a coupling constant of 0.1 ps. The pressure was controlled by the Parrienello-Rahman barostat at 1.0 bar, with an isothermal compressibility of 4.5 × 10-5 bar^{-1}.[302, 303]. Coordinates were saved every 10 ps, and the trajectories were concatenated for subsequent analysis when necessary.

Each complex was minimised by steepest descent minimization for a maximum of 50000 steps or until the maximum force was less than 1000 kJ mol^{-1} nm^{-1}. Velocities were taken from the Maxwell distribution at 298.15 K. The system was equilibrated under constant volume (NVT)

for 100 ps and a time step of 1 fs with position restraints. The system was then equilibrated under constant pressure (NPT) for 200 ps with a step of 1 fs. Production MD was run using the Parrinello-Rahman coupling for pressure coupling, a time step of 1 fs, and a length of 1 ns.

Each complex was subjected to the NVT-NPT-MD run before clustering. Clustering made use of the GROMOS clustering algorithm[304] with a cut off of 0.2 nm, allowing for PBC effects. The median of the most populated cluster was then used for carrying out DFT optimisations at the CPCM(Solvent)-PBE0-D3BJ/ ma-def2-TZVPP level of theory, before NMR calculations were carried out at the CPCM(Solvent)-M062X/def2-QZVP level of theory.

Attempts were made to use MD simulations to conformationally sample the HBPTC catalyst complexes, however, these either ended in complex disassociation or structures not in agreement with experimental data. This we believe is due to the forcefield not being optimised for this system and therefore was not investigated further.

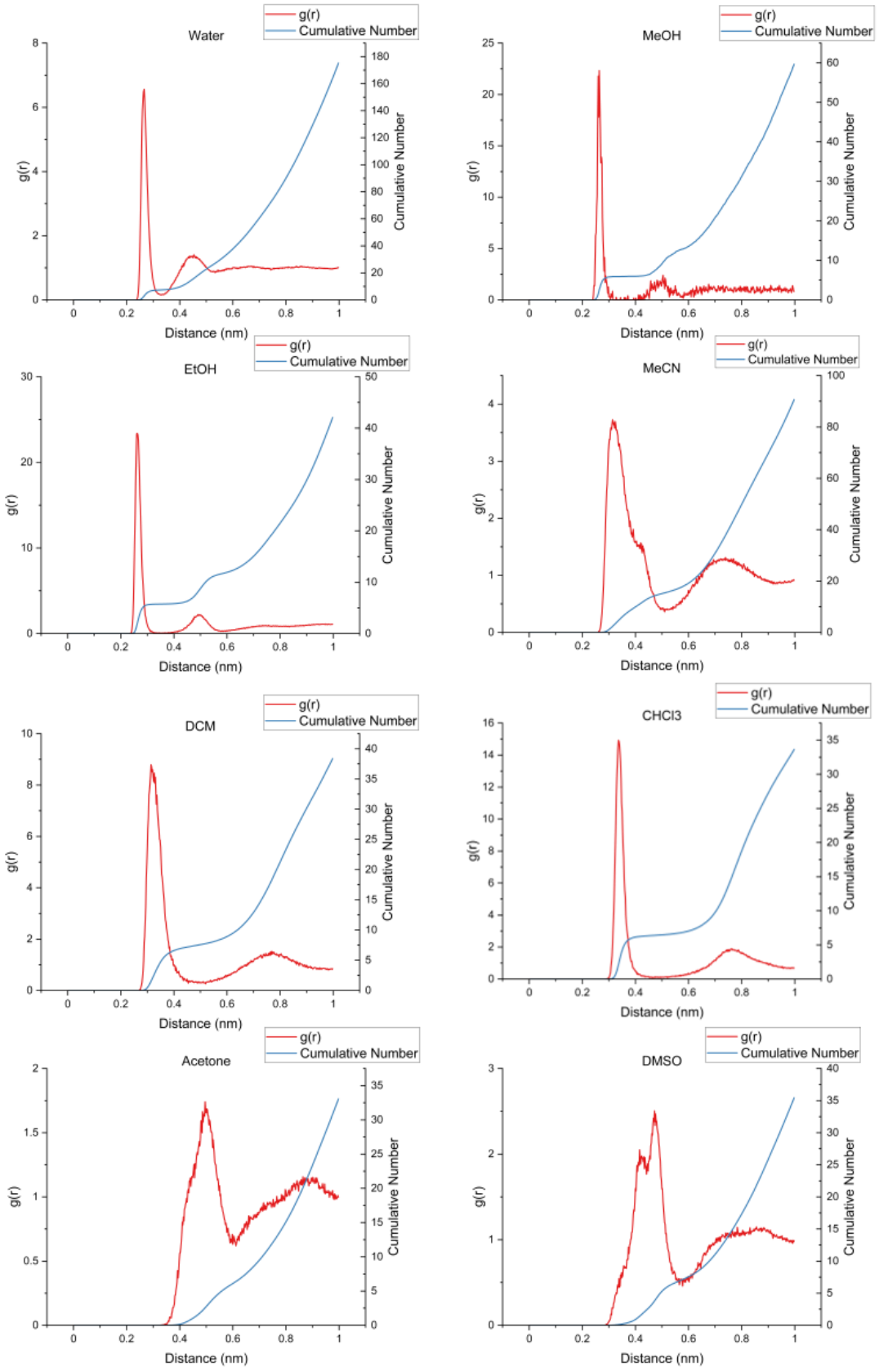


Figure 62 Radial distribution functions for MD simulations of fluoride in a box of solvent. Red line shows radial distribution as a function of distance; blue line shows the total number of molecules at that distance.

Chapter 5 Translating Chemical Structures: Using Machine Learning to Predict NMR

5.1 Abstract

Nuclear Magnetic Resonance (NMR) Spectroscopy is a key experimental technique for the structural elucidation of molecules. Computational NMR calculations can aid in the assignment of structures or even correct misassigned structures. While significant advances in computing power have enabled the wide use of high-level electronic structure methods for NMR shift prediction, these methods remain expensive. Machine Learning (ML) models promise to bridge this gap, reaching chemical accuracy at a fraction of the computational cost. Herein, a transformer-based ML-NMR model has been developed to predict ^{1}H,^{13}C and ^{19}F NMR with Root Mean Square Errors (RMSE) of 0.55, 3.3, and 8.2 ppm respectively. Similarity scores and Integrated Gradients (IGs) are used to interrogate the model and understand the chemistry that is learnt during training. Our results show that the NMRBERT model can identify key structural motifs and the effect they have on NMR shift, before combining the individual parts to give an overall prediction for the NMR shift of a given nucleus. The model is then applied to identify the correct product of regioselective late-stage fluorination reactions, for a range of fluorinated aromatic regioisomers alongside a confidence score for the prediction, however the model is unable to discern between diastereomers and aromatic trifluoromethyl groups.

5.2 Results and Discussion

As shown in Chapter 2, a large body of work on the prediction of ^{1}H and ^{13}C NMR using ML techniques has been carried out, which can achieve the accuracy of DFT methods at a fraction of the computational cost and time. As we found with our work described in Chapter 4, the calculation of ^{19}F NMR with a high level of accuracy can be both time-consuming and computationally expensive, and therefore, we were interested in developing our own method using ML techniques. Building on the work on ML screening, seen in Chapter 3 and ^{19}F NMR prediction using DFT in Chapter 4, we developed NMRBERT, a modified BERT transformer, to predict ^{19}F NMR shifts, purely from SMILES strings. We then use attribution techniques, introduced in Chapter 2; to understand the predictions that the model is making and identify limitations with using this type of architecture.

5.2.1 ^{19}F NMR prediction

Screening of Machine Learning Algorithms

Model development began using a subset of our ^{19}F dataset (358 mono-fluorinated molecules), further detail on dataset generation and model parameters, including dropout percentage and hyperparameters can be found in Sections 5.4.1-5.4.4. We began by testing two BERT-based models using either ensemble cross-validation or data-splitting to reduce model overfitting. The five-fold cross-validated model performed well with an initial error of 20 ppm (RMSE), while the five-fold split performed slightly worse with an error of 26 ppm (RMSE Figure 63).

To compare the error of the BERT models to other widely used machine learning models we then screened a range of ML models, using both Fingerprints and SOAP descriptors and standard ML algorithms as described in Section 5.5, the results are shown above in Figure 63.

For molecular fingerprints as descriptors, the best performing models were SVM Linear, Random Forests and Bagged Trees, all with errors of 20 ppm, whereas while higher order

SVM's such as Quadratic, Cubic, and Gaussians Processes all perform poorly with errors above 40 ppm. For models trained on SOAP descriptors, the lowest errors were obtained using a Fine KNN (RMSE = 16 ppm). Boosted Trees, Bagged Tree and Random Forests all have errors between 17 ppm and 20 ppm. Gaussian processes perform the worst on the SOAP dataset with errors above 40 ppm.

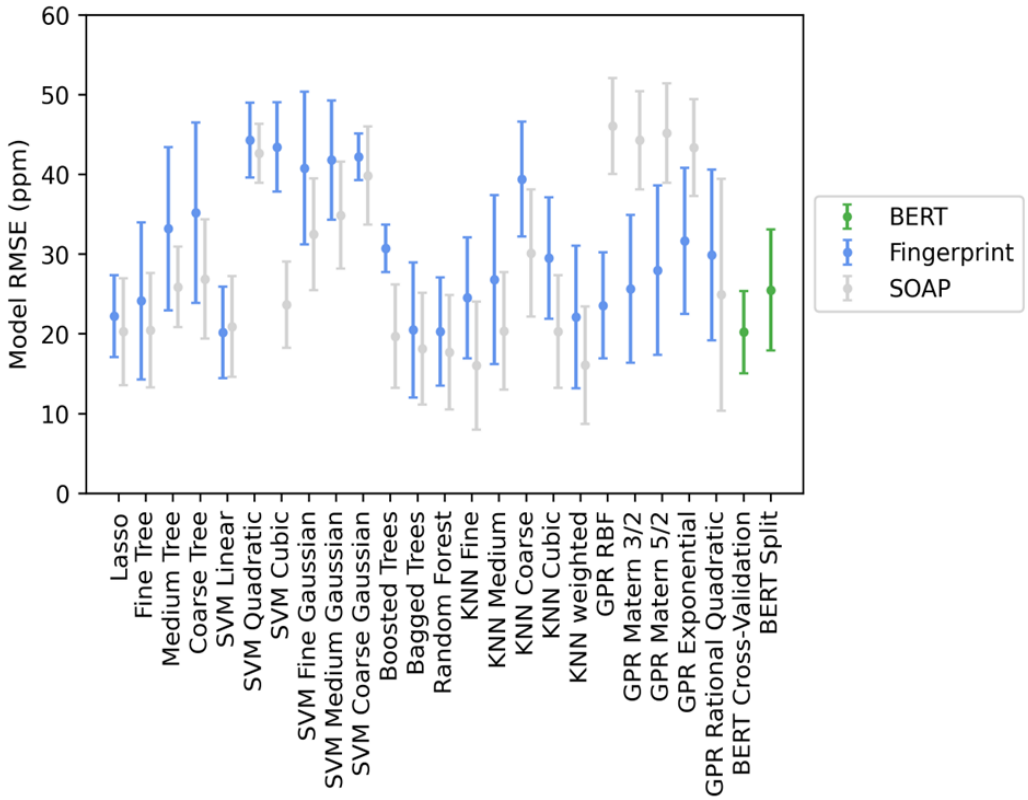


Figure 63 Comparison of ML algorithms using different descriptors. Fingerprints and SOAP descriptors were screened over a range of standard ML methods, while BERT was tested using two different dataset-splitting methods. Error is given in RMSE, and the error bars represent the Cross-Validation standard deviation.

Inspection of the training data for the BERT models showed that while the initial dataset covered a wide range of chemical groups, some such as SO_2F were sparsely populated. This group is uniquely placed in the 50-60 ppm range, and in the dataset, only five examples were present meaning that some of the five-fold splits had not seen this region of chemical space. This resulted in the model predicting the mean of the training data, -70 ppm which causes errors to be >100 ppm. To improve the model, we iteratively added more data points obtained from

a further literature search. This resulted in a reduction in the error of the model to 11 ppm (RMSE) on the mono-fluorinated dataset (Figure 64).

Molecular fingerprints while simple, are unable to be used for multi-fluorinated molecules as masking techniques are unable to generate different fingerprints for varying isotopes of the same molecule. SOAP descriptors show very good efficacy, however, require conformational sampling to ensure that the SOAP descriptors fully represent all possible environments which can be computationally expensive for large systems. The BERT model, however, rather than learning from scratch has already been pretrained on chemical data[152, 153, 305] and therefore it would only require finetuning to predict NMR, and furthermore only uses SMILES strings meaning there is no need for conformational sampling. Using a BERT transformer would allow for further study on the limits that SMILES based transformers could be applied to chemical problems, we therefore took the BERT model forward for further study.

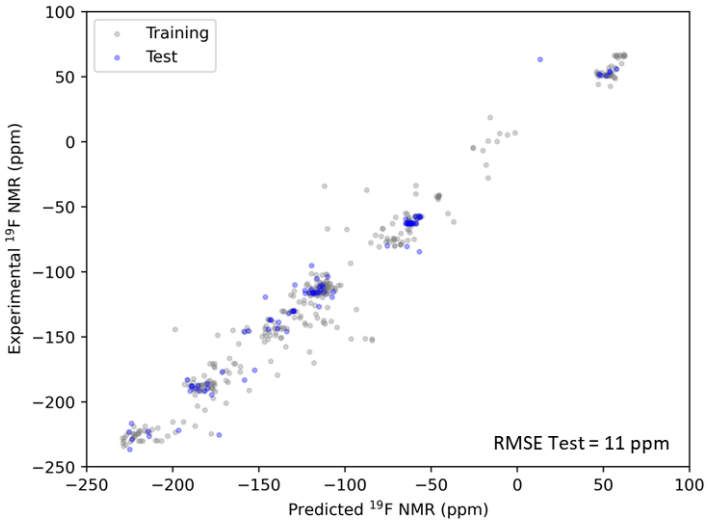


Figure 64 Optimised ^{19}F predictions on monofluorinated dataset 459:91 Train:Test Split RMSETest = 11 ppm

Development of a Multi-Fluorinated Model

While good performance was obtained with the monofluorinated model, one limitation is the explicit exclusion of multifluorinated molecules, which significantly reduces the chemical space covered by the model. Polyfluorinated aromatic and alkyl chains hold a privileged place

in ^{19}F NMR, neighbouring fluorine's can shield an adjacent ^{19}F nucleus resulting in shifts that appear further downfield than expected. Therefore, multifluorinated molecules needed to be included in our dataset.

We first wondered whether knowledge could be transferred directly to multifluorinated molecules, where the dataset included molecules with multiple environments and the model attempting to predict either of the two environments. However, this transfer task failed with the model unable to discern between different environments, with the model either predicting the average of the training data or only predicting one environment resulting in errors of more than 50 ppm.

The transformer tokenises each atom or bond in a SMILES string, so that the model learns the relationship between different tokens (atoms) and NMR shift, however, when multiple F tokens are present this function breaks down as there are multiple NMR shifts to be predicted. We decided to use masking techniques to ensure there is a 1-to-1 mapping between inputs and outputs.

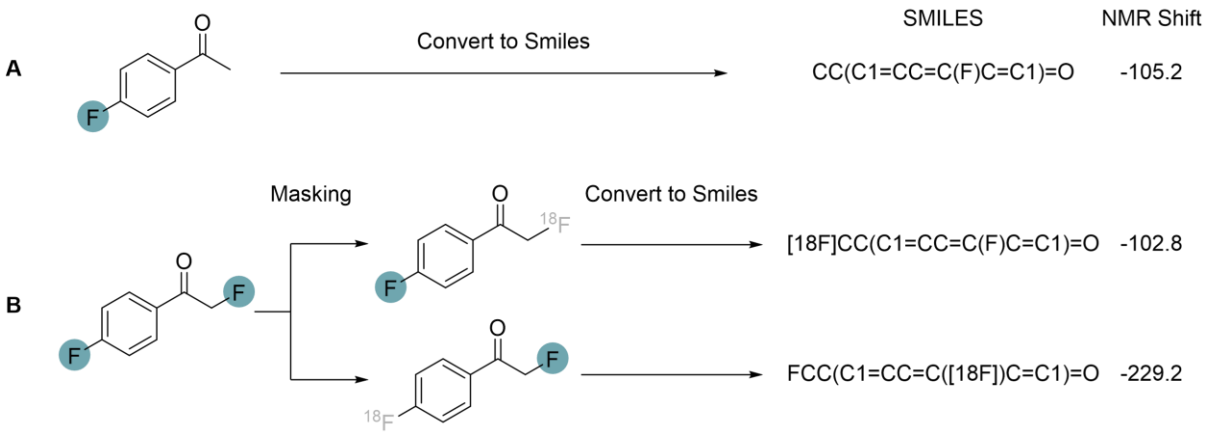


Figure 65 SMILES strings generated from 2D Structures for A) mono-fluorinated molecules and B) multi-fluorinated. Here the non-target environments are masked before generating SMILES strings

The masking group chosen was ^{18}F as it was already a token in YieldBERT. For structures where only one fluorine is present, there is therefore no need to change the structure, so the

SMILES string is left unchanged (Figure 65A). For structures with multiple fluorine environments, the non-target environments are masked as ^{18}F, and this masked structure is then converted into its SMILES string (Figure 65B).

Prediction on Multifluorinated Dataset

With a masked training dataset in hand, we trained the NMRBERT for ^{19}F NMR prediction using a five-fold split and the hyperparameters reported by Schwaller *et al.*[152] Pleasingly with a small dataset (565 training data points,113 test data points) a small RMSE is obtained, RMSE = 14 ppm corresponding to a 5% error over the range of data points (Figure 66). However, even after hyperparameter optimisation and the addition of a further 64 training data points alongside the expansion of the test data to 314 data points the model performed slightly worse at 17 ppm (RMSE). We therefore wanted to understand if the method of dataset preparation was having a significant effect on the model's performance.

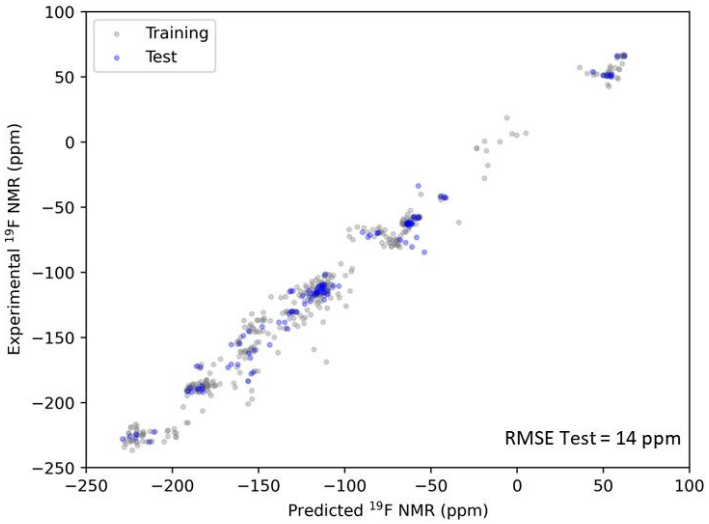


Figure 66 Initial ^{19}F NMR predictions for multifluorinated dataset 565:113 Train:Test Split RMSETest = 14 ppm

To understand the impact of training data splitting, we investigated the impact of data shuffling, data splitting, and cross-validation on the model's performance. The three methods behave significantly differently, with RMSEs of 17 ppm, 9.4 ppm and 8.5 ppm for data-splitting, data-shuffling and cross-validation respectively.

Here we employed Tanimoto similarity scores to understand chemical similarity in reaction datasets and the effects on the accuracy and generalisability of the model by separating the training and test sets by Tanimoto similarity. Training a model where the similarity between training and test data was less than 0.4 would show if the model was learning chemical principles which can be used to predict new compounds. Given the handcrafted nature of the dataset only a 93:7 training: test split could be achieved.

When data shuffling is used the RMSE increased from 9.4 to 14.7 ppm (Figure 67A vs Figure 67B), while for cross-validation, a smaller differences were observed (8.5 ppm to 13 ppm) showing that the model is less prone to overfitting during training. Given the improved performance of the cross-validation model, we choose this as our optimal data splitting method.

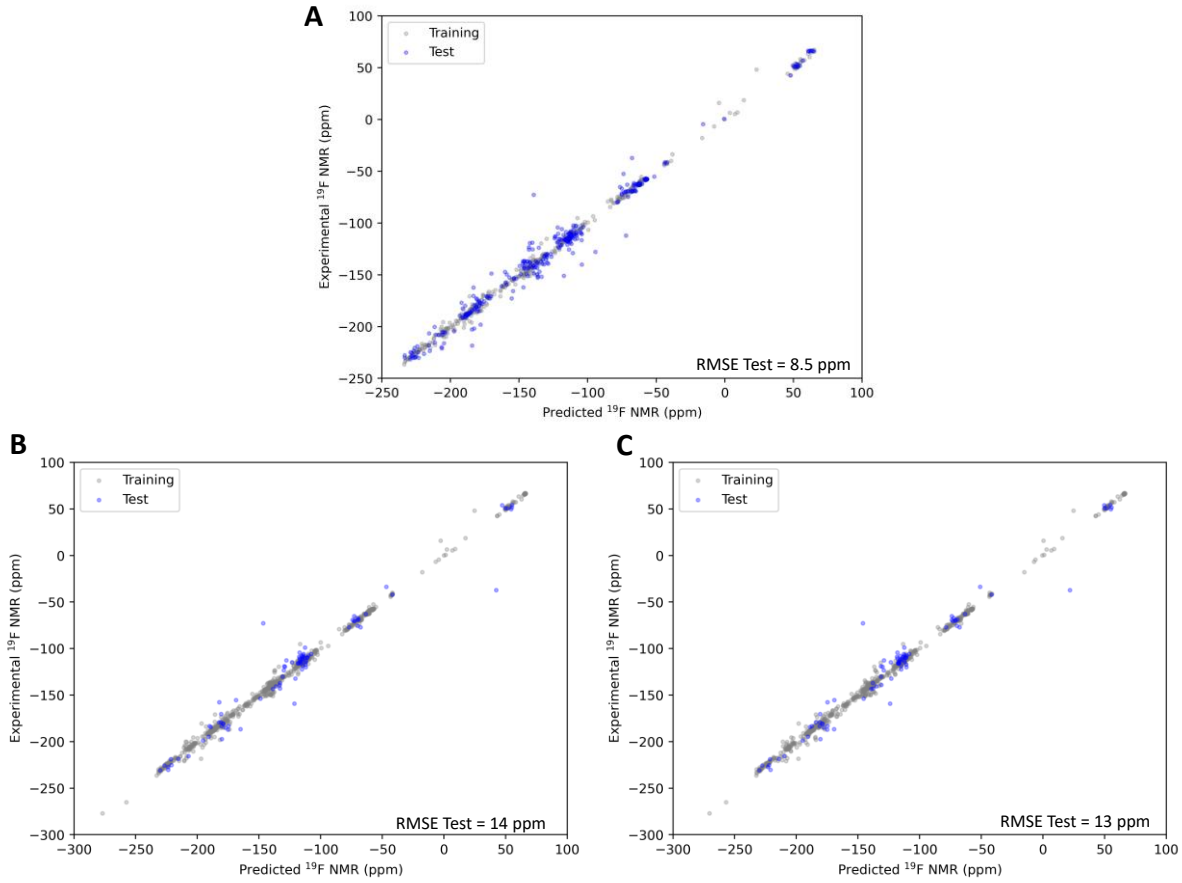


Figure 67 A) Final ^{19}F NMR predictions for multifluorinated dataset 629:314Train:Test Split RMSETest = 8.5 ppm B) Tanimoto 10-fold shuffle RMSETest = 14 ppm C) Tanimoto 10-fold cross-validation RMSETest = 13 ppm. For the Tanimoto splits, the train: test split was 845:89, carried out using Tanimoto criteria > 0.4.

Augmentation Effects

One important tuneable parameter in the use of SMILES-based transformers is augmentation. This process generates multiple valid but different SMILES strings for the same molecule effectively increasing the training dataset without increasing the amount of source data. However, it can lead to overfitting as there is a finite number of SMILES per molecule. For example, if there are only six valid SMILES strings for a given molecule, but the augmentation level is set to 15, then multiple copies of the same SMILES string will be generated during augmentation and added to the training data, resulting in multiple copies of the same string.

We investigated if duplication affected the model accuracy by comparing models where duplicates were allowed, to one where duplicates were removed. Increasing augmentation led to a decrease in the error of the training set (Figure 68A), but the error in validation increases after 20 rounds of augmentations. When duplicates are removed (Figure 68B) the error on the training, test and validation sets continues to decrease plateauing by 20 augmentations.

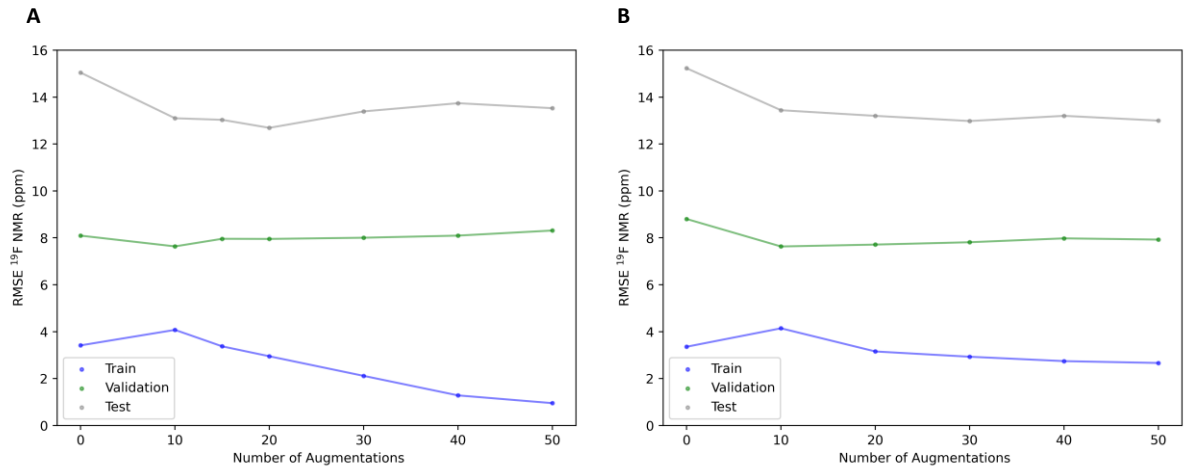


Figure 68 Comparison of a number of augmentations vs. model error A) without removing duplicates and B) removing duplicates.

The fact that the training error continues to decrease while the validation error stays constant with a larger number of augmentations in Figure 68A suggests that the model is overfitting,

and therefore it is not sensible to go to larger number of augmentation levels. For this reason, we used 15 augmentations only in line with previous work.[152]

Breaking down the Model Error

While the overall RMSE for the model was 8.5 ppm (Figure 67A), the error can be further broken down by the environment of the nuclei of interest. Groups that cover a large ppm range (e.g. R_2CHF, RMSE = 9.7 ppm, Range = 110 ppm) or have a small number of data points (BF_4, RMSE = 15.9 ppm, 4 training data points) are the main source of error. Primary alkyl and aromatic trifluoromethylated arenes perform well, with an error of 6.5 ppm and 1.75 ppm respectively.

Our ML method performance is similar to the best DFT method reported to date. Overall NMRBERT achieves an MAE of 5.1 ppm comparing this to the work of Bagno,[57] which obtained an MAE of 6.2 ppm using DFT methods. Tantillo[58] focused on predicting aromatic fluorine and reported an MAE of 2.1 ppm; by comparison, our model has an MAE of 5.6 ppm for aromatic fluorines, while higher this could be due to the model aiming to generalise over multiple environments rather than just being specific to one type of environment. Recently Whiting *et al.* used a computationally expensive DFT method to predict a small number of fluorine NMRs with varying solvents, achieving an RMSE of 3.75 ppm[59] after linear fitting, while far more accurate than our model this method was only tested on small molecules, and therefore could become more expensive to calculate for larger molecules. In comparison, our model takes 0.65 seconds per molecule on one CPU core.

5.2.2 Transfer to ^{1}H and ^{13}C NMR

To explore the transferability of this architecture to ^{1}H and ^{13}C NMR predictions, we used the experimental dataset NMRShiftDB and the DFT-generated dataset reported by Paton *et al.* (Paton set).[245] Using the NMRShiftDB set, with minimal modification of our masking workflow we obtained an RMSE (^{13}C) = 6.8 ppm and RMSE (^{1}H) = 0.8 ppm (Figure 69A). With a large increase in the number of environments per molecule, we decided that instead of masking all other environments, identifying the target by labelling it either as ^{1}H or ^{13}C would be preferential. While the ^{1}H NMR error is higher than the best-performing DFT methods (0.2 ppm) it is within the margin of error for ^{13}C NMR prediction (5 ppm).

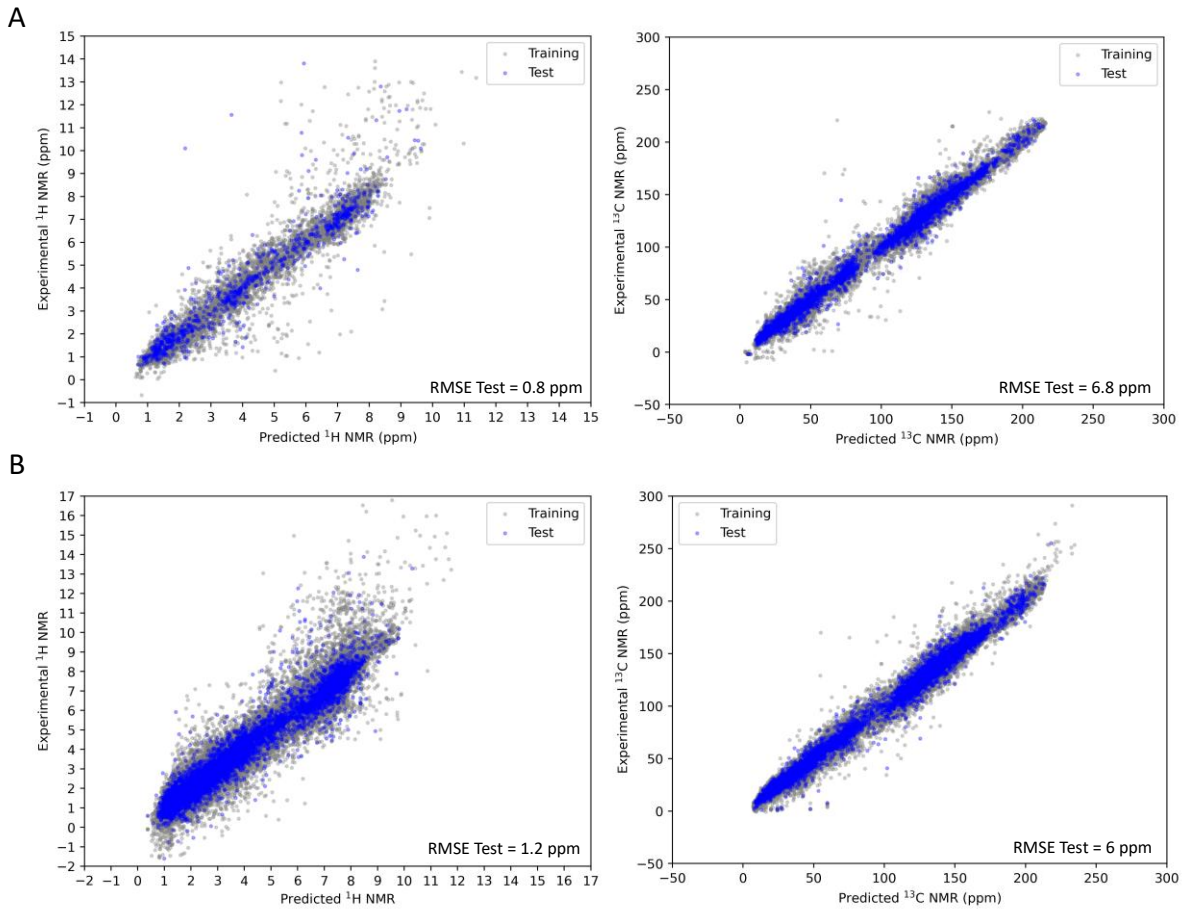


Figure 69 Prediction of ^{1}H and ^{13}C NMR on A) NMRShiftDB and B) Paton Dataset. Training was carried out using 10-fold cross-validation and a 2:1 Train:Test split.

To understand the impact of solvent inclusion we included data points from multiple solvents and appended the NMR solvent to the SMILES string of the compound, yet this extra information did not affect the error and therefore was not further explored.

To understand if the observed ^{1}H error was caused by misassigned data, a noted error within NMRShiftDB.[87, 245, 306] We used the Paton dataset, which was generated from computational calculations, and should be free from misassignment. After training on the Paton dataset, the ^{1}H prediction however performed worse with the CV error of 1.2 ppm, while the ^{13}C NMR predictions were unchanged at 6 ppm. Given the nature of this curated dataset, we, therefore, conclude that data misassignment is not the reason for the observed error (Figure 69B).

To further understand the nature of the large error obtained when training on the Paton dataset we studied the underlying ^{1}H data. We noticed that in some cases multiple copies of the same SMILES string were assigned to different NMR shifts. Further investigation noted that these strings originated from diastereotopic protons.

An example can be seen with the molecule Oculatol, where the highlighted protons on the A ring are diastereotopic (Figure 70A). Both protons have different ppm shifts, a separation of 0.9 ppm, by changing one of the diastereotopic protons to either ^{1}H or ^{2}H should result in the generation of a new chiral centre. However, when these structures are converted to canonical SMILES strings the differentiation between the two different environments is lost. The dataset now contains the same SMILES string, but the experimental NMR is different for each string. This means the model is unable to accurately predict both values, leading to larger errors.

To circumvent this issue, we substitute the diastereotopic proton for another atom, replacing the ^{1}H with an Lv atom as it is not contained in the training data. While chemically absurd, this generates unique SMILES strings for diastereotopic and chemically equivalent but magnetically inequivalent nuclei (Figure 70B). Retraining the ^{1}H models using this modified training data resulted in an error of 0.51 ppm and 0.59 ppm for the Paton and NMRShiftDB

datasets. To ensure that any error was not caused by the addition of these heavy atoms we changed the Lv back to ^{1}H in the SMILES representation, this has no impact when the molecule is read by the transformer, and the error for this model was unchanged.

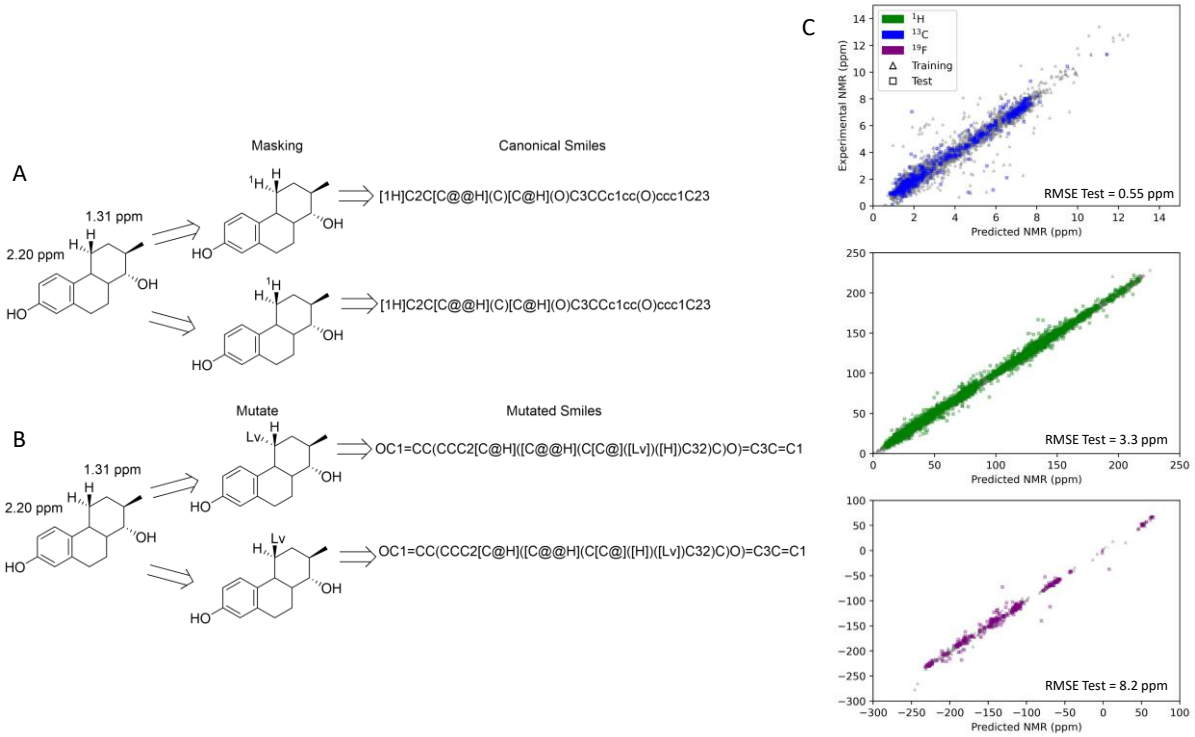


Figure 70 A) Limitation of SMILES strings to generate chiral centres for isotopes as shown on Octanol B) Mutating the target atom generates chiral centres and unique smiles strings C) Comparing predicted NMR against experimental NMR trained to predict ^{1}H,^{13}C and ^{19}F NMR in one model, test sets are identical to those in individual models.

5.2.3 Multi-nuclei prediction

With separate models that could predict the ^{1}H, ^{13}C and ^{19}F NMR with an error of 5% over the size of the spectrum we trained a single model on all three environments. Given the different ppm ranges for each nucleus, we hypothesised that the errors in ^{13}C and ^{19}F (6 ppm and 9 ppm respectively) would result in the model being unable to predict ^{1}H spectra entirely. However, while the average error for this model is 6 ppm, when we break down the errors per nuclei, we found the errors to be smaller than their respective trained models.

The combined NMRBERT model has errors (RMSE) of 0.55, 3.3, and 8.2 ppm for ^{1}H, ^{13}C and ^{19}F spectra respectively, compared to 0.59, 6.8, and 8.5 ppm for the individually trained models. The improvement in each nuclei subset suggests that the model is learning the subtle changes affecting NMR properties and that this knowledge can be shared to improve the error of all predictions (Figure 70C). Comparing the error of our model to those published our model performs similarly to others based on an experimental dataset. Other models tested on the NMRShiftDB achieve accuracies (MAE) between 0.2 and 0.3 ppm for ^{1}H and 1.3 – 1.5 ppm for ^{13}C predictions.

Kuhn et al[306] used a GNN and achieved similar accuracies, 0.22 ppm and 1.35 ppm (MAE) compared to NMRBERT, 0.29 and 2.0 for ^{1}H and ^{13}C respectively, Kang et al[206] achieved similar levels as well at 0.22 ppm and 1.36 ppm. Training on QM-derived data can achieve accuracies below 0.15 ppm and 1.3 ppm for ^{1}H and ^{13}C respectively,[84, 185, 245] however as noted there is then an added error from the QM method which must be considered, which could lead to errors similar to those achieved by our method.

5.2.4 Assessing the effect of model complexity on ^{19}F NMR shift prediction

During training, the NMRBERT model reached an RMSE of 0.59, 6.8, and 8.5 ppm for ^{1}H, ^{13}C and ^{19}F. The size of the error over the range of the predicted spectrum however was similar, 4.1 %, 2.4 % and 2.7 % error for H, C and F respectively. We, therefore, wished to investigate whether a more complex model, with a larger number of attention heads and hidden layers, would change the accuracy of the model. We, therefore, enlarged the YieldBERT model to have the same hyperparameters as that of the original BERT Model,[307] This larger model will be referred to as LargeNMRBERT.

On the prediction of ^{19}F NMR, the accuracies for NMRBERT and LargeNMRBERT are identical on the training data, 3.0 ppm. However, on the test set the accuracies are 8.5 ppm vs. 10.5 ppm, respectively. This slight increase in test set error implies that increasing the size of

the NMRBERT model has no impact on the ability to predict NMR signals, and therefore the model's size is not a limiting factor in its accuracy. This further justifies the use of a smaller model, which takes less time to train and evaluate alongside having the same level of error as a larger model. This further suggests that the source of the error is not coming from the size of the model, but either from the molecular representation or the dataset itself.

5.2.5 Explaining the NMRBERT predictions

To understand and interpret the models and identify potential sources of error we used two different techniques; fingerprint comparisons, and integrated gradients (IGs), both introduced in Chapter 2. Fingerprint comparisons allow us to understand how similar the model sees two molecules in chemical space. IGs allow for an attribution between each token in a smiles string to the overall prediction, giving insight into the importance of each token and its contribution to the overall prediction.

Can fingerprint similarity explain NMRBERT predictions?

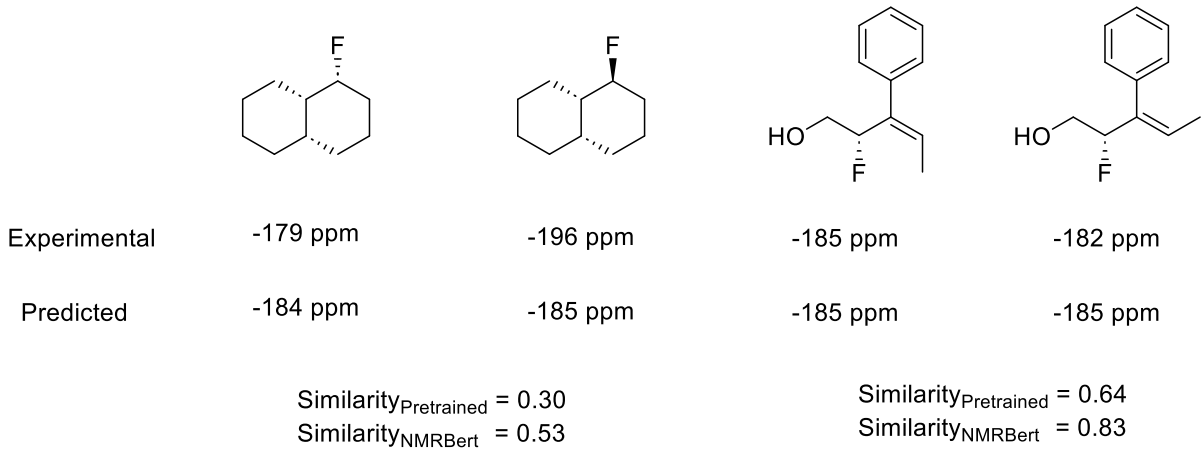
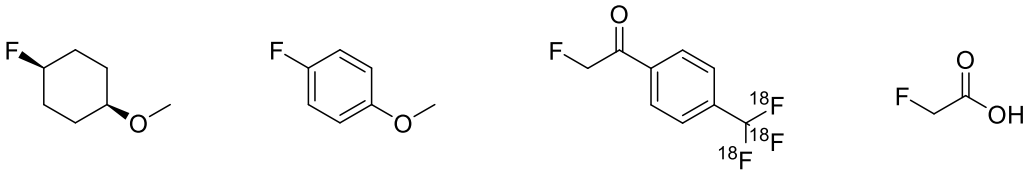


Figure 71 Comparing the BERT fingerprint similarity score between diastereoisomers and cis-trans isomerism in the NMRBert Model

To understand the relationship between fingerprint similarity and predicted chemical shift we selected a set of examples that covered structural isomerism (*cis*/*trans*), diastereomers, oxidation states and similar NMR shifts.

For *cis* and *trans* isomers and diastereomers (Figure 71), where the experimental difference in ^{19}F NMR signals are 3 ppm and 18 ppm with our model predicting a difference of < 1 ppm respectively, therefore unable to discern between these structures. However, when we look at the similarity between fingerprints, we see that during training the similarity scores increase for both diastereomers and cis-trans isomers by similar amounts, 0.23 and 0.19 respectively showing that these structures move closer together during training.

Analysis of fingerprint similarity indicates that while our model put these pairs of structures far apart in chemical space (0.53 vs 0.83) the model is unable to resolve that difference by predicting different NMR shifts (Figure 71).



Experimental	-179 ppm	-124 ppm	-230 ppm	-230 ppm
Predicted	-180 ppm	-121 ppm	-229 ppm	-230 ppm

$Similarity_{Pretrained}$ = 0.15
$Similarity_{NMRBert}$ = 0.086

$Similarity_{Pretrained}$ = 0.11
$Similarity_{NMRBert}$ = 0.16

Figure 72 Similarly scores between different oxidation states and structures with similar environments and NMR Shifts

A different trend is seen in when we compare two further examples; a change in oxidation state, aromatic to aliphatic, and structures which have very similar NMR shifts but different structures (Figure 72). Overall, the model can distinguish between the different structures as the difference in experimental NMR is greater than the error in our model.

When we calculate the similarity score between para-fluoroanisole and (1*s*,4*s*)-1-fluoro-4-methoxycyclohexane, we observed that it decreased during training with the similarity score being very low (0.086), therefore moving these different structures further apart in space as we

would expect. For the two compounds with similar fluorine environments but different structure sizes the similarity scores are higher, but still only 0.16 after training. The model however can take these two different structures and their fingerprints to predict very similar NMR shifts.

Overall, the similarity scores alone are not an indicator of the NMR shifts that the model predicts. Furthermore, while similarity scores do change during training this is not converging to the same fingerprint for molecules with the same fluorine environment. This suggests that the NMRBERT model is not just fitting structures to an NMR value, but that it retains knowledge about chemical space and the differences between chemical species, and therefore no relationship between predicted chemical similarity and predicted chemical shift.

Can integrated gradients give chemically rational explanations for NMRBERT predictions?

IGs break down the model's prediction into attributes for each token (atoms or bonds), allowing for the identification of which parts of a molecule are important for the prediction of NMR shifts. As mentioned in Chapter 2, the use of IGs requires a baseline from which to calculate the path integral. In this case, the baseline was uniformly made up of '.' tokens, which are used to represent the separation between compounds in SMILES strings and therefore are reasonable examples of empty molecules.

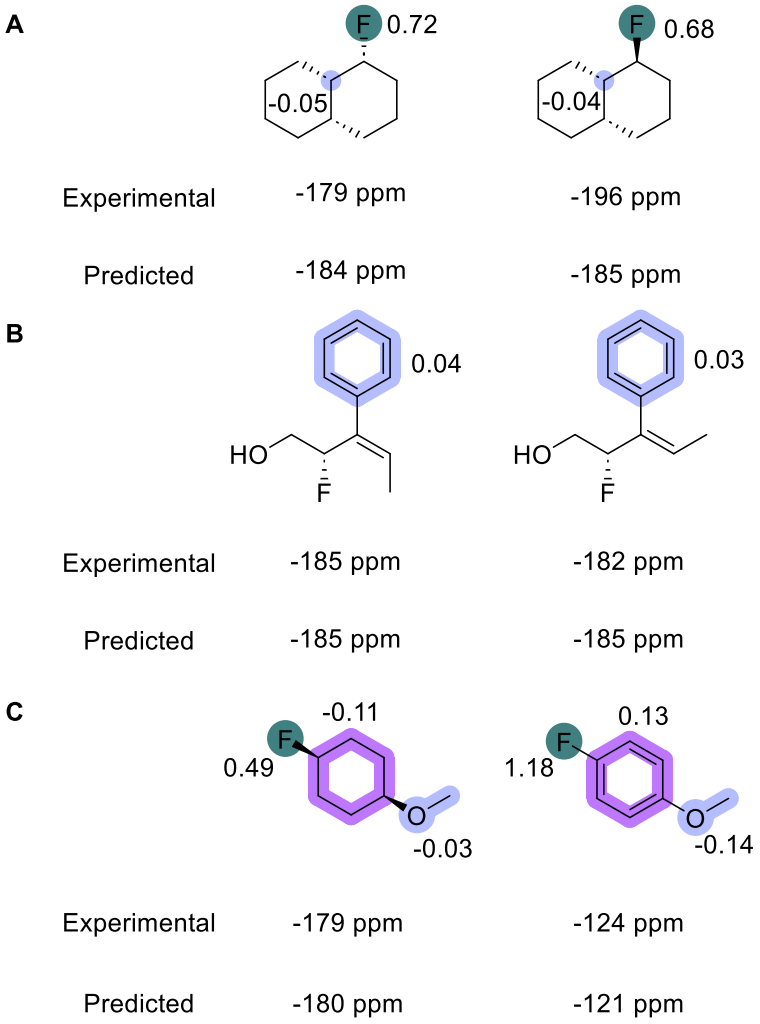


Figure 73 IGs of highlighted fragments are shown along with the experimental and predicted NMR is shown below each compound. A) Difference between Diastereomers B) Cis-Trans Isomerism C) Oxidation difference between aromatic and aliphatic groups

IGs for the diastereomers are similar with the only differences being on the ring junctions, and the fluorine and the β-carbon (Figure 73A) with less than 0.04 between any of the tokens. The small difference in IGs and the larger difference in fingerprint similarity implies that the model is unable to capture the subtle differences in chemical space from these SMILES strings and convert this into a difference in the ^{19}F NMR prediction. This could explain the error in our model, for this set of diastereomers.

The case of the *cis* and *trans* isomers (Figure 73B) is more complex, the SMILES for the *cis* or *trans* isomerism are encoded in the aromatic ring tokens, therefore, those tokens must be

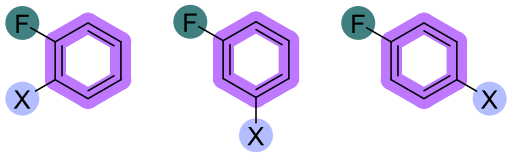
compared to understand what the model is learning. The difference in IG's of the aromatic ring is only 0.01, in line with a small experimental difference of 3 ppm between the two structures, and our model predicts a gap of < 1 ppm.

Finally, when comparing aliphatic and aromatic backbones, where the experimental difference is over 50 ppm, the IG values show significant differences (> 0.2) in both magnitude and sign, especially at the ring and F tokens (Figure 73C). This is in line with the fact that aromatic centres are more deshielded than aliphatic, while the fluorine atom mesomerically donates into the aromatic ring, reducing the electron density of the fluorine atom. Interestingly, there is a slight decrease in the importance of the OMe group, probably due to the mesomeric effects of the OMe donating into the ring and thereby shielding the F centre. In contrast, in the aliphatic case, only inductive effects are present (Figure 73C).

Overall, the IG results from these case studies are in line with the expected chemical reasoning, they also provide insight into the reason that the model is unable to discern between diastereomers, i.e. NMRBERT can separate the two in chemical space, but not convert that difference into a change in NMR. This results in diastereomers having a large error when compared to achiral structures. With that in mind, we further wished to extend our understanding of what NMRBERT was learning, by looking at electron donating and withdrawing effects and their effect on ^{19}F NMR.

Do integrated gradients identify electron donating or withdrawing effects

Finally, we apply IGs to a range of substituted fluorobenzenes to identify how the contributions from different structural motifs vary with the *ortho*, *meta* or *para* substitution patterns (Figure 74). IGs were calculated and broken down into three categories: those from the F token, those from the ring tokens and those from the X (substituent) group.



X = OH, Br, Me, Cl, NH_2, OMe, I,
SH, CHO, COOH, COMe, NO_2

Figure 74 Visualising the three types of token groups in substituted fluorobenznes. F tokens are shown in green, Ring tokens in purple and R groups in blue

For the X tokens two distinct groups appear independent of the position of the X group (Figure 75A), those that increase the NMR prediction (Cl, Br, I and SH) and those that decrease the prediction (OH, Me, NH_2, OMe, CHO, COOH, COMe, NO_2). This separation corresponds to the halogen effect, where a balance between mesomeric and inductive effects results in a net shielding effect across the series. Notably, the model also includes the SH group, for which experimental values follow a similar pattern to the halogen series and will be discussed further in the following section. All other tokens are negative and therefore contribute to shielding compared to those in the halogen class.

When analysing the IGs for the aromatic rings (Figure 75B) in the molecules containing SH, I and CHO, these values decrease as the substituent moves from the *ortho* to the *para* position. This reflects the increase in the shielding of the F atom from the aromatic rings. For Cl, Br, NH_2, OH and Me, the IGs increase from the *ortho* position to the *meta* position before decreasing back to the *para* position. This is characteristic of *ortho/para*-directing effects. Through mesomeric effects, the *ortho* and *para* positions are shielded, with the former being more shielded than the corresponding *para* position. Finally, the IGs for COOH, COMe and NO_2 also increase, suggesting a decrease in shielding during the series.

Finally, for the F tokens (Figure 75C) we see similar effects to those of the ring tokens: the IGs of COOH, CHO, Cl and SH all increase over the series suggesting a deshielding effect as the F moves away from the substitutions, implying an inductive-like relationship. For OMe, Me,

OH and NH_2 all behave as expected, the F token is shielded at the *ortho* and *para* positions, with the *ortho* being the most shielded. The more noticeable group, however, contains NO_2, COMe, Br and I. The IGs decrease from the *ortho* to the *meta* position before increasing back to the *para* position, with larger IGs than at the *ortho* position. This corresponds to a deshielding of the *ortho* and the *para* positions with the *para* being more deshielded. This would suggest a meta-directing mesomeric effect, however only half the group is *meta-directing*. This shows that the model is interpreting the effects on the F token within this group as similar, which is not what experimentally is observed. Furthermore, the four largest IG, contributing >1.1 to the NMR shift, are from *meta-directing* groups. The bottom four are also Cl, Br, I and SH, all under 0.9, the lowest at 0.5, therefore further showing the separation between different classes of compounds.

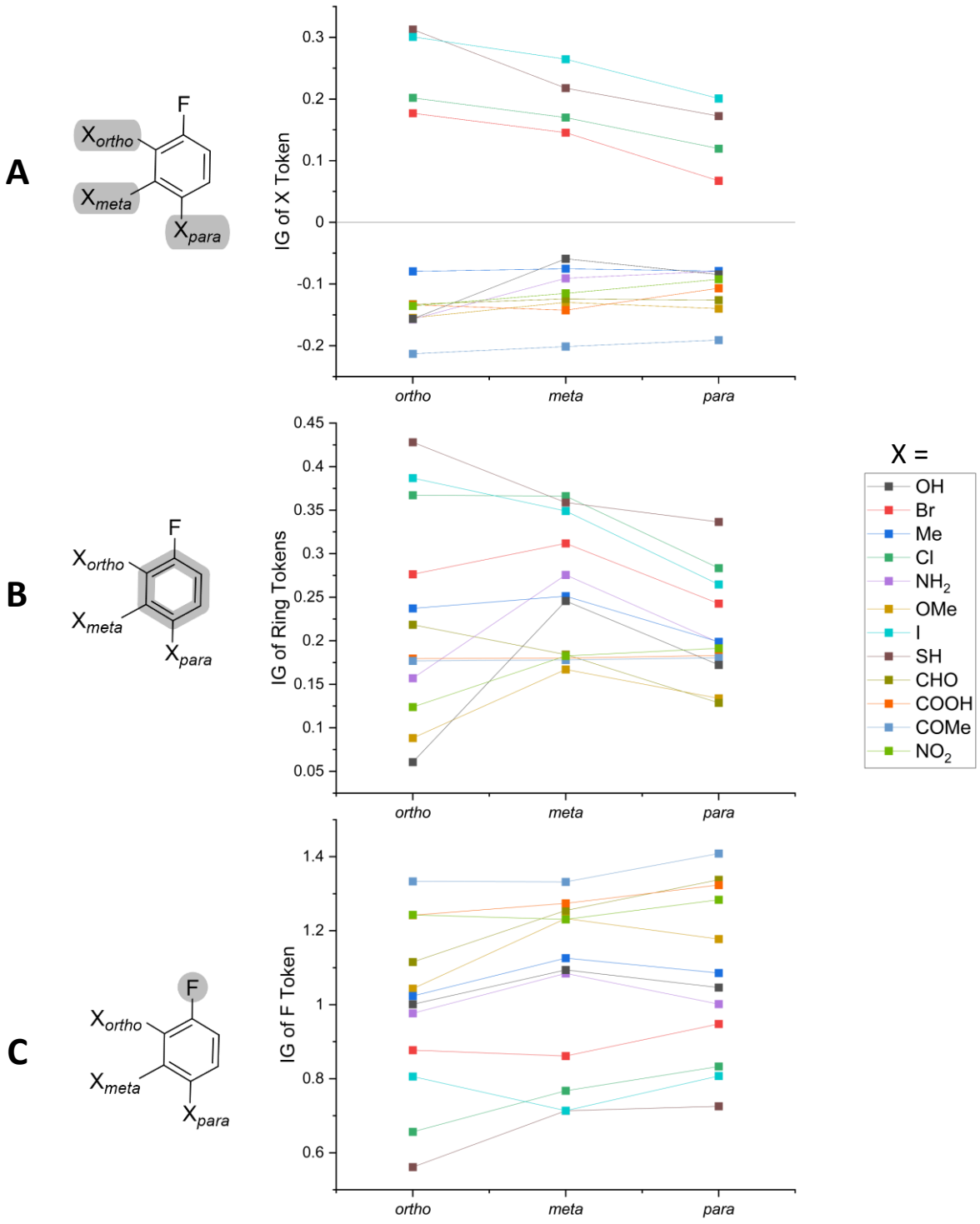


Figure 75 Outputs of the IG method for all mono-substituted compounds for the token are shown for *ortho*, *meta* and *para*-substituted fluorobenzene. A)IGs of the X tokens B) IGs of the Ring tokens C) IGs of the F tokens. X = OH, Br ,Me ,Cl , NH_2, OMe, I, SH, CHO, COOH, COMe, NO_2

Do BERT models learn chemistry or just fit to data

From the previous analysis, the following trends emerge. First, IGs decrease from the *ortho* to the *meta* position and then increase at the *para* position, similar to mesomeric electron-donating groups. For the second trend, IGs increase from *ortho* to *meta*, and then fall back down at the *para* position. This pattern is characteristic of mesomeric electron-withdrawing groups. In the third trend, IGs decrease from *ortho* to *para*, this shows the halogen effect. Through inductive effects the *ortho* positions are deshielded, this is a distance-dependent result and so the *meta* and *para* positions become more shielded. The final trend is the most surprising, IGs increase from *ortho* to *para*, which is inconsistent with either inductive or mesomeric effect. When we analysed the original data however for the ^{19}F NMR shifts these trends are represented in the data, both NO_2 and CHO exhibit this inverse relationship (Figure 76B) meaning the model is identifying this pattern from the data. The classification of SH as a halogen series, as seen in the previous section, is also backed up by data (Figure 76A) which shows the similar pattern even when *a priori* that would not have been predicted.

This subset of mono-substituted fluorobenzenes show that the model is learning patterns from the data and carefully balancing the competing factors to give an overall prediction for the NMR shift. Instead of just fitting all tokens to the prediction, this fine balancing between the different groups allows for a more nuanced approach to ^{19}F NMR predictions. IGs here are therefore key to understanding which parts of molecules are being identified with the most important being the fluorine itself, followed by the ring and X tokens, whose relative importance varies depending on the X group.

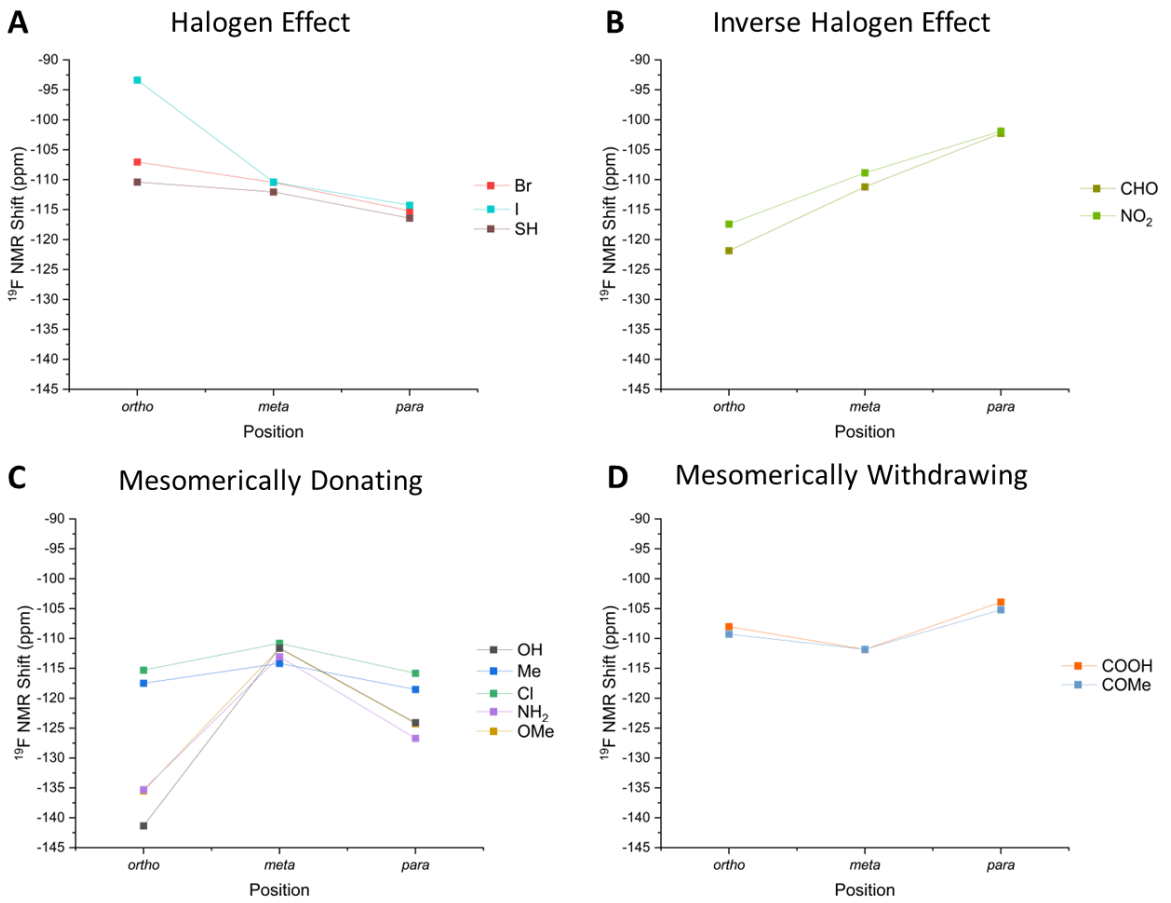


Figure 76 Experimental ^{19}F NMR trends as a function of position A) The Halogen effect was shielding increase across the trend. B) The inverse halogen effect where shielding decreases across the trend. C) Shows mesomeric donating groups D) Shoes mesomerically withdrawing groups

Variation of the baseline

Rather than using just a series of "." tokens as a baseline another option is to choose an actual molecule as a baseline, here we chose the unsubstituted fluorobenzene, meaning only the effects of the R group on the shift are studied. This method however limits the examples to groups which are represented as a single SMILES token such as CH_3, NH_2, and OH as a one-to-one replacement of the R group for an H must be carried out.

As expected, the only non-zero IGs obtained are those for tokens representing the R group. Furthermore, the change from *ortho, meta, and para* for all groups directly mirrors the changes in NMR shift for the separate environments (Figure 77A).

Plotting the IGs against the raw predicted NMR shifts (before scaling), shows a strong stratification into *ortho*, *meta* and *para*-substituted fluorobenzenes. The ordering of tokens within each class is the same as in Figure 75 but shows that within each subset the model learns the relative effects of placing a variety of functional groups at that position (Figure 77A). This stratification however appears to be an artefact due to the choice of baseline.

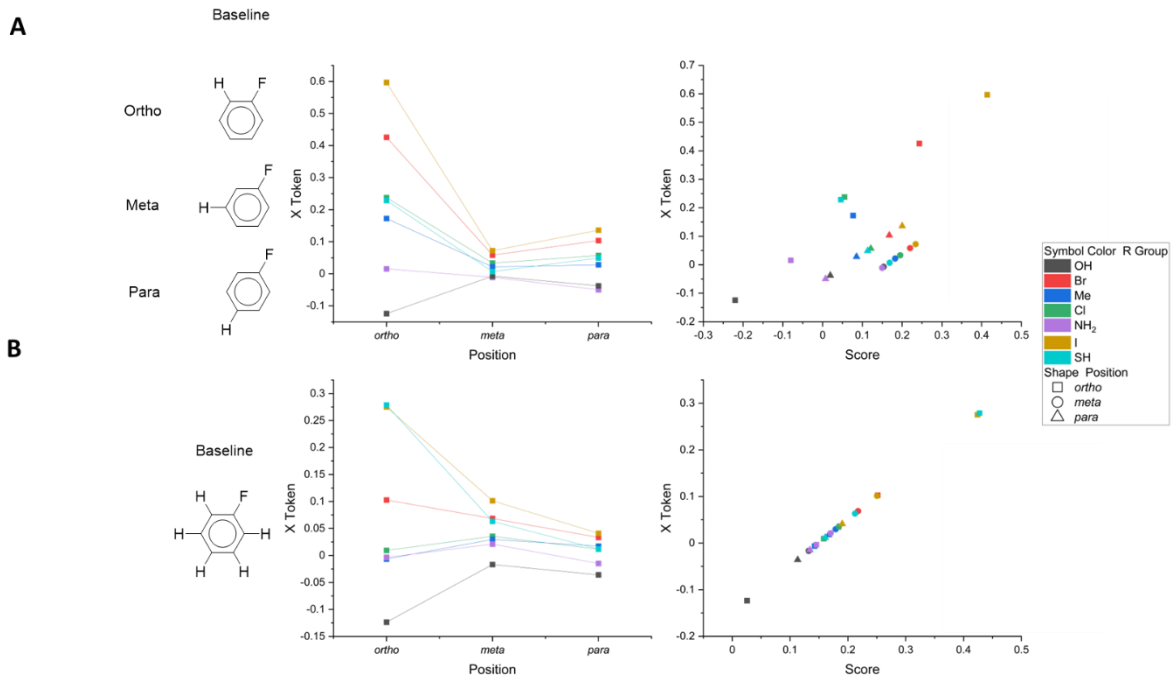


Figure 77 Comparing the effects of changing baseline on the IG Methods. A) The baseline is chosen according to the substituents. B) Universal Baseline is used. The score is the raw output of the NMRBERT Model before transforming back to NMR Shifts.

Using fluorobenzene as a baseline requires the use of explicit hydrogens; however, the model had not been trained with explicit hydrogens in the SMILES strings and therefore the accuracy of the interpretations could not be relied upon. To circumvent this issue, a model was retrained with explicit hydrogens for which a similar accuracy was obtained (test RMSE = 10 ppm vs 8.5 ppm without Hs) (Figure 77B). The IG results can be separated into two classes of compounds, those that show inductive behaviour, decreasing from the *ortho* to *para* positions (S, I, Br) and those that increase from *ortho* to *meta* and decrease at the *para* position (OH,

NH_2, Cl and C), showing a mesomeric electron donating effects. Finally, the relative IGs value between R groups correlates with the experimental order. The only outlier is the NH_2 R group, which seems to arise from the significant error obtained for this group during the retraining of NMRBERT with explicit hydrogens This would explain why the IGs for NH_2 substituted benzenes are more positive than would be anticipated when compared to OH. Attempts to increase the model's accuracy using augmentation did not improve the model predictions, showing that adding explicit hydrogens does not improve the model's predictions.

5.2.6 Predicting products of late-stage fluorination reactions using NMRBERT

We envisaged that a model that accurately predicts ^{19}F NMR signals could be used to help characterise products in reactions where two regio-isomers are possible. Examples include late-stage fluorination techniques reported by Groves[24, 308] and Ritter.[309, 310] We, therefore, wanted to explore if NMRBERT could be used to identify the products of these reactions purely by ^{19}F NMR. From the four papers, we assembled three separate test sets for, aromatic fluorinations (13 examples), aromatic diflouro and trifluoro methylations (10 examples), and for aliphatic fluorination (8 examples). To validate our predictions a confidence score is used, analogous to that developed by Goodman for the DP4 program (See 5.4.11 Calculating confidence values for regioselective reactions), to assess our prediction (Figure 78).

In general, the NMRBERT can easily differentiate between different substitution patterns on fluorinated aromatic rings with the correct product having the greater confidence in 8/13 cases. Figure 78 **A** and **B** show the predictions and confidence scores for two typical regioselective problems. For example, in **C,** our original predictions were significantly off; however, a re-examining of the literature uncovered that the original compound was isolated as the TFA salt, predicting the TFA salt results in much more accurate predictions in line with the experimentally reported value. Interestingly there is only one example of a protonated nitrogen in the training data and not a protonated pyridine. This information could only have been learnt

during the pre-training of the model, showing that pre-training can have important effects on the prediction in new chemical space. In contrast, the model shows low confidence in the predictive power of the model for difluoro and trifluoro-methylated aromatics, with only 2/10 products correctly predicted (Figure 78 B).

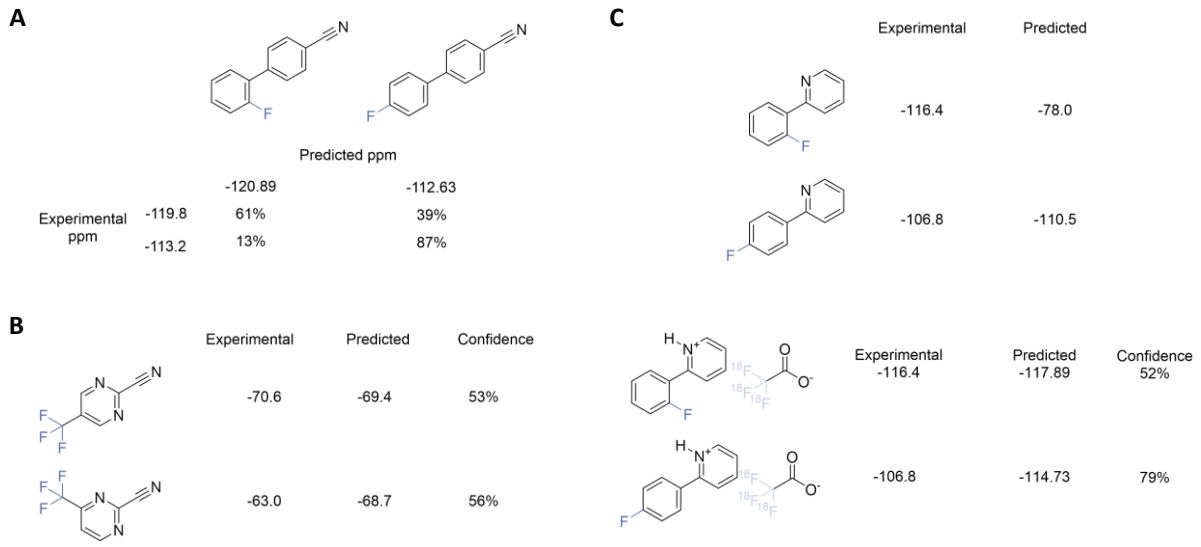


Figure 78 Three examples from our external test set with predictions along with confidence A and B show examples of the model both succeeding and failing to identify the difference between two regioisomers. C gives the model picks up on protonation states of the molecule and has a significant impact on the prediction.

Using regioselective fluorination data from Groves, 8 aliphatic fluorines were chosen, including regio- and diastereomers. Overall, the model is unable to resolve the difference between diastereomers even though the difference between the reported ^{19}F NMR shift is large. However, comparing the difference between regio-isomers the model can resolve between all 4 regio-isomers successfully. This does make sense however as in most of the examples the difference in ppm is less than the error in the model.

5.3 Conclusions and Future Work

Herein we presented a transformer-based ML model to predict NMR signals for ^{19}F. The model can make predictions with sub 10 ppm error on a range of fluorine environments. Compared to other works, which use graph neural networks, this is one of the first models that use text-based SMILES strings only to predict this molecule property without reverting to atomic coordinates. We also introduced masking of functional groups we do not wish to predict properties for and show that this allows for a one-to-one prediction for SMILES strings to NMR shifts.

The applicability of the framework was also demonstrated in ^{13}C NMR predictions, where an accuracy of 6.8 ppm was achieved. However, extension to ^{1}H NMR predictions proved more challenging, with errors of 0.59 ppm. Yet when we combined our data into one unified model that could predict ^{1}H, ^{13}C and ^{19}F NMR at the same time the errors decreased to 0.55, 3.3 and 8.2 ppm respectively. This decrease in the errors shows that when the amount and diversity of the data increased, the model was able to learn from each of the different environments to improve the prediction for the nucleus in question

In addition, the use of Integrated Gradients enabled us to interpret the predictions from our model. The model was able to identify chemical principles such as the effects of electron-donating and withdrawing groups on NMR signals. We believe this work will enable others to predict a range of chemical properties not only NMR signals using transformers as a base ML method.

As noted in the discussion on applications for regioselective predictions, our model currently is unable to discern between diastereomers. This could be an issue firstly in data collection and curation. In most cases for enantioselective or diastereoselective reactions only the major or the desired product is experimentally determined and the minor or the opposite diastereomer is not reported. We therefore are significantly at a disadvantage in terms of experimental data. This is an area of potential further work where supplemental DFT calculations could be used

to generate NMR shifts for a range of diastereomers so that our database will become more balanced.

The second problem is about the representation of 3D chemical space using a 1D string of characters. Given the effect that 3D conformations have on the NMR shifts of a molecule, it follows that SMILES strings may not be able to fully encompass the effect that local through-space effects have on NMR signals. This also could explain why our error in ^{1}H NMR is so high, ^{1}H NMR is much more influenced by the molecular confirmation than ^{19}F or ^{13}C and so if the model is unable to interpret these subtle differences from the SMILES strings resulting in the higher observed error.

Finally, pretraining of the BERT model is on the Masked Language task on reaction SMILES. While this may teach a large amount about general chemistry and how SMILES represent that chemistry, it might not be good enough to understand the subtle local effects that will dominate in NMR Prediction. Therefore, it might be more useful to pre-train on chemical data where very small local differences are present during the MLM task. This could be as simple as training a model to predict ^{13}C NMR and then using that learnt task to predict ^{1}H. A more complex example would be on training on chemical molecules where atoms were randomly changed to the labels used during training such as ^{1}H and ^{13}C so that the model had also learnt to take these into account.

5.4 Methods

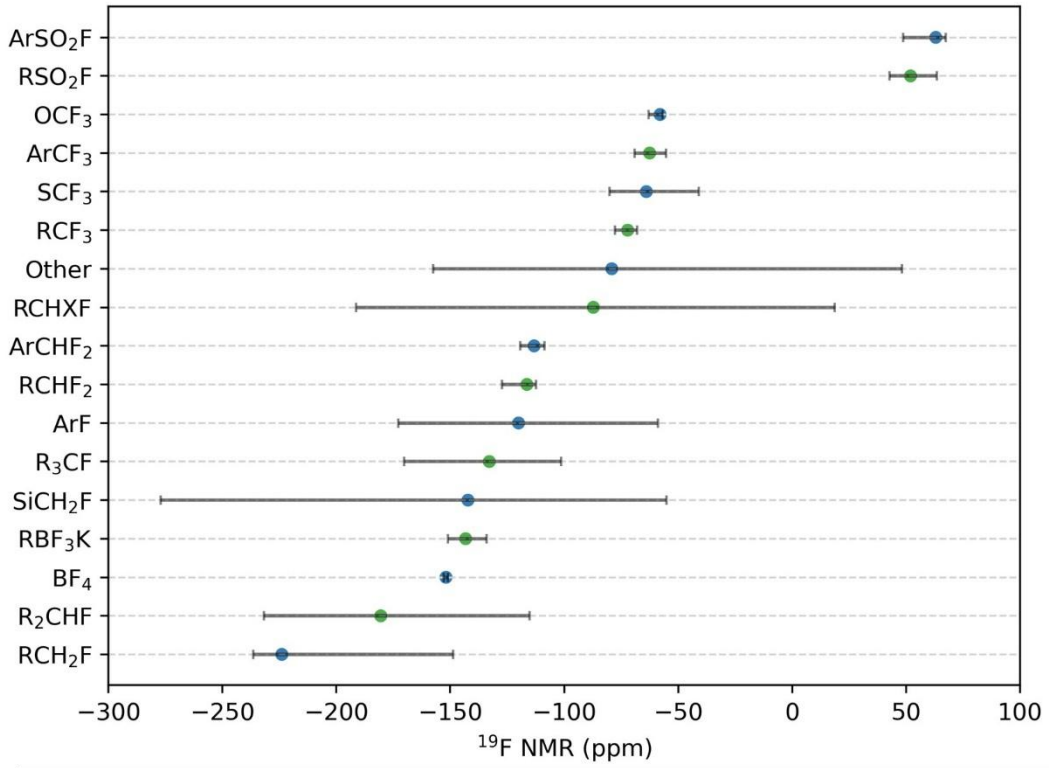
5.4.1 ^{19}F Dataset

Unlike ^{1}H and ^{13}C, currently, there are no extensive databases containing experimental or computed ^{19}F NMR. Small ^{19}F datasets can be found in NMRShiftDB, but they only represent a limited region of the chemical space. Therefore, we decided to generate a dataset based on

molecules for which ^{19}F NMR shifts have been measured experimentally and which cover a wide range of ^{19}F chemical environments.

Data was extracted from 24 papers (See Appendix 3), covering a wide range of chemical environments. Compounds were stored as SMILES strings generated in Chemdraw20, and the corresponding NMR shifts and solvent were noted alongside. Structures which could not be canonicalized by RDKIT were excluded from the dataset.

From this search, we collected 943 ^{19}F environments from 803 separate molecules including aromatic and aliphatic environments C-F, CF_3, CF_2H and SO_2F groups (Figure 79). The database can be further divided into subgroups based on the local chemical environment, *e.g.* aryl, alkenyl, and alkyl groups. SCF_3 and OCF_3 groups were added as a further subset.



Groups	Number of Examples	Average of Experimental	Min of Experimental	Max of Experimental
$ArCHF_2$	60	-113.23	-119.40	-108.70
$ArCF_3$	77	-62.55	-69.10	-55.42
ArF	201	-120.05	-172.80	-58.90
$ArSO_2F$	16	62.99	48.71	67.27
RBF_3K	17	-143.26	-151.00	-134.1
BF_4	6	-151.90	-152.86	-151.07
R_2CHF	281	-180.50	-231.70	-115.29
R_3CF	32	-132.88	-170.14	-101.45
RCF_3	20	-72.188	-77.66	-68.01
RCH_2F	56	-223.90	-236.46	-148.80
$RCHF_2$	27	-116.39	-127.43	-112.40
OCF_3	31	-57.97	-62.96	-56.88
SCF_3	27	-63.99	-80.09	-41.06
SiCF	7	-142.25	-277.00	-55.20
SO_2F	39	51.91	42.70	63.32
RCHXF	34	-87.25	-191.30	18.60
Other	12	-79.165	-157.64	48.16
Grand Total	943	-123.64	-277.00	67.27

Figure 79 Summary of ^{19}F NMR Dataset showing environment types, population, and experimental range. R signifies any atom but a H or Halogen. X signifies a Halogen atom. Other signifies any other functional group not covered by the classifications with less than 5 examples in the dataset.

5.4.2 Environment masking

For mono-fluorinated compounds, ChemDraw representations are directly converted into their corresponding SMILES string and added to the database along with the ^{19}F NMR Shift of the compound (Figure 65A). Structures with multiple fluorine environments were processed as follows. SMILES strings were visualised in ChemDraw, For each environment, other than the target environment, all other F atoms were replaced with ^{18}F, the structure was then converted back into a SMILES string and returned to the database with the matching NMR Shift (Figure 65B).

For the automated procedure of NMRShiftDB and Paton dataset, sdf files were read into RDKIT, and NMR shifts and assignments were extracted from the corresponding entry and extracted into a list. For every NMR shift our code converts the nucleus to the isotope label, ^{1}H, ^{13}C and ^{19}F for hydrogen, carbon, and fluorine respectively. The molecule is then saved as a SMILES string with the NMR shift before the process is repeated for each NMR shift in the molecule.

5.4.3 Molecule descriptor generation

For initial model testing a subset of monofluorinated compounds was extracted. To compare representation of molecules both Fingerprints and SOAP[195] descriptors were tested. Fingerprint generation was carried out using the Morgan fingerprint method in RDKIT[246] using a radius of 3 and a length of 2048. For the generation of SOAP descriptors, 3D confirmations were obtained using the gnff method in xTB v6.4[50] before SOAP descriptors were calculated for the fluorine in each atom using ASE[311] and Dscribe[312] with the following arguments; rcut = 3,nmax = 4,lmax = 4,rbf = 'polynomial'.

5.4.4 Model screening

To identify the best ML algorithm for ^{19}F NMR prediction a range of ML methods were screened including Linear Regression, Support Vector Machines (SVM), Decision Trees, K Nearest Neighbours and Gaussian Processes using an in-house Python script. Both Molecular Fingerprints and SOAP descriptors were screened using this method. A BERT model was also trained using SMILES strings as an input.

5.4.5 Data preparation

To study how different train:test splitting methods perform, three methods were investigated: cross-validation, data splitting and data shuffling.

The impacts of shuffling vs. splitting were investigated by training two models each with a 2:1 training: test split where the training data was either shuffled ten times or a five-fold data split was performed. All model parameters and data were kept consistent for this test.

For the Tanimoto Data splitting a modified script from Lee *et al*[313] was used in which each molecule in the test set has less than a 0.4 Tanimoto similarity score to the training data.

Table 6 Summary of all Models trained and their respective dataset, validation method and model test error. Models with identical datasets use the same test sets. Dataset size shows the size of train: test data sets. Tanimoto identifies if the train:test split was carried out with Tanimoto criteria or a random split

Model	*Dataset Size*	*Tanimoto*	*Dataset*	*Validation Method*	*Test Error (ppm)*
1	295:63	N	Mono-First	5-fold CV	20
2	295:63	N	Mono-First	5-fold Split	26
3	459:91	N	Mono-Final	5-fold Split	11
4	565:113	N	Multi-First	5-fold Split	14
5	629:314	N	Final	5-fold Split	17
6	629:314	N	Final	10-Fold Shuffle	9.4
7	629:314	N	Final	10-Fold CV	8.5
8	854:89	Y	Final	10-Fold CV	13
9	854:89	Y	Final	10-Fold Shuffle	14

5.4.6 BERT model hypermeter optimisation

To identify the best dropout rate and learning rate, both of which are variable hyperparameters, a Bayesian optimisation strategy was used. Utilising the wandb optimiser a single model was trained on one-fold of the dataset. The optimiser varies the hyperparameters using Bayesian inference to identify the optimum hyperparameters for the model. The optimum hyperparameters are outlined in Table 1.

Table 7 Hyperparameters used for the training of all models. Hyperparameters were chosen using a Bayesian Optimisation in wandb

Model	*Learning Rate*	*Drop Out*
^{19}F NMR Model	0.00009913	0.6153
^{19}F Large Bert	0.0002388	0.7837
^{1}H NMR Model	0.00009913	0.6153
^{13}C NMR Model	0.0001885	0.3261
Multi Nuclei NMR Model	0.00009913	0.6153

5.4.7 ^{1}H and ^{13}C datasets

For training on ^{1}H and ^{13}C data two different datasets were tested, NMRShiftDB and Paton Dataset. The NMRShiftDB is an open-source dataset where experimental spectra are uploaded along with assignments and metadata about the experiment which contains 46,000 molecules. The Paton dataset is an 8,000-molecule subset of NMRShiftDB where ^{1}H and ^{13}C shifts were computationally generated using conformational sampling and DFT calculations.

When pre-processing data in NMRShiftDB it was noted that while the data reports to be from experimental spectra the majority, 88%, come from computational calculations rather than experimental spectra. In cases where NMR shifts were computed the Lamour frequency value is set to 0; those data points were not included. To reduce noise due to solvent effects we only used data obtained in deuterated chloroform for our proton database.

For both NMRShiftDB and Paton Dataset, we extracted all ^{1}H and ^{13}C NMR shifts along with the corresponding molecules stored as SMILES strings. Atomic labelling was performed using an in-house script to match each environment to the atoms which were then labelled as either a ^{1}H or ^{13}C. No further data cleaning was performed, and the data was extracted assuming all assignments were correct and labelling consistent. For the NMRShiftDB we also included a

separate column for the NMR solvent, unless otherwise stated we only studied compounds measured in $CDCl_3$. After data curation, the NMRShiftDB dataset contained 5964 and 45594 datapoints for ^{1}H and ^{13}C respectively while the Paton dataset contained 117962 and 99061 datapoints for ^{1}H and ^{13}C.

5.4.8 LargeNMRBERT

For the generation of the of the LargeNMRBert model we increased the number of attention heads from 4 to 12 and the size of the intermediate and hidden layers, from 256 to 512 and 512 to 3072 respectively. The training for the Masked Language Model (MLM) was carried out according to *Schwaller et al*,[152] except the learning rate, which had to be reduced by a factor of 10 to avoid a local minimum. We performed a hyperparameter sweep on our model to identify the optimum learning rate and dropout probability before training the model to predict ^{19}F NMR shifts. The same train and test split was used to evaluate the model's accuracy.

5.4.9 Fingerprint comparisons

Using the CLS token as a molecular fingerprint of each compound, several examples were selected to compare how a range of subtle chemical differences were comprehended by the model. The similarity between two compounds is defined as,

$$similarity(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{1 + \|\boldsymbol{U} - \boldsymbol{V}\|}$$

Where **U** and **V** are the hidden representations (fingerprints) of the molecules.

5.4.10 Integrated gradient methods

One of the challenges when using NNs is the black-box nature of the models, meaning that understanding why NNs reach a given decision is challenging. Integrated Gradients (IG), as introduced in Chapter 2, aim to interpret the output of NNs. Recently Lee *et al.* used IGs to study how transformer models interpret chemical selectivity.[313]

As the NMRBERT has a regression output we directly use the NMR shift as our output, in line with other published work on IGs for regression tasks.[314] The IG baselines were generated as a string of '.' tokens the same length as our SMILES strings with the corresponding CLS and SEP tokens prepended and appended to the string respectively.

5.4.11 Calculating confidence values for regioselective reactions

To compare two possible products, as in the case of regioisomers, a confidence value was calculated using the method reported by Goodman in his DP4 approach.[46, 48, 49]

$$Error = Computed - Experimental \tag{5.1}$$

$$z = \frac{Error - MeanErrorF}{StdevF} \tag{5.2}$$

$$\rho = 2 \times cdf(-z) \tag{5.3}$$

$$Confidence = \frac{\rho_i}{\rho} \times 100 \tag{5.4}$$

Where MeanErrorF is the mean of the errors during on training data and the StdevF is the standard deviation of the errors on our training data.

Chapter 6 Conclusions

The work in this thesis broadly covers the use of Density Functional Theory (DFT) and Machine Learning (ML) to study fluorine-containing molecules. Chapter 3 and Chapter 4 focus on the hydrogen bonding phase transfer catalysis (HBPTC) complexes pioneered by the Gouverneur Group, while Chapter 5 focuses on general organofluorine compounds and the prediction of ^{19}F NMR .

6.1 Thesis breakdown

In Chapter 3 we firstly developed a workflow for the generation of molecular descriptors and then the prediction of enantioselectivity in the HBPTC reaction. This Python workflow can use molecular fingerprints, Mordred descriptors, and DFT-derived descriptors to encode the reaction, and then identifies the optimal ML algorithm for predicting enantioselectivity. We also developed *graphSterimol,* an adapted Python script which allows for the whole molecule to be included during the calculation of Sterimol parameters. The addition of the whole molecule allows for the bulk of a molecule to influence the conformations of the groups Sterimol values are calculated for, which can result in models with higher R^2 and lower RMSE when compared to literature examples.

We applied this new workflow to experimental datasets from the Gouverneur Group, which showed that the combination of DFT descriptors and LASSO algorithm was an optimal balance of computational time, accuracy and interpretability. The model was able to predict enantioselectivities with errors of less than 1.05 kJ/mol for different catalysts and substrates. We then interrogated the trained model, allowing for quantification of the contributions from different parts of each catalyst or substrate. Using the knowledge gained we computationally generated and screened a range of potential catalysts and substrates for our experimental collaborators to see if they were good potential candidates. Unfortunately, COVID-19 meant

our collaborators were unable to test our best-performing catalysts, but this remains an area of possible future work.

In Chapter 4 we delved further into the study of the HBPTC reaction by focusing on the development of an approach for the calculation of ^{19}F NMR and $^{1}J_{HF}$ coupling constants in the HBPTC catalyst complexes. This method uses conformational sampling in xTB using the CREST package before DFT optimisations and NMR calculations are performed. After benchmarking our DFT calculations on two available small datasets, we found that the method can calculate both ^{19}F NMR shifts and the $^{1}J_{HF}$ coupling constants across the H-F hydrogen bond easily. Yet the calculation of $^{1}J_{HF}$ coupling constants for the covalent H-F bond still remains a significant challenge for DFT methods with errors of more than 200 Hz.

When we applied our DFT approach to mono urea fluoride complexes (MUF) we found that both the ^{19}F NMR and $^{1}J_{HF}$ coupling constants do correlate well with the Hammett values for para-substituted benzenes. An important observation was the significant difference between calculations with implicit solvent vs explicit solvation. In the case of explicit solvation, we observed much smaller $^{1}J_{HF}$ coupling constants, due to the H-F hydrogen bonds lengthening by over 0.1 Å. Yet, while there are significant differences in the absolute values of the ^{19}F NMR shifts and $^{1}J_{HF}$ coupling constants between different solvation methods, the correlation with Hammett values remains. Therefore, comparison between different catalysts is viable, while the prediction of absolute values remains challenging.

We then further tested the approach on the BINAM catalyst complexes from the Gouverneur group. The $^{1}J_{HF}$ coupling constant RMSE of 3.4 Hz allows for differentiation between all three of the hydrogen bonds present in the complexes. We also observed that two possible binding conformations were possible: an open conformation where the fluoride is exposed to bulk solvent, and a closed conformation where the TBA cation sits above the fluoride. The former is the lowest energy binding mode. However, as seen with the MUF complexes, we observed

that the H-F hydrogen bond lengths here are shorter than experimentally determined. This is likely due to a lack of explicit solvent in the open conformation which would lengthen the hydrogen bonds resulting in them being more in line with experimental results.

In Chapter 5 we developed a transformer-based ML model which could predict the ^{19}F NMR shifts for a range of organo-fluorinated molecules with an RMSE of 8.5 ppm. This model was also extended to ^{13}C NMR where an accuracy of 6.8 ppm was achieved. However, when we extended to the prediction of ^{1}H NMR the model struggled with the best-performing model only achieving an accuracy of 0.59 ppm. We also discovered that a model which was trained to predict ^{1}H, ^{13}C and ^{19}F NMR at the same time improved on the individually trained models with RMSEs of 0.55, 3.3 and 8.2 ppm respectively. This shows that increasing the diversity of data and NMR environments shown during training can lead to reduced error and therefore could improve future model development.

We were also interested in investigating what chemistry was learnt by our models. Fingerprint comparisons showed that the model was not purely identifying structures which would have similar NMR and placing them close by in chemical space but rather would keep them apart in chemical space and then convert that representation into similar NMR shifts. When we used Integrated Gradients (IGs) to understand the importance of each token we see that with some compounds, such as diastereomers, the model is unable to resolve the difference in structure to predict a different NMR shift. IGs also showed that for different substituted benzenes the model was learning the subtle changes of electron donation and withdrawing effects and their effect on NMR shifts. We finally applied our model to the identification of late-stage fluorination reactions, to identify which regio-isomer was synthesised purely by its ^{19}F NMR shift. For aromatic fluorines, this proved to be very successful with 8/13 examples being correctly

identified. Further work is needed to identify aromatic trifluoromethyl groups, however, which only achieved 2/10 on our test data.

6.2 Future directions

With the development of *graphSterimol* in Chapter 3 we only focused on the calculation of Sterimol parameters for the lowest energy conformation for each catalyst or substrate as the DFT descriptors were calculated only on the lowest energy conformation. One interesting implementation would be to include conformationally averaged Sterimol values using this method. This has been implemented for the R groups alone, but not with the whole molecule included in the conformational sampling. Extending this code to include the whole conformational ensemble, such as one obtained from CREST, could lead to a more accurate description of steric bulk in both catalysts and substrates. To validate the model's prediction, it could also be useful to experimentally test several of the best-performing, and most synthetically accessible catalysts or substrates. This would allow for a further quantification of both the limits of the model and add in further data which could lead to a refinement in its predictions.

While the DFT approach we used for the prediction of ^{19}F NMR shifts and $^{1}J_{HF}$ coupling constants in Chapter 4 does allow for the relative prediction of both constants, the absolute values currently are outside the capabilities of our method. One area of possible further research is in the development of DFT functionals which are better able to predict the $^{1}J_{HF}$ coupling constants across the H-F covalent bond. From our work, the best functional seem to include a large amount of HF exchange and therefore further development in this direction could lead to improved prediction of the $^{1}J_{HF}$ coupling constants. A possible direction would be the development of DH-DFT functionals for coupling constant prediction. However, this is currently not implemented in ORCA. A further interesting inclusion in ORCA 6.0, which was

recently released, is the ability to calculate zero-point vibrational corrections to both NMR shifts and coupling constants. The inclusion of this could further lead to an improvement in the DFT prediction of these properties especially in the small H-F clusters. However, the most challenging area for future work is in the modelling of explicit solvents in the HBPTC catalysts. As the major class of conformations leave the fluoride anion exposed to bulk solvent, accurate representation of explicit solvent would lead to less overbinding. While our own attempts to use MD simulations resulted in complex instability, future work could focus on forcefield development which would be able to reproduce the stability of the complex and the expected binding. This forcefield could then be used to sample catalyst conformations and then DFT calculations could be performed on each cluster with explicit solvent. Another option is the use of machine learning potentials (MLPs) instead of classical MD simulations. This could allow for more accurate and faster conformational sampling of the catalyst complexes, however, the size of the system involved could mean the expense of MLPs is not computationally viable.

For the prediction of ^{19}F NMR using our NMRBERT model, we found that our model performed well on a variety of different fluorine environments such as aromatic fluorines, primary alkyl fluorines, and aromatic trifluoromethyl groups. However, the model struggles to perform on both secondary alkyl fluorines and diastereomers. One problem is data collection, as in a lot of examples only one possible diastereomer is reported and therefore experimental data is limited. This could be supplemented through the use of DFT calculations, such as those carried out in Chapter 4. This would allow for a greater amount of data to be used during training but could introduce further errors based on the level of DFT theory used. A more subtle question and one well outside the scope of this thesis is whether or not transformer-based models are able to capture the subtle 3D conformations and through-space effects that can be crucial in NMR prediction. The work in Chapter 5 suggests that the models struggle with ^{1}H prediction more than both ^{13}C and ^{19}F. This, therefore, implies that the model is not capturing

these local effects, which are a significant contribution in ^{1}H NMR. This could unfortunately be a current limitation with SMILES-based transformers but could be the focus of further transformer-based model design. An initial test of this hypothesis could focus on training a BERT-based model firstly on asymmetric reactions where chirality is key for the reaction prediction. This would expose the model to a lot more chiral information during its pre-training phase which could lead to a better understanding of chiral local environments when it comes to prediction on ^{1}H data.

6.3 Final thoughts

This thesis shows an overview of the different techniques available to computational organic chemists to study a range of chemical problems. We aim that the work in the thesis does show areas which could be potential pitfalls in future studies. We have shown that DFT and ML can be an aid in the study of both enantioselectivity prediction and NMR prediction, but careful management of data, code, and a knowledge of chemistry are key to being able to solve these challenging problems in the future.

Chapter 7 References

(1) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc Chem Res* **2017**, *50* (3), 539-543. DOI: 10.1021/acs.accounts.6b00532.
(2) Houk, K. N.; Cheong, P. H.-Y. Computational prediction of small-molecule catalysts. *Nature* **2008**, *455* (7211), 309-313. DOI: 10.1038/nature07368.
(3) Cheng, G. J.; Zhang, X.; Chung, L. W.; Xu, L.; Wu, Y. D. Computational organic chemistry: bridging theory and experiment in establishing the mechanisms of chemical reactions. *J Am Chem Soc* **2015**, *137* (5), 1706-1725. DOI: 10.1021/ja5112749.
(4) Bachrach, S. M. *Computational organic chemistry*; John Wiley & Sons, 2014.
(5) Bachrach, S. M. Computational organic chemistry. *Annual Reports Section "B" (Organic Chemistry)* **2008**, *104*, 394. DOI: 10.1039/b719311b.
(6) Hopmann, K. H. Quantum chemical studies of asymmetric reactions: Historical aspects and recent examples. *Int J Quantum Chem* **2015**, *115* (18), 1232-1249. DOI: 10.1002/qua.24882.
(7) Sperger, T.; Sanhueza, I. A.; Schoenebeck, F. Computation and Experiment: A Powerful Combination to Understand and Predict Reactivities. *Acc Chem Res* **2016**, *49* (6), 1311-1319. DOI: 10.1021/acs.accounts.6b00068.
(8) Peng, Q.; Duarte, F.; Paton, R. S. Computing organic stereoselectivity – from concepts to quantitative calculations and predictions. *Chem Soc Rev* **2016**, *45* (22), 6093-6107. DOI: 10.1039/c6cs00573j.
(9) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational prediction of 1H and 13C chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem Rev* **2012**, *112* (3), 1839-1862. DOI: 10.1021/cr200106v.
(10) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J. L. Prediction of chemical reaction yields using deep learning. *Mach Learn-Sci Techn* **2021**, *2* (1), 015016-015016. DOI: 10.1088/2632-2153/abc81d.
(11) Inoue, M.; Sumii, Y.; Shibata, N. Contribution of Organofluorine Compounds to Pharmaceuticals. *ACS Omega* **2020**, *5* (19), 10633-10640. DOI: 10.1021/acsomega.0c00830.
(12) Ogawa, Y.; Tokunaga, E.; Kobayashi, O.; Hirai, K.; Shibata, N. Current Contributions of Organofluorine Compounds to the Agrochemical Industry. *iScience* **2020**, *23* (9), 101467. DOI: 10.1016/j.isci.2020.101467.
(13) O'Hagan, D. Understanding organofluorine chemistry. An introduction to the C–F bond. *Chem. Soc. Rev.* **2008**, *37* (2), 308-319. DOI: 10.1039/b711844a.
(14) Gillis, E. P.; Eastman, K. J.; Hill, M. D.; Donnelly, D. J.; Meanwell, N. A. Applications of Fluorine in Medicinal Chemistry. *J Med Chem* **2015**, *58* (21), 8315-8359. DOI: 10.1021/acs.jmedchem.5b00258.
(15) Patel, C.; André-Joyaux, E.; Leitch, J. A.; De Irujo-Labalde, X. M.; Ibba, F.; Struijs, J.; Ellwanger, M. A.; Paton, R.; Browne, D. L.; Pupo, G.; Aldridge, S.; Hayward, M. A.; Gouverneur, V. Fluorochemicals from fluorspar via a phosphate-enabled mechanochemical process that bypasses HF. *Science* **2023**, *381* (6655), 302-306. DOI: 10.1126/science.adi1557.
(16) Pupo, G.; Ibba, F.; Ascough, D. M. H.; Vicini, A. C.; Ricci, P.; Christensen, K. E.; Pfeifer, L.; Morphy, J. R.; Brown, J. M.; Paton, R. S.; Gouverneur, V. Asymmetric nucleophilic fluorination under hydrogen bonding phase-transfer catalysis. *Science* **2018**, *360* (6389), 638-642. DOI: 10.1126/science.aar7941.

(17) Lozano, O.; Blessley, G.; Martinez del Campo, T.; Thompson, A. L.; Giuffredi, G. T.; Bettati, M.; Walker, M.; Borman, R.; Gouverneur, V. Organocatalyzed enantioselective fluorocyclizations. *Angew Chem Int Ed Engl* **2011**, *50* (35), 8105-8109. DOI: 10.1002/anie.201103151.
(18) Katcher, M. H.; Sha, A.; Doyle, A. G. Palladium-catalyzed regio- and enantioselective fluorination of acyclic allylic halides. *J Am Chem Soc* **2011**, *133* (40), 15902-15905. DOI: 10.1021/ja206960k.
(19) Lovett, G. H.; Chen, S.; Xue, X. S.; Houk, K. N.; MacMillan, D. W. C. Open-Shell Fluorination of Alkyl Bromides: Unexpected Selectivity in a Silyl Radical-Mediated Chain Process. *J Am Chem Soc* **2019**, *141* (51), 20031-20036. DOI: 10.1021/jacs.9b11434.
(20) Nagib, D. A.; MacMillan, D. W. Trifluoromethylation of arenes and heteroarenes by means of photoredox catalysis. *Nature* **2011**, *480* (7376), 224-228. DOI: 10.1038/nature10647.
(21) Miller, E.; Kim, S.; Gibson, K.; Derrick, J. S.; Toste, F. D. Regio- and Enantioselective Bromocyclization of Difluoroalkenes as a Strategy to Access Tetrasubstituted Difluoromethylene-Containing Stereocenters. *J Am Chem Soc* **2020**, *142* (19), 8946-8952. DOI: 10.1021/jacs.0c02331.
(22) Yang, X.; Wu, T.; Phipps, R. J.; Toste, F. D. Advances in Catalytic Enantioselective Fluorination, Mono-, Di-, and Trifluoromethylation, and Trifluoromethylthiolation Reactions. *Chem Rev* **2015**, *115* (2), 826-870. DOI: 10.1021/cr500277b.
(23) Honjo, T.; Phipps, R. J.; Rauniyar, V.; Toste, F. D. A doubly axially chiral phosphoric acid catalyst for the asymmetric tandem oxyfluorination of enamides. *Angew Chem Int Ed Engl* **2012**, *51* (38), 9684-9688. DOI: 10.1002/anie.201205383.
(24) Liu, W.; Huang, X.; Cheng, M. J.; Nielsen, R. J.; Goddard, W. A., 3rd; Groves, J. T. Oxidative aliphatic C-H fluorination with fluoride ion catalyzed by a manganese porphyrin. *Science* **2012**, *337* (6100), 1322-1325. DOI: 10.1126/science.1222327.
(25) Li, J.; Chen, J.; Sang, R.; Ham, W. S.; Plutschack, M. B.; Berger, F.; Chabbra, S.; Schnegg, A.; Genicot, C.; Ritter, T. Photoredox catalysis with aryl sulfonium salts enables site-selective late-stage fluorination. *Nat Chem* **2020**, *12* (1), 56-62. DOI: 10.1038/s41557-019-0353-3.
(26) Dolbier Jr, W. R. *Guide to fluorine NMR for organic chemists*; John Wiley & Sons, 2016.
(27) Claridge, T. D. *High-resolution NMR techniques in organic chemistry*; Elsevier, 2016.
(28) Senn, H. M.; O'Hagan, D.; Thiel, W. Insight into enzymatic C-F bond formation from QM and QM/MM calculations. *J Am Chem Soc* **2005**, *127* (39), 13643-13655. DOI: 10.1021/ja053875s.
(29) Dong, C.; Huang, F.; Deng, H.; Schaffrath, C.; Spencer, J. B.; O'Hagan, D.; Naismith, J. H. Crystal structure and mechanism of a bacterial fluorinating enzyme. *Nature* **2004**, *427* (6974), 561-565. DOI: 10.1038/nature02280.
(30) Raheem, I. T.; Thiara, P. S.; Peterson, E. A.; Jacobsen, E. N. Enantioselective Pictet-Spengler-type cyclizations of hydroxylactams: H-bond donor catalysis by anion binding. *J Am Chem Soc* **2007**, *129* (44), 13404-13405. DOI: 10.1021/ja076179w.
(31) Taylor, M. S.; Jacobsen, E. N. Asymmetric Catalysis by Chiral Hydrogen-Bond Donors. *Angew Chem Int Ed* **2006**, *45* (10), 1520-1543. DOI: 10.1002/anie.200503132.
(32) Schreiner, P. R. Metal-free organocatalysis through explicit hydrogen bonding interactions. *Chem Soc Rev* **2003**, *32* (5), 289-296. DOI: 10.1039/b107298f.
(33) Maruoka, K. Practical Aspects of Recent Asymmetric Phase-Transfer Catalysis. *Org Process Res Dev* **2008**, *12* (4), 679-697. DOI: 10.1021/op7002979.
(34) Hashimoto, T.; Maruoka, K. Recent Development and Application of Chiral Phase-Transfer Catalysts. *Chem Rev* **2007**, *107* (12), 5656-5682. DOI: 10.1021/cr068368n.

(35) Maruoka, K.; Ooi, T. Enantioselective amino acid synthesis by chiral phase-transfer catalysis. *Chem Rev* **2003**, *103* (8), 3013-3028. DOI: 10.1021/cr020020e.
(36) Pirkle, W. H.; Snyder, S. E. Two-component chiral phase transfer catalysts: enantioselective esterification of an N-acylated amino acid. *Org Lett* **2001**, *3* (12), 1821-1823. DOI: 10.1021/ol015823n.
(37) Corey, E. J.; Xu, F.; Noe, M. C. A Rational Approach to Catalytic Enantioselective Enolate Alkylation Using a Structurally Rigidified and Defined Chiral Quaternary Ammonium Salt under Phase Transfer Conditions. *J Am Chem Soc* **1997**, *119* (50), 12414-12415. DOI: 10.1021/ja973174y.
(38) Pupo, G.; Vicini, A. C.; Ascough, D. M. H.; Ibba, F.; Christensen, K. E.; Thompson, A. L.; Brown, J. M.; Paton, R. S.; Gouverneur, V. Hydrogen Bonding Phase-Transfer Catalysis with Potassium Fluoride: Enantioselective Synthesis of beta-Fluoroamines. *J Am Chem Soc* **2019**, *141* (7), 2878-2883. DOI: 10.1021/jacs.8b12568.
(39) Ibba, F.; Pupo, G.; Thompson, A. L.; Brown, J. M.; Claridge, T. D. W.; Gouverneur, V. Impact of Multiple Hydrogen Bonds with Fluoride on Catalysis: Insight from NMR Spectroscopy. *J Am Chem Soc* **2020**, *142* (46), 19731-19744. DOI: 10.1021/jacs.0c09832.
(40) Benassi, E. Benchmarking of density functionals for a soft but accurate prediction and assignment of (1) H and (13)C NMR chemical shifts in organic and biological molecules. *J Comput Chem* **2017**, *38* (2), 87-92. DOI: 10.1002/jcc.24521.
(41) Flaig, D.; Maurer, M.; Hanni, M.; Braunger, K.; Kick, L.; Thubauville, M.; Ochsenfeld, C. Benchmarking Hydrogen and Carbon NMR Chemical Shifts at HF, DFT, and MP2 Levels. *J Chem Theory Comput* **2014**, *10* (2), 572-578. DOI: 10.1021/ct400780f.
(42) Laskowski, R.; Blaha, P.; Tran, F. Assessment of DFT functionals with NMR chemical shifts. *Phys Rev B* **2013**, *87* (19). DOI: 10.1103/PhysRevB.87.195130.
(43) Lacerda, E. G., Jr.; Kamounah, F. S.; Coutinho, K.; Sauer, S. P. A.; Hansen, P. E.; Hammerich, O. Computational Prediction of (1) H and (13) C NMR Chemical Shifts for Protonated Alkylpyrroles: Electron Correlation and Not Solvation is the Salvation. *Chemphyschem* **2019**, *20* (1), 78-91. DOI: 10.1002/cphc.201801066.
(44) Auer, A. A.; Gauss, J.; Stanton, J. F. Quantitative prediction of gas-phase C13 nuclear magnetic shielding constants. *J Chem Phys* **2003**, *118* (23), 10407-10417. DOI: 10.1063/1.1574314.
(45) Iron, M. A. Evaluation of the Factors Impacting the Accuracy of (13)C NMR Chemical Shift Predictions using Density Functional Theory-The Advantage of Long-Range Corrected Functionals. *J Chem Theory Comput* **2017**, *13* (11), 5798-5819. DOI: 10.1021/acs.jctc.7b00772.
(46) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem Sci* **2020**, *11* (17), 4351-4359. DOI: 10.1039/d0sc00442a.
(47) Howarth, A.; Goodman, J. M. The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. *Chem Sci* **2022**, *13* (12), 3507-3518. DOI: 10.1039/d1sc04406k.
(48) Smith, S. G.; Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J Am Chem Soc* **2010**, *132* (37), 12946-12959. DOI: 10.1021/ja105035r.
(49) Ermanis, K.; Parkes, K. E. B.; Agback, T.; Goodman, J. M. The optimal DFT approach in DP4 NMR structure analysis - pushing the limits of relative configuration elucidation. *Org Biomol Chem* **2019**, *17* (24), 5886-5890. DOI: 10.1039/c9ob00840c.
(50) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB-An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole

Electrostatics and Density-Dependent Dispersion Contributions. *J Chem Theory Comput* **2019**, *15* (3), 1652-1671. DOI: 10.1021/acs.jctc.8b01176.
(51) Grimme, S.; Bannwarth, C.; Dohm, S.; Hansen, A.; Pisarek, J.; Pracht, P.; Seibert, J.; Neese, F. Fully Automated Quantum-Chemistry-Based Computation of Spin-Spin-Coupled Nuclear Magnetic Resonance Spectra. *Angew Chem Int Ed Engl* **2017**, *56* (46), 14763-14769. DOI: 10.1002/anie.201708266.
(52) Bagno, A. Complete Prediction of the1H NMR Spectrum of Organic Molecules by DFT Calculations of Chemical Shifts and Spin-Spin Coupling Constants. *Chemistry* **2001**, *7* (8), 1652-1661. DOI: 10.1002/1521-3765(20010417)7:8<1652::Aid-chem16520>3.0.Co;2-v.
(53) Smith, S. G.; Goodman, J. M. Assigning the stereochemistry of pairs of diastereoisomers using GIAO NMR shift calculation. *J Org Chem* **2009**, *74* (12), 4597-4607. DOI: 10.1021/jo900408d.
(54) Smith, S. G.; Paton, R. S.; Burton, J. W.; Goodman, J. M. Stereostructure assignment of flexible five-membered rings by GIAO 13C NMR calculations: prediction of the stereochemistry of elatenyne. *J Org Chem* **2008**, *73* (11), 4053-4062. DOI: 10.1021/jo8003138.
(55) Rosenau, C. P.; Jelier, B. J.; Gossert, A. D.; Togni, A. Exposing the Origins of Irreproducibility in Fluorine NMR Spectroscopy. *Angew Chem Int Ed Engl* **2018**, *57* (30), 9528-9533. DOI: 10.1002/anie.201802620.
(56) Saielli, G.; Bini, R.; Bagno, A. Computational 19F NMR. 1. General features. *Theor Chem Acc* **2012**, *131* (3), 1-11. DOI: 10.1007/s00214-012-1140-z.
(57) Saielli, G.; Bini, R.; Bagno, A. Computational 19F NMR. 2. Organic compounds. *RSC Adv.* **2014**, *4* (78), 41605-41611. DOI: 10.1039/c4ra08290g.
(58) Saunders, C.; Khaled, M. B.; Weaver, J. D., 3rd; Tantillo, D. J. Prediction of (19)F NMR Chemical Shifts for Fluorinated Aromatic Compounds. *J Org Chem* **2018**, *83* (6), 3220-3225. DOI: 10.1021/acs.joc.8b00104.
(59) Dumon, A. S.; Rzepa, H. S.; Alamillo-Ferrer, C.; Bures, J.; Procter, R.; Sheppard, T. D.; Whiting, A. A computational tool to accurately and quickly predict 19F NMR chemical shifts of molecules with fluorine–carbon and fluorine–boron bonds. *Phys Chem Chem Phys* **2022**, *24* (34), 20409-20425, 10.1039/D2CP02317B. DOI: 10.1039/D2CP02317B.
(60) Gerken, M.; Boatz, J. A.; Kornath, A.; Haiges, R.; Schneider, S.; Schroer, T.; Christe, K. O. The NMR shifts are not a measure for the nakedness of the fluoride anion. *J Fluorine Chem* **2002**, *116* (1), 49-58. DOI: 10.1016/s0022-1139(02)00101-x.
(61) Richter, W. E.; Pontes, R. M.; Abiko, L. A.; Gauze, G. F.; Basso, E. A. Computation of 3JHH coupling constants with a combination of density functional theory and semiempirical calculations. Application to complex molecules. *Comput Theor Chem* **2012**, *1001*, 7-14. DOI: 10.1016/j.comptc.2012.10.019.
(62) Lopez-Vallejo, F.; Fragoso-Serrano, M.; Suarez-Ortiz, G. A.; Hernandez-Rojas, A. C.; Cerda-Garcia-Rojas, C. M.; Pereda-Miranda, R. Vicinal 1H-1H NMR coupling constants from density functional theory as reliable tools for stereochemical analysis of highly flexible multichiral center molecules. *J Org Chem* **2011**, *76* (15), 6057-6066. DOI: 10.1021/jo200637g.
(63) Li, Y. Structural revision of glabramycins B and C, antibiotics from the fungus Neosartorya glabra by DFT calculations of NMR chemical shifts and coupling constants. *Rsc Adv* **2015**, *5* (46), 36858-36864. DOI: 10.1039/c5ra01753j.
(64) Buevich, A. V.; Sauri, J.; Parella, T.; De Tommasi, N.; Bifulco, G.; Williamson, R. T.; Martin, G. E. Enhancing the utility of (1)JCH coupling constants in structural studies through optimized DFT analysis. *Chem Commun (Camb)* **2019**, *55* (41), 5781-5784. DOI: 10.1039/c9cc02469g.

(65) Bally, T.; Rablen, P. R. Quantum-chemical simulation of 1H NMR spectra. 2. Comparison of DFT-based procedures for computing proton-proton coupling constants in organic molecules. *J Org Chem* **2011**, *76* (12), 4818-4830. DOI: 10.1021/jo200513q.
(66) San Fabian, J.; Garcia de la Vega, J. M.; Suardiaz, R.; Fernandez-Oliva, M.; Perez, C.; Crespo-Otero, R.; Contreras, R. H. Computational NMR coupling constants: shifting and scaling factors for evaluating 1JCH. *Magn Reson Chem* **2013**, *51* (12), 775-787. DOI: 10.1002/mrc.4014.
(67) Shenderovich, I. G.; Smirnov, S. N.; Denisov, G. S.; Gindin, V. A.; Golubev, N. S.; Dunger, A.; Reibke, R.; Kirpekar, S.; Malkina, O. L.; Limbach, H. H. Nuclear magnetic resonance of hydrogen bonded clusters between F- and (HF)(n): Experiment and theory. *Ber Bunsen Phys Chem* **1998**, *102* (3), 422-428. DOI: DOI 10.1002/bbpc.19981020322.
(68) Comeau, D. C.; Bartlett, R. J. The Equation-of-Motion Coupled-Cluster Method - Applications to Open-Shell and Closed-Shell Reference States. *Chem Phys Lett* **1993**, *207* (4-6), 414-423. DOI: Doi 10.1016/0009-2614(93)89023-B.
(69) Nooijen, M.; Bartlett, R. J. Equation-of-Motion Coupled-Cluster Method for Electron-Attachment. *J Chem Phys* **1995**, *102* (9), 3629-3647. DOI: Doi 10.1063/1.468592.
(70) Perera, S. A.; Nooijen, M.; Bartlett, R. J. Electron correlation effects on the theoretical calculation of nuclear magnetic resonance spin-spin coupling constants. *J Chem Phys* **1996**, *104* (9), 3290-3305. DOI: Doi 10.1063/1.471092.
(71) Perera, S. A.; Bartlett, R. J. NMR Spin−Spin Coupling Constants for Hydrogen Bonds of [F(HF)n]-, n = 1−4, Clusters. *J Am Chem Soc* **2000**, *122* (6), 1231-1232. DOI: 10.1021/ja993275r.
(72) Pecul, M.; Leszczynski, J.; Sadlej, J. Comprehensive ab initio studies of nuclear magnetic resonance shielding and coupling constants in XH···O hydrogen-bonded complexes of simple organic molecules. *J Chem Phys* **2000**, *112* (18), 7930-7938. DOI: 10.1063/1.481394.
(73) Pecul, M.; Sadlej, J.; Leszczynski, J. The 19F–1H coupling constants transmitted through covalent, hydrogen bond, and van der Waals interactions. *J Chem Phys* **2001**, *115* (12), 5498-5506. DOI: 10.1063/1.1398099.
(74) Alkorta, I.; Blanco, F.; Elguero, J. A theoretical structural analysis of the factors that affect (1)J(NH), (1h)J(NH) and (2h)J(NN) in N-H...N hydrogen-bonded complexes. *Magn Reson Chem* **2009**, *47* (3), 249-256. DOI: 10.1002/mrc.2382.
(75) Del Bene, J. E.; Elguero, J. Ab initio study of complexes with two cations as N-H donors to F-: structures and spin-spin coupling constants across N-H-F hydrogen bonds. *J Phys Chem A* **2005**, *109* (47), 10753-10758. DOI: 10.1021/jp0547515.
(76) Shenderovich, I. G.; Burtsev, A. P.; Denisov, G. S.; Golubev, N. S.; Limbach, H. H. Influence of the temperature-dependent dielectric constant on the H/D isotope effects on the NMR chemical shifts and the hydrogen bond geometry of the collidine-HF complex in CDF3/CDClF2 solution. *Magn Reson Chem* **2001**, *39* (S1), S91-S99. DOI: DOI 10.1002/mrc.938.
(77) Alkorta, I.; Elguero, J.; Limbach, H. H.; Shenderovich, I. G.; Winkler, T. A DFT and AIM analysis of the spin-spin couplings across the hydrogen bond in the 2-fluorobenzamide and related compounds. *Magn Reson Chem* **2009**, *47* (7), 585-592. DOI: 10.1002/mrc.2433.
(78) Sychrovský, V. r.; Gräfenstein, J.; Cremer, D. Nuclear magnetic resonance spin–spin coupling constants from coupled perturbed density functional theory. *J Chem Phys* **2000**, *113* (9), 3530-3547. DOI: 10.1063/1.1286806.
(79) Wu, A. A.; Grafenstein, J.; Cremer, D. Analysis of the transmission mechanism of NMR spin-spin coupling constants using Fermi contact spin density distribution, Partial Spin Polarization, and orbital currents: XHn molecules. *J Phys Chem A* **2003**, *107* (36), 7043-7056. DOI: 10.1021/jp0305411.

(80) Tuttle, T.; Grafenstein, J.; Wu, A.; Kraka, E.; Cremer, D. Analysis of the NMR spin-spin coupling mechanism across a H-bond: Nature of the H-bond in proteins. *J Phys Chem B* **2004**, *108* (3), 1115-1129. DOI: 10.1021/jp0363951.
(81) Tuttle, T.; Kraka, E.; Wu, A.; Cremer, D. Investigation of the NMR spin-spin coupling constants across the hydrogen bonds in ubiquitin: the nature of the hydrogen bond as reflected by the coupling mechanism. *J Am Chem Soc* **2004**, *126* (16), 5093-5107. DOI: 10.1021/ja030246e.
(82) Cremer, D.; Grafenstein, J. Calculation and analysis of NMR spin-spin coupling constants. *Phys Chem Chem Phys* **2007**, *9* (22), 2791-2816. DOI: 10.1039/b700737j.
(83) Gao, P.; Zhang, J.; Peng, Q.; Zhang, J.; Glezakou, V. A. General Protocol for the Accurate Prediction of Molecular (13)C/(1)H NMR Chemical Shifts via Machine Learning Augmented DFT. *J Chem Inf Model* **2020**, *60* (8), 3746-3754. DOI: 10.1021/acs.jcim.0c00388.
(84) Unzueta, P. A.; Greenwell, C. S.; Beran, G. J. O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Delta-Machine Learning. *J Chem Theory Comput* **2021**, *17* (2), 826-840. DOI: 10.1021/acs.jctc.0c00979.
(85) Guan, Y.; Shree Sowndarya, S. V.; Gallegos, L. C.; St. John, P. C.; Paton, R. S. Real-time prediction of 1H and 13C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem Sci* **2021**, *12* (36), 12012-12026. DOI: 10.1039/d1sc03343c.
(86) Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB constructing a free chemical information system with open-source components. *J Chem Inf Comput Sci* **2003**, *43* (6), 1733-1739.
(87) Blinov, K. A.; Smurnyy, Y. D.; Elyashberg, M. E.; Churanova, T. S.; Kvasha, M.; Steinbeck, C.; Lefebvre, B. A.; Williams, A. J. Performance validation of neural network based (13)c NMR prediction using a publicly available data source. *J Chem Inf Model* **2008**, *48* (3), 550-555. DOI: 10.1021/ci700363r.
(88) Schrödinger, E. An Undulatory Theory of the Mechanics of Atoms and Molecules. *Phys Rev* **1926**, *28* (6), 1049-1070. DOI: 10.1103/PhysRev.28.1049.
(89) Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Annalen der Physik* **1927**, *389* (20), 457-484. DOI: 10.1002/andp.19273892002.
(90) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys Rev* **1934**, *46* (7), 618-622. DOI: 10.1103/physrev.46.618.
(91) Coester, F.; Kümmel, H. Short-range correlations in nuclear wave functions. *Nuc Phys* **1960**, *17*, 477-485. DOI: 10.1016/0029-5582(60)90140-1.
(92) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem Phys Lett* **1989**, *157* (6), 479-483. DOI: 10.1016/S0009-2614(89)87395-6.
(93) Sengupta, A.; Ramabhadran, R. O.; Raghavachari, K. Breaking a bottleneck: Accurate extrapolation to “gold standard” CCSD(T) energies for large open shell organic radicals at reduced computational cost. *J Comp Chem* **2016**, *37* (2), 286-295. DOI: 10.1002/jcc.24050.
(94) Řezáč, J.; Hobza, P. Describing Noncovalent Interactions beyond the Common Approximations: How Accurate Is the “Gold Standard,” CCSD(T) at the Complete Basis Set Limit? *J Chem Theory Comput* **2013**, *9* (5), 2151-2155. DOI: 10.1021/ct400057w.
(95) Ramabhadran, R. O.; Raghavachari, K. Extrapolation to the Gold-Standard in Quantum Chemistry: Computationally Efficient and Accurate CCSD(T) Energies for Large Molecules Using an Automated Thermochemical Hierarchy. *J Chem Theory Comput* **2013**, *9* (9), 3986-3994. DOI: 10.1021/ct400465q.
(96) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Physical Review* **1964**, *136* (3B), B864-B871. DOI: 10.1103/physrev.136.b864.

(97) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys Rev* **1965**, *140* (4A), A1133-A1138. DOI: 10.1103/physrev.140.a1133.
(98) Goerigk, L.; Mehta, N. A Trip to the Density Functional Theory Zoo: Warnings and Recommendations for the User. *Aust J Chem* **2019**, *72* (8). DOI: 10.1071/ch19023.
(99) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys Chem Chem Phys* **2017**, *19* (48), 32184-32215. DOI: 10.1039/c7cp04913g.
(100) Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings* **2001**, *577* (1), 1-20. DOI: 10.1063/1.1390175.
(101) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J Chem Phys* **2006**, *124* (3), 034108. DOI: 10.1063/1.2148954.
(102) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J Comput Chem* **2011**, *32* (7), 1456-1465. DOI: 10.1002/jcc.21759.
(103) Grimme, S. Density functional theory with London dispersion corrections. *Wires Comput Mol Sci* **2011**, *1* (2), 211-228. DOI: 10.1002/wcms.30.
(104) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010**, *132* (15), 154104. DOI: 10.1063/1.3382344.
(105) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **2005**, *7* (18), 3297-3305. DOI: 10.1039/b508541a.
(106) Jensen, F. Basis Set Convergence of Nuclear Magnetic Shielding Constants Calculated by Density Functional Methods. *J Chem Theory Comput* **2008**, *4* (5), 719-727. DOI: 10.1021/ct800013z.
(107) Jensen, F. The optimum contraction of basis sets for calculating spin-spin coupling constants. *Theor Chem Acc* **2010**, *126* (5-6), 371-382. DOI: 10.1007/s00214-009-0699-5.
(108) Barone, V. Structure, Magnetic Properties and Reactivities of Open-Shell Species From Density Functional and Self-Consistent Hybrid Methods. In *Recent Advances in Density Functional Methods*, Recent Advances in Computational Chemistry, 1995; pp 287-334.
(109) Mulliken, R. S. Electronic Population Analysis on LCAO–MO Molecular Wave Functions. I. *J Chem Phys* **1955**, *23* (10), 1833-1840. DOI: 10.1063/1.1740588.
(110) Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J Chem Phys* **1950**, *18* (3), 365-375. DOI: 10.1063/1.1747632.
(111) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theoretica Chimica Acta* **1977**, *44* (2), 129-138. DOI: 10.1007/bf00549096.
(112) Mayer, I. Charge, bond order and valence in the AB initio SCF theory. *Chem Phys Lett* **1983**, *97* (3), 270-274. DOI: 10.1016/0009-2614(83)80005-0.
(113) Wolff, S. K.; Ziegler, T. Calculation of DFT-GIAO NMR shifts with the inclusion of spin-orbit coupling. *J Chem Phys* **1998**, *109* (3), 895-905. DOI: Doi 10.1063/1.476630.
(114) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient implementation of the gauge-independent atomic orbital method for NMR chemical shift calculations. *J Am Chem Soc* **1990**, *112* (23), 8251-8260. DOI: 10.1021/ja00179a005.
(115) Ramsey, N. F. Electron Coupled Interactions between Nuclear Spins in Molecules. *Phys Rev* **1953**, *91* (2), 303-307. DOI: 10.1103/physrev.91.303.
(116) Pyykkö, P. Perspective on Norman Ramsey's theories of NMR chemical shifts and nuclear spin-spin coupling. *Theoretical Chemistry Accounts: Theory, Computation, and*

Modeling (Theoretica Chimica Acta) **2000**, *103* (3-4), 214-216. DOI: 10.1007/s002149900011.
(117) Helgaker, T.; Jaszuński, M.; Ruud, K. Ab Initio Methods for the Calculation of NMR Shielding and Indirect Spin−Spin Coupling Constants. *Chem Rev* **1999**, *99* (1), 293-352. DOI: 10.1021/cr960017t.
(118) Malkin, V. G.; Malkina, O. L.; Eriksson, L. A.; Salahub, D. R. The calculation of NMR and ESR spectroscopy parameters using density functional theory. In *Theoretical and Computational Chemistry*, Seminario, J. M., Politzer, P. Eds.; Vol. 2; Elsevier, 1995; pp 273-347.
(119) Bally, T.; Rablen, P. R. Quantum-Chemical Simulation of 1H NMR Spectra. 2. Comparison of DFT-Based Procedures for Computing Proton–Proton Coupling Constants in Organic Molecules. *J Org Chem* **2011**, *76* (12), 4818-4830. DOI: 10.1021/jo200513q.
(120) Gräfenstein, J.; Cremer, D. Analysis of the spin-dipole transmission mechanism for NMR spin–spin coupling constants using orbital contributions, spin polarization, and spin-dipole energy density distribution. *Chem Phys Lett* **2004**, *387* (4), 415-427. DOI: 10.1016/j.cplett.2004.01.120.
(121) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *J Comput Chem* **2003**, *24* (6), 669-681. DOI: 10.1002/jcc.10189.
(122) Takano, Y.; Houk, K. N. Benchmarking the Conductor-like Polarizable Continuum Model (CPCM) for Aqueous Solvation Free Energies of Neutral and Ionic Organic Molecules. *J Chem Theory Comput* **2005**, *1* (1), 70-77. DOI: 10.1021/ct049977a.
(123) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J Phys Chem B* **2009**, *113* (18), 6378-6396. DOI: 10.1021/jp810292n.
(124) Pascual-ahuir, J. L.; Silla, E.; Tuñon, I. GEPOL: An improved description of molecular surfaces. III. A new algorithm for the computation of a solvent-excluding surface. *J Comp Chem* **1994**, *15* (10), 1127-1138. DOI: 10.1002/jcc.540151009.
(125) Cancès, E.; Mennucci, B. The escaped charge problem in solvation continuum models. *J Chem Phys* **2001**, *115* (13), 6130-6135. DOI: 10.1063/1.1401157.
(126) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc* **1985**, *107* (13), 3902-3909. DOI: 10.1021/ja00299a024.
(127) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J Comp Chem* **1989**, *10* (2), 209-220. DOI: 10.1002/jcc.540100208.
(128) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Mode* **2007**, *13* (12), 1173-1213. DOI: 10.1007/s00894-007-0233-4.
(129) Seifert, G.; Porezag, D.; Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *Int J Quantum Chem* **1996**, *58* (2), 185-192. DOI: 10.1002/(sici)1097-461x(1996)58:2<185::aid-qua7>3.0.co;2-u.
(130) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys Rev B* **1995**, *51* (19), 12947-12957. DOI: 10.1103/PhysRevB.51.12947.
(131) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J Chem Theory Comput* **2017**, *13* (5), 1989-2009. DOI: 10.1021/acs.jctc.7b00118.

(132) Gaus, M.; Cui, Q.; Elstner, M. DFTB3: Extension of the self-consistent-charge density-functional tight-binding method (SCC-DFTB). *J Chem Theory Comput* **2012**, *7* (4), 931-948. DOI: 10.1021/ct100684s.
(133) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comp Chem* **2003**, *24* (16), 1999-2012. DOI: 10.1002/jcc.10349.
(134) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J Am Chem Soc* **1995**, *117* (19), 5179-5197. DOI: 10.1021/ja00124a002.
(135) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* **1998**, *102* (18), 3586-3616. DOI: 10.1021/jp973084f.
(136) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. Uff, a Full Periodic-Table Force-Field for Molecular Mechanics and Molecular-Dynamics Simulations. *J Am Chem Soc* **1992**, *114* (25), 10024-10035. DOI: DOI 10.1021/ja00051a040.
(137) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J Comp Chem* **2004**, *25* (9), 1157-1174. DOI: 10.1002/jcc.20035.
(138) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J Am Chem Soc* **1996**, *118* (45), 11225-11236. DOI: 10.1021/ja9621760.
(139) Robertson, M. J.; Tirado-Rives, J.; Jorgensen, W. L. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J Chem Theory Comput* **2015**, *11* (7), 3499-3509. DOI: 10.1021/acs.jctc.5b00356.
(140) Liakos, D. G.; Guo, Y.; Neese, F. Comprehensive Benchmark Results for the Domain Based Local Pair Natural Orbital Coupled Cluster Method (DLPNO-CCSD(T)) for Closed- and Open-Shell Systems. *J Phys Chem A* **2020**, *124* (1), 90-100. DOI: 10.1021/acs.jpca.9b05734.
(141) Neese, F. Efficient and accurate approximations to the molecular spin-orbit coupling operator and their use in molecular g-tensor calculations. *J Chem Phys* **2005**, *122* (3), 34107. DOI: 10.1063/1.1829047.
(142) Sancho-García, J. C.; Adamo, C. Double-hybrid density functionals: merging wavefunction and density approaches to get the best of both worlds. *Phys Chem Chem Phys* **2013**, *15* (35), 14581. DOI: 10.1039/c3cp50907a.
(143) Dittmer, A.; Stoychev, G. L.; Maganas, D.; Auer, A. A.; Neese, F. Computation of NMR Shielding Constants for Solids Using an Embedded Cluster Approach with DFT, Double-Hybrid DFT, and MP2. *J Chem Theory Comput* **2020**, *16* (11), 6950-6967. DOI: 10.1021/acs.jctc.0c00067.
(144) Noga, J.; Kedzuch, S.; Simunek, J.; Ten-No, S. Explicitly correlated coupled cluster F12 theory with single and double excitations. *J Chem Phys* **2008**, *128* (17), 174103. DOI: 10.1063/1.2907741.
(145) Wykes, M.; Su, N. Q.; Xu, X.; Adamo, C.; Sancho-Garcia, J. C. Double hybrid functionals and the Pi-system bond length alternation challenge: rivaling accuracy of post-HF methods. *J Chem Theory Comput* **2015**, *11* (2), 832-838. DOI: 10.1021/ct500986b.

(146) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. *Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors*; ChemRxiv, 2020. DOI: 10.26434/CHEMRXIV.12907316.V1.
(147) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent Sci* **2018**, *4* (11), 1465-1476. DOI: 10.1021/acscentsci.8b00357.
(148) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604-610. DOI: 10.1038/nature25978.
(149) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent Sci* **2019**, *5* (9), 1572-1583. DOI: 10.1021/acscentsci.9b00576.
(150) Grzybowski, B. A.; Szymkuc, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wolos, A.; Klucznik, T. Chematica: A Story of Computer Code That Started to Think like a Chemist. *Chem* **2018**, *4* (3), 390-397. DOI: 10.1016/j.chempr.2018.02.024.
(151) Stocker, S.; Csanyi, G.; Reuter, K.; Margraf, J. T. Machine learning in chemical reaction space. *Nat Commun* **2020**, *11* (1), 5505. DOI: 10.1038/s41467-020-19267-x.
(152) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach Learn-Sci Techn* **2021**, *2* (1), 15016-15016. DOI: 10.1088/2632-2153/abc81d.
(153) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Data augmentation strategies to improve reaction yield predictions and estimate uncertainty. *ChemRxiv* **2020**. DOI: 10.26434/chemrxiv.13286741.v1.
(154) Zuranski, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting Reaction Yields via Supervised Learning. *Acc Chem Res* **2021**, *54* (8), 1856-1865. DOI: 10.1021/acs.accounts.0c00770.
(155) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186-190. DOI: 10.1126/science.aar5169.
(156) 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. DOI: 10.1038/s41586-021-03213-y.
(157) Gallegos, L. C.; Luchini, G.; St John, P. C.; Kim, S.; Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc Chem Res* **2021**, *54* (4), 827-836. DOI: 10.1021/acs.accounts.0c00745.
(158) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363* (6424), eaau5631-eaau5631. DOI: 10.1126/science.aau5631.
(159) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J Am Chem Soc* **2018**, *140* (15), 5004-5008. DOI: 10.1021/jacs.8b01523.
(160) Anantpinijwatna, A.; Sales-Cruz, M.; Kim, S. H.; O'Connell, J. P.; Gani, R. A systematic modelling framework for phase transfer catalyst systems. *Chem Eng Res Des* **2016**, *115*, 407-422. DOI: 10.1016/j.cherd.2016.07.011.
(161) Henle, J. J.; Zahrt, A. F.; Rose, B. T.; Darrow, W. T.; Wang, Y.; Denmark, S. E. Development of a Computer-Guided Workflow for Catalyst Optimization. Descriptor Validation, Subset Selection, and Training Set Analysis. *J Am Chem Soc* **2020**, *142* (26), 11578-11592. DOI: 10.1021/jacs.0c04715.

(162) Bakowies, D.; von Lilienfeld, O. A. Density Functional Geometries and Zero-Point Energies in Ab Initio Thermochemical Treatments of Compounds with First-Row Atoms (H, C, N, O, F). *J Chem Theory Comput* **2021**, *17* (8), 4872-4890. DOI: 10.1021/acs.jctc.1c00474.
(163) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat Chem* **2020**, *12* (10), 945-951. DOI: 10.1038/s41557-020-0527-z.
(164) Dragoni, D.; Daff, T. D.; Csanyi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys Rev Mater* **2018**, *2* (1). DOI: ARTN 013808
10.1103/PhysRevMaterials.2.013808.
(165) Borgis, D.; Gendre, L.; Ramirez, R. Molecular density functional theory: application to solvation and electron-transfer thermodynamics in polar solvents. *J Phys Chem B* **2012**, *116* (8), 2504-2512. DOI: 10.1021/jp210817s.
(166) Wen, M.; Blau, S. M.; Spotte-Smith, E. W. C.; Dwaraknath, S.; Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem Sci* **2020**, *12* (5), 1858-1868. DOI: 10.1039/d0sc05251e.
(167) St John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat Commun* **2020**, *11* (1), 2328. DOI: 10.1038/s41467-020-16201-z.
(168) Folmsbee, D.; Koes, D. R.; Hutchison, G. Evaluation of Thermochemical Machine Learning for Potential Energy Curves and Geometry Optimization. **2020**. DOI: 10.26434/chemrxiv.13029437.v1.
(169) Mones, L.; Bernstein, N.; Csanyi, G. Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *J Chem Theory Comput* **2016**, *12* (10), 5100-5110. DOI: 10.1021/acs.jctc.6b00553.
(170) Lee, E.; Ludwig, T.; Yu, B.; Singh, A.; Gygi, F.; Nørskov, J. K.; de Pablo, J. Neural Network Sampling of the Free Energy Landscape for Nitrogen Dissociation on Ruthenium. *ChemRxiv* **2020**. DOI: 10.26434/chemrxiv.13374059.v1.
(171) Lahey, S. J.; Thien Phuc, T. N.; Rowley, C. N. Benchmarking Force Field and the ANI Neural Network Potentials for the Torsional Potential Energy Surface of Biaryl Drug Fragments. *J Chem Inf Model* **2020**, *60* (12), 6258-6268. DOI: 10.1021/acs.jcim.0c00904.
(172) Meyer, R.; Weichselbaum, M.; Hauser, A. W. Machine Learning Approaches toward Orbital-free Density Functional Theory: Simultaneous Training on the Kinetic Energy Density Functional and Its Functional Derivative. *J Chem Theory Comput* **2020**, *16* (9), 5685-5694. DOI: 10.1021/acs.jctc.0c00580.
(173) Gandolfi, M.; Rognoni, A.; Aieta, C.; Conte, R.; Ceotto, M. Machine learning for vibrational spectroscopy via divide-and-conquer semiclassical initial value representation molecular dynamics with application to N-methylacetamide. *J Chem Phys* **2020**, *153* (20), 204104. DOI: 10.1063/5.0031892.
(174) Choudhary, K.; Garrity, K. F.; Sharma, V.; Biacchi, A. J.; Walker, A. R. H.; Tavazza, F. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *Npj Comput Mater* **2020**, *6* (1), 1-13. DOI: ARTN 64
10.1038/s41524-020-0337-2.
(175) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem Sci* **2017**, *8* (10), 6924-6935. DOI: 10.1039/c7sc02267k.
(176) Kananenka, A. A.; Yao, K.; Corcelli, S. A.; Skinner, J. L. Machine Learning for Vibrational Spectroscopic Maps. *J Chem Theory Comput* **2019**, *15* (12), 6850-6858. DOI: 10.1021/acs.jctc.9b00698.

(177) Lam, J.; Abdul-Al, S.; Allouche, A. R. Combining Quantum Mechanics and Machine-Learning Calculations for Anharmonic Corrections to Vibrational Frequencies. *J Chem Theory Comput* **2020**, *16* (3), 1681-1689. DOI: 10.1021/acs.jctc.9b00964.
(178) Meiler, J.; Maier, W.; Will, M.; Meusinger, R. Using neural networks for (13)c NMR chemical shift prediction-comparison with traditional methods. *J Magn Reson* **2002**, *157* (2), 242-252. DOI: 10.1006/jmre.2002.2599.
(179) Binev, Y.; Aires-de-Sousa, J. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *J Chem Inf Comput Sci* **2004**, *44* (3), 940-945. DOI: 10.1021/ci034228s.
(180) Kwon, Y.; Lee, D.; Choi, Y. S.; Kang, M.; Kang, S. Neural Message Passing for NMR Chemical Shift Prediction. *J Chem Inf Model* **2020**, *60* (4), 2024-2030. DOI: 10.1021/acs.jcim.0c00195.
(181) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. Prediction of 1H NMR chemical shifts using neural networks. *Anal Chem* **2002**, *74* (1), 80-90. DOI: 10.1021/ac010737m.
(182) Jonas, E.; Kuhn, S. Rapid prediction of NMR spectral properties with quantified uncertainty. *J Cheminform* **2019**, *11* (1), 50. DOI: 10.1186/s13321-019-0374-3.
(183) Martinez-Trevino, S. H.; Uc-Cetina, V.; Fernandez-Herrera, M. A.; Merino, G. Prediction of Natural Product Classes Using Machine Learning and (13)C NMR Spectroscopic Data. *J Chem Inf Model* **2020**, *60* (7), 3376-3386. DOI: 10.1021/acs.jcim.0c00293.
(184) Gupta, A.; Chakraborty, S.; Ramakrishnan, R. Revving up 13C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach Learn-Sci Techn* **2021**, *2* (3), 035010-035010. DOI: 10.1088/2632-2153/abe347.
(185) Gerrard, W.; Bratholm, L. A.; Packer, M. J.; Mulholland, A. J.; Glowacki, D. R.; Butts, C. P. IMPRESSION - prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem Sci* **2020**, *11* (2), 508-515. DOI: 10.1039/c9sc03854j.
(186) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571* (7765), 343-348. DOI: 10.1038/s41586-019-1384-z.
(187) Hansch, C.; Kurup, A.; Garg, R.; Gao, H. Chem-Bioinformatics and QSAR: A Review of QSAR Lacking Positive Hydrophobic Terms. *Chem Rev* **2001**, *101* (3), 619-672. DOI: 10.1021/cr0000067.
(188) Belfield, S. J.; Firman, J. W.; Enoch, S. J.; Madden, J. C.; Erik Tollefsen, K.; Cronin, M. T. D. A review of quantitative structure-activity relationship modelling approaches to predict the toxicity of mixtures. *CompTox* **2023**, *25*, 100251. DOI: 10.1016/j.comtox.2022.100251.
(189) Kwon, S.; Bae, H.; Jo, J.; Yoon, S. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics* **2019**, *20* (1). DOI: 10.1186/s12859-019-3135-4.
(190) Matsumoto, K.; Miyao, T.; Funatsu, K. Ranking-Oriented Quantitative Structure–Activity Relationship Modeling Combined with Assay-Wise Data Integration. *ACS Omega* **2021**, *6* (18), 11964-11973. DOI: 10.1021/acsomega.1c00463.
(191) Rosell-Hidalgo, A.; Young, L.; Moore, A. L.; Ghafourian, T. QSAR and molecular docking for the search of AOX inhibitors: a rational drug discovery approach. *J Comput Aided Mol Des* **2021**, *35* (2), 245-260. DOI: 10.1007/s10822-020-00360-8.
(192) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys Rev B* **2019**, *99* (1), 014104. DOI: 10.1103/PhysRevB.99.014104.
(193) Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Modeling & Simulation* **2016**, *14* (3), 1153-1173. DOI: 10.1137/15m1054183.

(194) Gaussian approximation potentials: A brief tutorial introduction. John Wiley and Sons Inc.: 2015; Vol. 115, pp 1051-1057.
(195) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **2010**, *104* (13), 136403. DOI: 10.1103/PhysRevLett.104.136403.
(196) Li, S.; Xi, L.; Wang, C.; Li, J.; Lei, B.; Liu, H.; Yao, X. A novel method for protein-ligand binding affinity prediction and the related descriptors exploration. *J Comp Chem* **2009**, *30* (6), 900-909. DOI: 10.1002/jcc.21078.
(197) Ain, Q. U.; Aleksandrova, A.; Roessler, F. D.; Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip Rev Comput Mol Sci* **2015**, *5* (6), 405-424. DOI: 10.1002/wcms.1225.
(198) Born, J.; Huynh, T.; Stroobants, A.; Cornell, W. D.; Manica, M. Active Site Sequence Representations of Human Kinases Outperform Full Sequence Representations for Affinity Prediction and Inhibitor Generation: 3D Effects in a 1D Model. *J Chem Inf Model* **2022**, *62* (2), 240-257. DOI: 10.1021/acs.jcim.1c00889.
(199) Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Substituting random forest for multiple linear regression improves binding affinity prediction of scoring functions: Cyscore as a case study. *BMC Bioinformatics* **2014**, *15* (1), 291. DOI: 10.1186/1471-2105-15-291.
(200) Zubatyuk, R.; Smith, J. S.; Nebgen, B. T.; Tretiak, S.; Isayev, O. Teaching a neural network to attach and detach electrons from molecules. *Nat Commun* **2021**, *12* (1), 4870. DOI: 10.1038/s41467-021-24904-0.
(201) Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J Chem Theory Comput* **2019**, *15* (6), 3678-3693. DOI: 10.1021/acs.jctc.9b00181.
(202) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *J Chem Inf Model* **2020**, *60* (7), 3408-3415. DOI: 10.1021/acs.jcim.0c00451.
(203) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J Chem Theory Comput* **2020**, *16* (7), 4192-4202. DOI: 10.1021/acs.jctc.0c00121.
(204) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **2019**, *10* (1), 2903. DOI: 10.1038/s41467-019-10827-4.
(205) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci* **2017**, *8* (4), 3192-3203. DOI: 10.1039/c6sc05720a.
(206) Kang, S.; Kwon, Y.; Lee, D.; Choi, Y. S. Predictive Modeling of NMR Chemical Shifts without Using Atomic-Level Annotations. *J Chem Inf Model* **2020**, *60* (8), 3765-3769. DOI: 10.1021/acs.jcim.0c00494.
(207) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc Chem Res* **2021**, *54* (2), 263-270. DOI: 10.1021/acs.accounts.0c00699.
(208) Karimi, M.; Wu, D.; Wang, Z.; Shen, Y. Explainable Deep Relational Networks for Predicting Compound-Protein Affinities and Contacts. *J Chem Inf Model* **2021**, *61* (1), 46-66. DOI: 10.1021/acs.jcim.0c00866.

(209) Wang, M. W. H.; Goodman, J. M.; Allen, T. E. H. Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models. *Chem Res Toxicol* **2021**, *34* (2), 217-239. DOI: 10.1021/acs.chemrestox.0c00316.
(210) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions Using Graph Convolutional Networks. *J Chem Inf Model* **2019**, *59* (12), 5026-5033. DOI: 10.1021/acs.jcim.9b00538.
(211) Zhong, W.; Yang, Z.; Chen, C. Y.-C. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nat Com* **2023**, *14* (1). DOI: 10.1038/s41467-023-38851-5.
(212) Liu, C.-H.; Korablyov, M.; Jastrzębski, S.; Włodarczyk-Pruszyński, P.; Bengio, Y.; Segler, M. RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software. *J Chem Inf Model* **2022**, *62* (10), 2293-2300. DOI: 10.1021/acs.jcim.1c01476.
(213) Stuyver, T.; Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *J Chem Phys* **2022**, *156* (8). DOI: 10.1063/5.0079574.
(214) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J Chem Inf Model* **2019**, *59* (3), 1253-1268. DOI: 10.1021/acs.jcim.8b00785.
(215) Yang, Z.; Chakraborty, M.; White, A. D. Predicting chemical shifts with graph neural networks. *Chem Sci* **2021**, *12* (32), 10802-10809. DOI: 10.1039/d1sc01895g.
(216) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv Neur In* **2017**, *30*.
(217) Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
(218) Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. Language models are few-shot learners. *Adv Neur In* **2020**, *33*, 1877-1901.
(219) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J. L. Mapping the space of chemical reactions using attention-based neural networks. *Nat Mach Intell* **2021**, *3* (2), 144-152. DOI: 10.1038/s42256-020-00284-w.
(220) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit Discov* **2022**, *1* (2), 91-97. DOI: 10.1039/d1dd00006c.
(221) M. Bran, A.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nat Mach Intell* **2024**, *6* (5), 525-535. DOI: 10.1038/s42256-024-00832-8.
(222) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem Sci* **2020**, *11* (12), 3316-3325. DOI: 10.1039/c9sc05704h.
(223) Zhang, J.; Mercado, R.; Engkvist, O.; Chen, H. Comparative Study of Deep Generative Models on Chemical Space Coverage. *J Chem Inf Model* **2021**, *61* (6), 2572-2581. DOI: 10.1021/acs.jcim.0c01328.
(224) He, J.; You, H.; Sandström, E.; Nittinger, E.; Bjerrum, E. J.; Tyrchan, C.; Czechtizky, W.; Engkvist, O. Molecular optimization by capturing chemist's intuition using deep neural networks. *J Cheminfom* **2021**, *13* (1). DOI: 10.1186/s13321-021-00497-0.

(225) Arus-Pous, J.; Patronov, A.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J. L.; Chen, H.; Engkvist, O. SMILES-based deep generative scaffold decorator for de-novo drug design. *J Cheminform* **2020**, *12* (1), 38. DOI: 10.1186/s13321-020-00441-8.
(226) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J.-L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *J Cheminfom* **2019**, *11* (1). DOI: 10.1186/s13321-019-0341-z.
(227) Bagal, V.; Aggarwal, R.; Vinod, P. K.; Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J Chem Inf Model* **2022**, *62* (9), 2064-2076. DOI: 10.1021/acs.jcim.1c00600.
(228) Fu, N.; Wei, L.; Song, Y.; Li, Q.; Xin, R.; Omee, S. S.; Dong, R.; Siriwardane, E. M. D.; Hu, J. Material transformers: deep learning language models for generative materials design. *Mach Learn-Sci Techn* **2023**, *4* (1), 015001.
(229) Kim, H.; Lee, J.; Ahn, S.; Lee, J. R. A merged molecular representation learning for molecular properties prediction with a web-based service. *Sci Rep* **2021**, *11* (1). DOI: 10.1038/s41598-021-90259-7.
(230) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nat Mach Intell* **2022**, *4* (12), 1256-1264. DOI: 10.1038/s42256-022-00580-7.
(231) Wen, N.; Liu, G.; Zhang, J.; Zhang, R.; Fu, Y.; Han, X. A fingerprints based molecular property prediction method using the BERT model. *J Cheminfom* **2022**, *14* (1). DOI: 10.1186/s13321-022-00650-3.
(232) Yu, J.; Zhang, C.; Cheng, Y.; Yang, Y.-F.; She, Y.-B.; Liu, F.; Su, W.; Su, A. SolvBERT for solvation free energy and solubility prediction: a demonstration of an NLP model for predicting the properties of molecular complexes. *Digit Discov* **2023**, *2* (2), 409-421. DOI: 10.1039/d2dd00107a.
(233) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Laino, T.; Reymond, J.-L. Data-Driven Chemical Reaction Classification, Fingerprinting and Clustering using Attention-Based Neural Networks Fingerprinting and Clustering using Attention-Based Neural Networks. *ChemRxiv* **2019**. DOI: 10.26434/chemrxiv.9897365.v2.
(234) Janet, J. P.; Tomberg, A.; Boström, J. Reusability report: Learning the language of synthetic methods used in medicinal chemistry. *Nat Mach Intell* **2021**, *3* (7), 572-575. DOI: 10.1038/s42256-021-00367-2.
(235) Santiago, C. B.; Guo, J. Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem Sci* **2018**, *9* (9), 2398-2412. DOI: 10.1039/c7sc04679k.
(236) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat Chem* **2012**, *4* (5), 366-374. DOI: 10.1038/nchem.1297.
(237) Neel, A. J.; Hilton, M. J.; Sigman, M. S.; Toste, F. D. Exploiting non-covalent π interactions for catalyst design. *Nature* **2017**, *543* (7647), 637-646. DOI: 10.1038/nature21701.
(238) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A. A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. **2021**. DOI: 10.26434/chemrxiv.12996665.v1.
(239) Miro, J.; Gensch, T.; Ellwart, M.; Han, S. J.; Lin, H. H.; Sigman, M. S.; Toste, F. D. Enantioselective Allenoate-Claisen Rearrangement Using Chiral Phosphate Catalysts. *J Am Chem Soc* **2020**, *142* (13), 6390-6399. DOI: 10.1021/jacs.0c01637.

(240) Levin, M. D.; Ovian, J. M.; Read, J. A.; Sigman, M. S.; Jacobsen, E. N. Catalytic Enantioselective Synthesis of Difluorinated Alkyl Bromides. *J Am Chem Soc* **2020**, *142* (35), 14831-14837. DOI: 10.1021/jacs.0c07043.
(241) Werth, J.; Sigman, M. S. Connecting and Analyzing Enantioselective Bifunctional Hydrogen Bond Donor Catalysis Using Data Science Tools. *J Am Chem Soc* **2020**, *142* (38), 16382-16391. DOI: 10.1021/jacs.0c06905.
(242) Rinehart, N. I.; Zahrt, A. F.; Henle, J. J.; Denmark, S. E. Dreams, False Starts, Dead Ends, and Redemption: A Chronicle of the Evolution of a Chemoinformatic Workflow for the Optimization of Enantioselective Catalysts. *Acc Chem Res* **2021**, *54* (9), 2041-2054. DOI: 10.1021/acs.accounts.0c00826.
(243) Denmark, S.; Zahrt, A.; Darrow, W.; Rose, B.; Henle, J. Computational Methods for Training Set Selection and Error Assessment Applied to Catalyst Design: Guidelines for Deciding Which Reactions to Run First and Which to Run Next. *ChemRxiv* **2020**. DOI: 10.26434/chemrxiv.13239422.v1.
(244) Han, H.; Choi, S. Transfer Learning from Simulation to Experimental Data: NMR Chemical Shift Predictions. *J Phys Chem Lett* **2021**, *12* (14), 3662-3668. DOI: 10.1021/acs.jpclett.1c00578.
(245) Guan, Y.; Sowndarya Sv, S.; Gallegos, L.; St. John, P.; Paton, R. Real-time Prediction of 1H and 13C Chemical Shifts with DFT accuracy using a 3D Graph Neural Network. **2021**. DOI: 10.26434/chemrxiv-2021-zmp0l.
(246) *RDKit: Open-source cheminformatics*. https://www.rdkit.org.
(247) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J Chem Phys* **2018**, *148* (24), 241722. DOI: 10.1063/1.5019779.
(248) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* **2017**, *8* (4), 3192-3203. DOI: 10.1039/c6sc05720a.
(249) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* **1988**, *28* (1), 31-36. DOI: 10.1021/ci00057a005.
(250) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J Cheminfom* **2018**, *10* (1). DOI: 10.1186/s13321-018-0258-y.
(251) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J Chem Doc* **1965**, *5* (2), 107-113. DOI: 10.1021/c160017a018.
(252) TT, T. An Elementary Mathematical theory of Classification and Prediction. *Internal IBM Technical Report* **1958**.
(253) Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société vaudoise des sciences naturelles* **1901**, *37*, 547–579.
(254) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323* (6088), 533-536. DOI: DOI 10.1038/323533a0.
(255) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* **2014**.
(256) Hinton, G.; Srivastava, N.; Swersky, K. *Neural networks for machine learning lecture 6a overview of mini-batch gradient descent*. 2012. https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
(257) Duchi, J.; Hazan, E.; Singer, Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J Mach Learn Res* **2011**, *12* (61), 2121-2159.
(258) Vaswani, A.; Brain, G.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. *Attention Is All You Need*.

(259) Byrne, R. M. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI*, 2019; California, CA: pp 6276-6282.
(260) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model agnostic generation of counterfactual explanations for molecules. *ChemRxiv* **2021**. DOI: 10.26434/chemrxiv-2021-4qkg8-v2.
(261) Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; Swami, A. Practical Black-Box Attacks against Machine Learning. *ASIA CCS 2017 - Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security* **2016**, 506-519.
(262) Huang, S.; Papernot, N.; Goodfellow, I.; Duan, Y.; Abbeel, P. Adversarial Attacks on Neural Network Policies. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings* **2017**.
(263) Parekh, S.; Kumar Singla, Y.; Chen, C.; Jessy Li, J.; Ratn Shah, R. My Teacher Thinks The World Is Flat! Interpreting Automatic Essay Scoring Mechanism. *arXiv* **2020**. DOI: 10.48550/arXiv.2012.13872.
(264) Lundberg, S. M.; Lee, S. I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30 (Nips 2017)* **2017**, *30*, 4766-4775.
(265) Ribeiro, M. T.; Singh, S.; Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *Thirty-Second Aaai Conference on Artificial Intelligence / Thirtieth Innovative Applications of Artificial Intelligence Conference / Eighth Aaai Symposium on Educational Advances in Artificial Intelligence* **2018**, *32* (1), 1527-1535.
(266) Sundararajan, M.; Taly, A.; Yan, Q. Q. Axiomatic Attribution for Deep Networks. *Pr Mach Learn Res* **2017**, *70*.
(267) Lu, Y.; Murzakhanov, I.; Chatzivasileiadis, S. Neural network interpretability for forecasting of aggregated renewable generation. In *2021 IEEE International conference on communications, control, and computing technologies for Smart Grids (SmartGridComm)*, 2021; IEEE: pp 282-288.
(268) Janizek, J. D.; Sturmfels, P.; Lee, S. I. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *J Mach Learn Res* **2021**, *22*, 1-54.
(269) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, *13-17-August-2016*, 1135-1144. DOI: 10.1145/2939672.2939778.
(270) Neese, F. Software update: the ORCA program system, version 4.0. *Wires Comput Mol Sci* **2018**, *8* (1), 1-6. DOI: ARTN e1327
10.1002/wcms.1327.
(271) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew Chem Int Ed* **2021**, *60* (8), 4266-4274. DOI: 10.1002/anie.202011941.
(272) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors: The Case of Sterimol Steric Parameters. *Acs Catal* **2019**, *9* (3), 2313-2323. DOI: 10.1021/acscatal.8b04043.
(273) Zheng, J. J.; Xu, X. F.; Truhlar, D. G. Minimally augmented Karlsruhe basis sets. *Theor Chem Acc* **2011**, *128* (3), 295-305. DOI: 10.1007/s00214-010-0846-z.
(274) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J Chem Inf Model* **2015**, *55* (2), 460-473. DOI: 10.1021/ci500588j.
(275) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J Chem Theory Comput* **2019**, *15* (5), 2847-2862. DOI: 10.1021/acs.jctc.9b00143.

(276) Shepherd, E. D.; Dyson, B. S.; Hak, W. E.; Nguyen, Q. N. N.; Lee, M.; Kim, M. J.; Sohn, T. I.; Kim, D.; Burton, J. W.; Paton, R. S. Structure Determination of a Chloroenyne from Laurencia majuscula Using Computational Methods and Total Synthesis. *J Org Chem* **2019**, *84* (9), 4971-4991. DOI: 10.1021/acs.joc.8b02975.
(277) R, V. V.; Ducati, L. C.; Tormena, C. F.; Autschbach, J. The halogen effect on the (13)C NMR chemical shift in substituted benzenes. *Phys Chem Chem Phys* **2018**, *20* (16), 11247-11259. DOI: 10.1039/c8cp01249k.
(278) Kaupp, M.; Malkina, O. L.; Malkin, V. G.; Pyykkö, P. How Do Spin-Orbit-Induced Heavy-Atom Effects on NMR Chemical Shifts Function? Validation of a Simple Analogy to Spin-Spin Coupling by Density Functional Theory (DFT) Calculations on Some Iodo Compounds. *Chem Eur J* **1998**, *4* (1), 118-126. DOI: 10.1002/(sici)1521-3765(199801)4:1<118::Aid-chem118>3.0.Co;2-6.
(279) Vicha, J.; Svec, P.; Ruzickova, Z.; Samsonov, M. A.; Bartova, K.; Ruzicka, A.; Straka, M.; Dracinsky, M. Experimental and Theoretical Evidence of Spin-Orbit Heavy Atom on the Light Atom (1) H NMR Chemical Shifts Induced through HI(-) Hydrogen Bond. *Chemistry* **2020**, *26* (40), 8698-8702. DOI: 10.1002/chem.202001532.
(280) Vicha, J.; Komorovsky, S.; Repisky, M.; Marek, R.; Straka, M. Relativistic Spin-Orbit Heavy Atom on the Light Atom NMR Chemical Shifts: General Trends Across the Periodic Table Explained. *J Chem Theory Comput* **2018**, *14* (6), 3025-3039. DOI: 10.1021/acs.jctc.8b00144.
(281) Berger, R. J.; Repisky, M.; Komorovsky, S. How does relativity affect magnetically induced currents? *Chem Commun (Camb)* **2015**, *51* (73), 13961-13963. DOI: 10.1039/c5cc05732a.
(282) Di Remigio, R.; Steindal, A. H.; Mozgawa, K.; Weijo, V.; Cao, H.; Frediani, L. PCMSolver: An open-source library for solvation modeling. *Int J Quantum Chem* **2019**, *119* (1), e25685. DOI: 10.1002/qua.25685.
(283) Ascough, D. M. H. Novel fluorination methodology through computational insight : tuning fluoride through hydrogen bonding. University of Oxford, 2019.
(284) Del Bene, J. E.; Bartlett, R. J.; Elguero, J. Interpreting 2J(F,N),1J(H,N) and 1J(F,H) in the hydrogen-bonded FH-collidine complex. *Magn Reson Chem* **2002**, *40* (12), 767-771. DOI: 10.1002/mrc.1103.
(285) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput Mol Sci* **2022**. DOI: 10.1002/wcms.1606.
(286) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree-Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree-Fock exchange. *Chem Phys* **2009**, *356* (1-3), 98-109. DOI: 10.1016/j.chemphys.2008.10.036.
(287) Neese, F. An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. *J Comput Chem* **2003**, *24* (14), 1740-1747. DOI: 10.1002/jcc.10318.
(288) Stoychev, G. L.; Auer, A. A.; Neese, F. Automatic Generation of Auxiliary Basis Sets. *J Chem Theory Comput* **2017**, *13* (2), 554-562. DOI: 10.1021/acs.jctc.6b01041.
(289) Weigend, F. A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys Chem Chem Phys* **2002**, *4* (18), 4285-4291. DOI: 10.1039/b204199p.
(290) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J Phys Chem A* **1998**, *102* (11), 1995-2001. DOI: 10.1021/jp9716997.
(291) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum mechanical continuum solvation models. *Chem Rev* **2005**, *105* (8), 2999-3093. DOI: 10.1021/cr9904009.

(292) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. Energies, structures, and electronic properties of molecules in solution with the C-PCM solvation model. *Journal of Computational Chemistry* **2003**, *24* (6), 669-681. DOI: 10.1002/jcc.10189.
(293) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011**, *32* (7), 1456-1465. DOI: 10.1CHap002/jcc.21759.
(294) Ditchfield, R. Self-Consistent Perturbation-Theory of Diamagnetism .1. Gauge-Invariant Lcao Method for Nmr Chemical-Shifts. *Mol Phys* **1974**, *27* (4), 789-807. DOI: Doi 10.1080/00268977400100711.
(295) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys Chem Chem Phys* **2020**, *22* (14), 7169-7192. DOI: 10.1039/c9cp06869d.
(296) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A Robust Non-Self-Consistent Tight-Binding Quantum Chemistry Method for large Molecules. *ChemRxiv* **2019**. DOI: 10.26434/chemrxiv.8326202.v1.
(297) Lindahl; Abraham; Hess; Spoel, v. d. GROMACS 2020 Manual. **2020**. DOI: 10.5281/ZENODO.3562512.
(298) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1-2*, 19-25. DOI: 10.1016/j.softx.2015.06.001.
(299) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J Chem Phys* **1995**, *103* (19), 8577-8593. DOI: 10.1063/1.470117.
(300) Hess, B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput* **2008**, *4* (1), 116-122. DOI: 10.1021/ct700200b.
(301) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **2007**, *126* (1), 014101. DOI: 10.1063/1.2408420.
(302) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **1981**, *52* (12), 7182-7190. DOI: 10.1063/1.328693.
(303) Nosé, S.; Klein, M. L. Constant pressure molecular dynamics for molecular systems. *Mol Phys* **1983**, *50* (5), 1055-1076. DOI: 10.1080/00268978300102851.
(304) Daura, X.; Gademann, K.; Jaun, B.; Seebach, D.; Van Gunsteren, W. F.; Mark, A. E. Peptide Folding: When Simulation Meets Experiment. *Angew Chem Int Ed* **1999**, *38* (1-2), 236-240. DOI: 10.1002/(sici)1521-3773(19990115)38:1/2<236::aid-anie236>3.0.co;2-m.
(305) Kreutter, D.; Schwaller, P.; Reymond, J. L. Predicting enzymatic reactions with a molecular transformer. *Chem Sci* **2021**, *12* (25), 8648-8659. DOI: 10.1039/d1sc02362d.
(306) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building blocks for automated elucidation of metabolites: machine learning methods for NMR prediction. *BMC Bioinformatics* **2008**, *9* (1), 400. DOI: 10.1186/1471-2105-9-400.
(307) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019; pp 4171-4186.
(308) Liu, W.; Groves, J. T. Manganese-catalyzed oxidative benzylic C-H fluorination by fluoride ions. *Angew Chem Int Ed Engl* **2013**, *52* (23), 6024-6027. DOI: 10.1002/anie.201301097.
(309) Berger, F.; Plutschack, M. B.; Riegger, J.; Yu, W.; Speicher, S.; Ho, M.; Frank, N.; Ritter, T. Site-selective and versatile aromatic C-H functionalization by thianthrenation. *Nature* **2019**, *567* (7747), 223-228. DOI: 10.1038/s41586-019-0982-0.

(310) Yamamoto, K.; Li, J.; Garber, J. A. O.; Rolfes, J. D.; Boursalian, G. B.; Borghs, J. C.; Genicot, C.; Jacq, J.; van Gastel, M.; Neese, F.; Ritter, T. Palladium-catalysed electrophilic aromatic C-H fluorination. *Nature* **2018**, *554* (7693), 511-514. DOI: 10.1038/nature25749.
(311) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *J Condens Matter Phys* **2017**, *29* (27), 273002. DOI: 10.1088/1361-648x/aa680e.
(312) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput Phys Commun* **2020**, *247*, 106949. DOI: 10.1016/j.cpc.2019.106949.
(313) Kovacs, D. P.; McCorkindale, W.; Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat Commun* **2021**, *12* (1), 1695. DOI: 10.1038/s41467-021-21895-w.
(314) Lu, Y.; Murzakhanov, I.; Chatzivasileiadis, S. Neural network interpretability for forecasting of aggregated renewable generation. *arXiv* **2021**. DOI: 10.48550/arXiv.2106.10476.

Appendix 1 Hypermeters for ML Screening Script

Lasso: Default,

Fine Tree : min_samples_leaf=4

Medium Tree : min_samples_leaf=12

Coarse Tree : min_samples_leaf=36

SVM Linear : kernel='linear'

SVM Quadratic : kernel='poly',degree=2

SVM Cubic : kernel='poly',degree=3

SVM Fine Gaussian : kernel='rbf', gamma=0.25

SVM Medium Gaussian : kernel='rbf',gamma=1

SVM Coarse Gaussian : kernel='rbf',gamma=4

Boosted Trees : min_samples_leaf=8

Bagged Trees : n_estimators=30

Random Forest: Default

KNN Fine : n_neighbors=1,weights='uniform',metric='euclidean'

KNN Medium : n_neighbors=10,weights='uniform',metric='euclidean'

KNN Coarse : n_neighbors=100,weights='uniform',metric='euclidean'

KNN Cubic : n_neighbors=10,weights='uniform',metric='minkowski',p=3

KNN weighted : n_neighbors=10,weights='distance',metric='euclidean'

GPR RBF : kernel=RBF(),n_restarts_optimizer=50,normalize_y=True

GPR Matern 3/2 : kernel=Matern(),n_restarts_optimizer=50,normalize_y=True

GPR Matern 5/2 : kernel=Matern(nu=2.5),n_restarts_optimizer=50,normalize_y=True

GPR Exponential : kernel=Matern(nu=1.2),n_restarts_optimizer=50,normalize_y=True

GPR Rational Quadratic: kernel=RationalQuadratic(),n_restarts_optimizer=50,normalize_y=True

Appendix 2 Data from Chapter 3

	Sigman			*unSterimol*			*graphSterimol*		
Group	L	B1	B5	L	B1	B5	L	B1	B5
Me	2.87	1.52	2.04	3.03	1.94	6.82	3.03	1.50	2.03
Et	4.11	1.52	3.17	4.18	2.24	6.80	4.18	1.53	3.15
Ph	6.28	1.71	3.11	6.38	2.55	6.87	6.38	1.70	3.21
Bn	4.62	1.52	6.02	4.64	2.90	6.74	4.64	1.52	6.08
iPr	4.11	1.9	3.17	4.16	2.73	6.78	4.16	1.89	3.16
tBu	4.11	2.6	3.17	4.27	2.76	6.65	4.27	2.73	3.18
Cy	6.17	1.91	3.49	6.24	3.13	6.78	6.24	1.90	3.46
CH2tBu	4.89	1.52	4.18	5.29	3.24	6.72	5.29	1.54	4.36
CHEt2	4.72	2.13	4.01	5.07	3.12	6.77	5.07	1.91	4.44
CH2iPr	4.92	1.52	4.45	5.17	2.90	6.81	5.17	1.53	4.36
CHPh2	5.15	2.01	6.02	6.10	4.40	6.77	6.10	1.97	6.12

Ad	6.17	3.16	3.49	6.30	3.31	6.66	6.30	3.09	3.62

Table 8 Sterimol Values for the NHK Reaction

	Sigman			**unSterimol**			**graphSterimol**		
Group	**L**	**B1**	**B5**	**L**	**B1**	**B5**	**L**	**B1**	**B5**
Me	**2.87**	**1.52**	**2.04**	**6.24**	**1.93**	**10.06**	**2.92**	**1.50**	**2.05**
Et	**4.11**	**1.52**	**3.17**	**6.18**	**2.60**	**10.10**	**3.99**	**1.54**	**3.15**
iPr	**4.11**	**1.9**	**3.17**	**5.76**	**2.73**	**10.21**	**4.04**	**1.90**	**3.16**
tBu	**4.11**	**2.6**	**3.17**	**5.39**	**2.75**	**10.34**	**4.07**	**2.73**	**3.16**
$CHPr_2$	**6.17**	**1.9**	**4.54**	**5.71**	**2.70**	**10.31**	**4.20**	**2.06**	**5.69**
CEt_3	**4.92**	**2.94**	**4.18**	**5.20**	**3.32**	**10.35**	**4.76**	**2.76**	**4.48**
$CHiPr_2$	**4.12**	**2.08**	**4.19**	**5.64**	**3.13**	**10.44**	**4.93**	**2.15**	**4.47**
Ad	**6.17**	**3.16**	**3.49**	**6.59**	**3.14**	**14.61**	**6.19**	**3.09**	**3.67**

Below are the SMILES strings for each of the Substrates and Catalysts using in the Training data along with the assigned ID number

Substrate	*Smiles*
1	Br[C@H]1[C@H](N(CC2=CC=CC=C2)CC3=CC=CC=C3)CCCC1
2	Br[C@H]1[C@H](N(CC2=CC=CC=C2)CC3=CC=CC=C3)CCCC1
3	Cl[C@H](C1=CC=CC=C1)[C@H](N(CC2=CC=CC=C2)CC3=CC=CC=C3)C4=CC=CC=C4
4	Br[C@H]1[C@H](N(CC2=CC=C(OC)C=C2)CC3=CC=C(OC)C=C3)CCCC1
5	Cl[C@H](C1=CC=CC=C1)[C@H](N(CC2=CC=C(OC)C=C2)CC3=CC=C(OC)C=C3)C4=CC=CC=C4
6	Cl[C@H](C1=CC=CC=C1)[C@H](N(CC=C)CC=C)C2=CC=CC=C2
7	Cl[C@H](C1=CC=CC=C1)[C@H](N(C)C)C2=CC=CC=C2
8	Br[C@H]1[C@H](SCCC2=CC=CC=C2)CCCC1
9	Br[C@H]1[C@H](SCCC2=CC=CC=C2)CCC1
10	Br[C@H](C1=CC=CC=C1)[C@H](SCCC2=CC=CC=C2)C3=CC=CC=C3
11	Br[C@H]1[C@@H](CC(C=CC=C2)=C2C1)SCCC3=CC=CC=C3
12	Br[C@H]1[C@H](SCCC2=CC=CC=C2)COC1
13	Br[C@H]1[C@H](SCCC2=CC=CC=C2)CN(C(OC(C)(C)C)=O)C1

14	Br[C@H](COCC1=CC=CC=C1)[C@@H](COCC2=CC=CC=C2)SCCC3=CC=CC=C3
15	Br[C@H](C1=CC=C(Cl)C=C1)[C@H](SCCC2=CC=CC=C2)C3=CC=C(Cl)C=C3
16	Br[C@H](C1=CC=C(F)C=C1)[C@H](SCCC2=CC=CC=C2)C3=CC=C(F)C=C3
17	Br[C@H]1[C@@H](CCCCC1)SCCC2=CC=CC=C2
18	Br[C@H](C1=CC=C(C)C=C1)[C@H](SCCC2=CC=CC=C2)C3=CC=C(C)C=C3
19	Br[C@H](C1=CC=CC=C1)[C@H](SC)C2=CC=CC=C2
20	Br[C@H](C1=CC=CC=C1)[C@H](SC2=CC=CC=C2)C3=CC=CC=C3
21	Br[C@H](C1=CC=CC=C1)[C@H](SC2CCCCC2)C3=CC=CC=C3
22	Br[C@H](C1=CC=CC=C1)[C@H](SCC2=CC=CC=C2)C3=CC=CC=C3
23	Br[C@H](C1=CC=CC=C1)[C@H](SCCCCCC)C2=CC=CC=C2
24	Br[C@H](C1=CC=CC=C1)[C@H](SC2=CC=C(OC)C=C2)C3=CC=CC=C3
25	Br[C@H](C1=CC=CC=C1)[C@H](SC2=CC=C(C(C)(C)C)C=C2)C3=CC=CC=C3
26	Br[C@H](C1=CC=CC=C1)[C@H](SCC)C2=CC=CC=C2
27	Br[C@@H]([C@H](SC1=CC=CS1)C2=CC=CC=C2)C3=CC=CC=C3
28	Br[C@@H](C1=CC=C(C=CC=C2)C2=C1)[C@@H](SCCC3=CC=CC=C3)C4=CC(C=CC=C5)=C5C=C4

29	Br[C@@H]([C@H](SC1=CC=CC=N1)C2=CC=CC=C2)C3=CC=CC=C3
30	Br[C@@H]([C@H](SC1=NC2=C(C=CC=C2)S1)C3=CC=CC=C3)C4=CC=CC=C4
31	Br[C@H](C1=CC=CC(Cl)=C1)[C@H](SCCC2=CC=CC=C2)C3=CC(Cl)=CC=C3
32	Br[C@H](C1=CC=CC(C)=C1)[C@H](SCCC2=CC=CC=C2)C3=CC(C)=CC=C3
33	Br[C@H](C1=CC=CC(F)=C1)[C@H](SCCC2=CC=CC=C2)C3=CC(F)=CC=C3
34	Br[C@H](C1=CC=CC(OC)=C1)[C@H](SCCC2=CC=CC=C2)C3=CC(OC)=CC=C3

Catalysts	*Smiles*
1	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2
2	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=CC=C6
3	O=C(N[C@H]1[C@H](NC(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=O)CCCC1)NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3
4	O[C@H]1CC2=CC=CC=C2[C@H]1NC(NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=O
5	O=C(NC1=CC=CC=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC=CC=C5)=O)C6=C2C=CC=C6

6	O=C(NC1=CC=C([N+]([O-])=O)C=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC=C([N+]([O-])=O)C=C5)=O)C6=C2C=CC=C6
7	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@@]([C@@]3=C(NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O)C=CC5=C3CCCC5)C6=C2CCCC6
8	OC1=[C@]([C@]2=C(C=CC=C3)C3=CC=C2NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O)C5=CC=CC=C5C=C1
9	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@H](C2=CC=CC=C2)[C@@H](C3=CC=CC=C3)NS(=O)(C[C@]4(C5(C)CCCCC5CC4=O)=O
10	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@H](C2=CC=CC=C2OCCCC)[C@@H](C3=CC=CC=C3OCCCC)NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O
11	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@@H](C2=CC=CC=C2O)[C@H](C3=CC=CC=C3O)NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O
12	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(Br)=C2)=[C@]([C@]3=C(CCCC4)C4=CC(Br)=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2CCCC6
13	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(I)=C2)=[C@]([C@]3=C(CCCC4)C4=CC(I)=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2CCCC6

14	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C=CC=C3)C3=C2)=C4)=[C@]([C@]5=C(CCCC6)C6=CC(C7=CC=C(C=CC=C8)C8=C7)=C5NC(NC9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=O)C%10=C4CCCC%10
15	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=CC=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=CC=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
16	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=CC=C2C)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=CC=C6C)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
17	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(C)C=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
18	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
19	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C(F)(F)F)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(C(F)(F)F)C=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
20	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C3=CC=CC=C3)C=C2)=C4)=[C@]([C@]5=C(CCCC6)C6=CC(C7=CC=C(C8=CC=CC=C8)C=C7)=C5NC(NC9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=O)C%10=C4CCCC%10

21	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C(C)(C)C)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(C(C)(C)C)C=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
22	O=C(NC1CCCCC1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5CCCCC5)=O)C6=C2C=CC=C6
23	O=C(NC1=CC=CC2=C1C=CC=C2)NC(C=C3)=[C@]([C@]4=C(C=CC=C5)C5=CC=C4NC(NC6=CC=CC7=C6C=CC=C7)=O)C8=C3C=CC=C8
24	O=C(NC1=CC=CC=C1C(F)(F)F)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC=CC=C5C(F)(F)F)=O)C6=C2C=CC=C6
25	O=C(NC1=C(Cl)C=CC=C1Cl)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=C(Cl)C=CC=C5Cl)=O)C6=C2C=CC=C6
26	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=C(C)C=CC(C)=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=C(C)C=CC(C)=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
27	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC(C)=CC(C)=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC(C)=CC(C)=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
28	NC1=CC=C2C(C=CC=C2)=[C@]1[C@]3=C(NC(NC4=C(Cl)C=CC=C4Cl)=O)C=CC5=C3C=CC=C5
29	NC1=CC=C2C(C=CC=C2)=[C@]1[C@]3=C(NC(NC4CCCCC4)=O)C=CC5=C3C=CC=C5

30	O=C(NC1=C(F)C(F)=C(F)C(F)=C1F)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=C(F)C(F)=C(F)C(F)=C5F)=O)C6=C2C=CC=C6
31	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)=CC(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
32	CC1=CC=CC(NC(NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=O)=[C@]1[C@]3=C(C)C=CC=C3NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O
33	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC(Br)=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=C(Br)C=C6
34	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C3=CC=CC4=C3C=CC=C4)C=C2)=C5)=[C@]([C@]6=C(CCCC7)C7=CC(C8=CC=C(C9=CC=CC%10=C9C=CC=C%10)C=C8)=C6NC(NC%11=CC(C(F)(F)F)=CC(C(F)(F)F)=C%11)=O)C%12=C5CCCC%12
35	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C3=CC(C=CC=C4)=C4C=C3)C=C2)=C5)=[C@]([C@]6=C(CCCC7)C7=CC(C8=CC=C(C9=CC(C=CC=C%10)=C%10C=C9)C=C8)=C6NC(NC%11=CC(C(F)(F)F)=CC(C(F)(F)F)=C%11)=O)C%12=C5CCCC%12

36	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC(C4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=C5)C5=CC=C3NC(NC6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=O)C7=C2C=C(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)C=C7
37	O=C(NC1=CC(C2=CC=CC=C2)=CC(C3=CC=CC=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC=CC=C8)=CC(C9=CC=CC=C9)=C7)=O)C%10=C4C=CC=C%10
38	NC1=CC=C2C(C=CC(Br)=C2)=[C@]1[C@]3=C(NC(NC4=CC(C5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=CC(C6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=C4)=O)C=CC7=C3C=CC(Br)=C7
39	NC1=CC=C2C(CCCC2)=[C@]1[C@]3=C(NC(NC4=CC(C5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=CC(C6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=C4)=O)C=CC7=C3CCCC7
40	O=C(NC1=CC(C2=CC(C)=CC(C)=C2)=CC(C3=CC(C)=CC(C)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C)=CC(C)=C8)=CC(C9=CC(C)=CC(C)=C9)=C7)=O)C%10=C4C=CC=C%10
41	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(CCCC6)C6=CC=C5NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C4CCCC8
42	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C(C4=CC=C(C5=CC=CC=C5)C=C4)=C6)=[C@]([C@]7=C(CCCC8)C8=CC(C9=CC=C(C%10=CC=CC=C%10)C=C9)=C7NC(NC%11=CC(C%12=CC(C(F)(F)F)=CC(C(F)(F)F)=C%12)=CC(C%13=CC(C(F)(F)F)=CC(C(F)(F)F)=C%13)=C%11)=O)C%14=C6CCCC%14

43	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(CCCC6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)=CC(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=C7)=O)C%10=C4CCCC%10
44	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC(O)=CC=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC(O)=CC=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
45	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=C(O)C=CC=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=C(O)C=CC=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
46	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C3=CC=C(CC)C=C3)C=C2)=C4)=[C@]([C@]5=C(CCCC6)C6=CC(C7=CC=C(C8=CC=C(CC)C=C8)C=C7)=C5NC(NC9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=O)C%10=C4CCCC%10
47	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(OC)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(OC)C=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
48	O=C(NC1=CC=C(S(F)(F)(F)(F)F)C=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC=C(S(F)(F)(F)(F)F)C=C5)=O)C6=C2C=CC=C6

49	O=C(NC1=CC(C2=CC=CC=C2)=CC(C3=CC=CC=C3)=C1)NC(C(C4=CC=C(C5=CC=CC=C5)C=C4)=C6)=[C@]([C@]7=C(CCCC8)C8=CC(C9=CC=C(C%10=CC=CC=C%10)C=C9)=C7NC(NC%11=CC(C%12=CC=CC=C%12)=CC(C%13=CC=CC=C%13)=C%11)=O)C%14=C6CCCC%14
50	O=C(NC1=CC(C)=CC(C)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C)=CC(C)=C5)=O)C6=C2C=CC=C6
51	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(C)C(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)=C(C)C(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
52	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=C(F)C(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)=C(F)C(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
53	O=C(NC1=CC(C(F)(C(F)(F)F)C(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(C(F)(F)F)C(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C5)=O)C6=C2C=CC=C6

54	O=C(NC1=CC(C2=CC(C(F)(F)F)=C(F)C(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=C(F)C(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=C(F)C(C(F)(F)F)=C8)=CC(C9=CC(C(F)(F)F)=C(F)C(C(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
55	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC=CC3=C2[C@@]4(CCC5=C4C(NC(NC6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=O)=CC=C5)CC3
56	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC4=CC=CC5=C4[C@@]6(CCC7=C6C(NC(NC8=CC(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=CC(C%10=CC(C(F)(F)F)=CC(C(F)(F)F)=C%10)=C8)=O)=CC=C7)CC5
57	O=C(NC1=CC(C)=CC(C)=C1)NC(C(C2=CC=C(OC)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(OC)C=C6)=C4NC(NC7=CC(C)=CC(C)=C7)=O)C8=C3CCCC8
58	O=C(NC1=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C5)=O)C6=C2C=CC=C6
59	O=C(NC1=CC(C2=CC(C(F)(C(F)(F)F)C(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C2)=CC(C3=CC(C(F)(C(F)(F)F)C(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(C(F)(F)F)C

	(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C8)=CC(C9=CC(C(F)(C(F)(F)F)C(F)(F)F)=CC(C(F)(C(F)(F)F)C(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
60	O=C(NC1=CC(C2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)=CC(C3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC(C)=C6)C6=CC=C5NC(NC7=CC(C8=CC(C(F)(F)F)=CC(C(F)(F)F)=C8)=CC(C9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=C7)=O)C%10=C4C=C(C)C=C%10
61	O=C(NC1=CC(C2=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C2)=CC(C3=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C3)=C1)NC(C=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC=C5NC(NC7=CC(C8=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C8)=CC(C9=CC(S(F)(F)(F)(F)F)=CC(S(F)(F)(F)(F)F)=C9)=C7)=O)C%10=C4C=CC=C%10
62	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(C)=O)C5=C2C=CC=C5
63	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(C3=CC=CC=C3)C=C2)=C4)=[C@]([C@]5=C(C=CC=C6)C6=CC(C7=CC=C(C8=CC=CC=C8)C=C7)=C5NC(NC9=CC(C(F)(F)F)=CC(C(F)(F)F)=C9)=O)C%10=C4C=CC=C%10
64	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C(C2=CC=C(OC(C)(C)C)C=C2)=C3)=[C@]([C@]4=C(CCCC5)C5=CC(C6=CC=C(OC(C)(C)C)C=C6)=C4NC(NC7=CC(C(F)(F)F)=CC(C(F)(F)F)=C7)=O)C8=C3CCCC8
65	O=C(C1NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)C(C1N[C@H]([C@H](NC3C(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)C(C3=O)=O)C5=CC=CC=C5)C6=CC=CC=C6)=O

66	O=C(C1NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)C(C1N[C@H]([C@@H](O)C3=CC=CC=C3)C4=CC=CC=C4)=O
67	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@@H](C2=CC=CC=C2O)[C@H](C3=CC=CC=C3O)NC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=S
68	O=C(NC1=C(OC)C=CC=C1)N[C@H](C2=CC=CC=C2)[C@@H](C3=CC=CC=C3)NC(NC4=C(OC)C=CC=C4)=O
69	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N[C@H](C2=CC=CC=C2)[C@@H](C3=CC=CC=C3)NS(C4=CC=C(C)C=C4)(=O)=O
70	O=C(C1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC=C3C(C=CC=C3)=[C@]2[C@]4=C(NC(C5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C=CC6=CC=CC=C64
71	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=CC=C3C(C=CC=C3)=[C@]2[C@]4=C(NS(C5=CC=C(C)C=C5)(=O)=O)C=CC6=CC=CC=C64
72	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC2=C([C@@H](NC(NC3=CC(C(F)(F)F)=CC(C(F)(F)F)=C3)=O)C4=CC=CC=C4)C=CC=C2
73	OC1=[C@@]([C@@]2=C(C=CC=C3)C3=CC(CNC(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)=O)=C2O)C5=CC=CC=C5C=C1CNC(NC6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=O

74	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NCC2=CC3=CC=CC=C3[C@]([C@]4=C(C=CC=C5)C5=CC(CNC(NC6=CC(C(F)(F)F)=CC(C(F)(F)F)=C6)=O)=C4OCOCC)=C2OCOCC
75	O=C(C1NC2=CC(C(F)(F)F)=CC(C(F)(F)F)=C2)C(C1N[C@H]([C@H](NC3C(NC4=CC(C(F)(F)F)=CC(C(F)(F)F)=C4)C(C3=O)=O)C5=CC=CC=C5)C6=CC=CC=C6)=O
76	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N(C(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=CC=C6)CC
77	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N(C(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=CC=C6)C(C)C
78	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N(C)C(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=CC=C6
79	O=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)N(C(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=O)C6=C2C=CC=C6)CCC
80	S=C(NC1=CC(C(F)(F)F)=CC(C(F)(F)F)=C1)NC(C=C2)=[C@]([C@]3=C(C=CC=C4)C4=CC=C3NC(NC5=CC(C(F)(F)F)=CC(C(F)(F)F)=C5)=S)C6=C2C=CC=C6

Below are the SMILES strings for each of the 131 optimised catalysts in the Handcrafted Dataset along with their id numbers

Number	*Smiles*
H1	CCCC(CCC)(CCC)N(C(Nc(cc1)ccc1-c1cccs1)=O)c(c(C=C)cc1ccccc11)c1-c(c1ccccc1cc1C=C)c1NC(Nc(cc1)ccc1-c1cccs1)=O
H2	CCCC(CCC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CCC)CCC)c2N(C2CCCC2)C(Nc2cc([C@H](C/C=C(/C=C/F)\F)c3cc(F)cc(F)c3)cc([C@H](C/C=C(/C=C\F)\F)c3cc(F)cc(F)c3)c2)=O)c1NC(Nc1cc([C@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc([C@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)c1)=O
H3	O=C(Nc1cc(-c2cocc2)cc(-c2cocc2)c1)Nc(ccc1cc(-c2cccs2)ccc11)c1-c1c(ccc(-c2cccs2)c2)c2ccc1N(C(Nc1cc(-c2cocc2)cc(-c2cocc2)c1)=O)c1ccco1
H4	CCCC(CCC)c1cc(NC(Nc(ccc2cc(C3CCCC3)ccc22)c2-c(c2ccc(C3CCCC3)cc2cc2)c2N(C(F)(F)F)C(Nc2cc(C(CCC)CCC)cc(C(CCC)CCC)c2)=O)=O)cc(C(CCC)CCC)c1
H5	CCCc1cc(NC(Nc(ccc2cc(C3=CCCCC3)ccc22)c2-c(c(c(cc2)c3)ccc3C3=CCCCC3)c2N(C(Nc2cc(CCC)cc(CCC)c2)=O)c2ccccc2)=O)cc(CCC)c1
H6	O=C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)Nc(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1N(CCl)C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O

H7	O=C(Nc(cc1)ccc1-c1ccco1)Nc(ccc1cc(C(c2ccccc2)c2ccccc2)ccc11)c1-c(c1ccc(C(c2ccccc2)c2ccccc2)cc1cc1)c1N(C(Nc(cc1)ccc1-c1ccco1)=O)c1cccc2ccccc12
H9	CCC(CC)(CC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CC)(CC)CC)c2N(C(Nc2ccc(CC3CCCCC3)cc2)=O)c2ccco2)c1NC(Nc1ccc(CC2CCCCC2)cc1)=O
H10	CC(C)Cc(cc1)ccc1NC(Nc(c(-c1cscc1)cc1ccccc11)c1-c(c1ccccc1cc1-c2cscc2)c1N(C(C)C)C(Nc1ccc(CC(C)C)cc1)=O)=O
H11	CCC(CC)(CC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CC)(CC)CC)c2N(CC=C)C(Nc(cc2)ccc2-c2cccs2)=O)c1NC(Nc(cc1)ccc1-c1cccs1)=O
H12	O=C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)Nc(c(-c1c(C=CC2)c2ccc1)cc1ccccc11)c1-c(c1ccccc1cc1-c2c(C=CC3)c3ccc2)c1N(C=C1CCCCC1)C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O
H13	CC(C)N(C(Nc1ccc(CC=C)cc1)=O)c(ccc1cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)ccc11)c1-c(c1ccc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)cc1cc1)c1NC(Nc1ccc(CC=C)cc1)=O
H14	CCC(CC)(CC)N(C(Nc1ccc(CC(C)C)cc1)=O)c(c(/C=C/C)cc1ccccc11)c1-c(c1ccccc1cc1/C=C/C)c1NC(Nc1ccc(CC(C)C)cc1)=O
H15	CC(C)N(C(Nc1cc(C2=CCCCC2)cc(C2=CCCCC2)c1)=O)c(ccc1cc(-c2ccco2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccco2)c1NC(Nc1cc(C2=CCCCC2)cc(C2=CCCCC2)c1)=O
H16	O=C(Nc(cc1)ccc1C1=CCCCC1)Nc(c(C(c1cc(C(F)(F)F)cc(C(F)(F)F)c1)c1cc(C(F)(F)F)cc(C(F)(F)F)c1)cc1ccccc11)c1-c(c1ccccc1cc1C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c1N(Cc1ccccc1)C(Nc(cc1)ccc1C1=CCCCC1)=O

H17	CCCc1cc2ccccc2c(-c(c2ccccc2cc2CCC)c2N(/C=C/c2ccccc2)C(Nc2ccc(Cc3ccccc3)cc2)=O)c1NC(Nc1ccc(Cc2ccccc2)cc1)=O
H18	CC(C)(C)Cc(cc1)ccc1NC(Nc(c(-c1cccs1)cc1ccccc11)c1-c(c1ccccc1cc1-c2cccs2)c1N(CC(F)(F)F)C(Nc1ccc(CC(C)(C)C)cc1)=O)=O
H19	CCCc(cc1cc2)ccc1c(-c(c1ccc(CCC)cc1cc1)c1N(Cc1cccc3ccccc13)C(Nc(cc1)ccc1-c1cccs1)=O)c2NC(Nc(cc1)ccc1-c1cccs1)=O
H20	C[C@@H](CC1)CC/C1=C/c(cc1cc2)ccc1c(-c(c1ccc(/C=C3/CC[C@@H](C)CC3)cc1cc1)c1N(/C=C1/CC[C@H](C)CC1)C(Nc1ccc(C3CCC(C)(C)CC3)cc1)=O)c2NC(Nc1ccc(C2CCC(C)(C)CC2)cc1)=O
H22	CCCC(CCC)(CCC)c(cc1)ccc1NC(Nc(c(-c1cocc1)cc1ccccc11)c1-c(c1ccccc1cc1-c2cocc2)c1N(Cc1cc2ccccc2cc1)C(Nc1ccc(C(CCC)(CCC)CCC)cc1)=O)=O
H23	O=C(Nc1ccc(Cc2cc3ccccc3cc2)cc1)Nc(ccc1cc(CC(F)(F)F)ccc11)c1-c(c1ccc(CC(F)(F)F)cc1cc1)c1N(CCl)C(Nc1ccc(Cc2cc3ccccc3cc2)cc1)=O
H24	CC(C)(CC1)CCC1N(C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O)c(ccc1cc(C2=CCCCC2)ccc11)c1-c(c(c(cc1)c2)ccc2C2=CCCCC2)c1NC(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O
H25	CCCC(CCC)(CCC)N(C(Nc1cc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)c1)=O)c(ccc1cc(C(c2ccc(C)cc2)c2ccc(C)cc2)ccc11)c1-

	c(c1ccc(C(c2ccc(C)cc2)c2ccc(C)cc2)cc1cc1)c1NC(Nc1cc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)c1)=O
H26	CN(C(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O)c(c(Cc1cc2ccccc2cc1)cc1ccccc11)c1-c(c1ccccc1cc1Cc2cc3ccccc3cc2)c1NC(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O
H27	CC1(C)CC(c(cc2)ccc2NC(Nc(ccc2cc(C3=CCCC(C)(C)C3)ccc22)c2-c(c(c(cc2)c3)ccc3C3=CCCC(C)(C)C3)c2N(Cc2cc3ccccc3cc2)C(Nc(cc2)ccc2C2=CCCC(C)(C)C2)=O)=O)=CCC1
H28	CC(C)[C@@H](C)N(C(Nc1cc(/C=C\C)cc(/C=C\C)c1)=O)c(ccc1cc(C(c2cc(C)cc(C)c2)c2cc(C)cc(C)c2)ccc11)c1-c(c1ccc(C(c2cc(C)cc(C)c2)c2cc(C)cc(C)c2)cc1cc1)c1NC(Nc1cc(/C=C\C)cc(/C=C/C)c1)=O
H29	CCN(C(Nc1ccc(Cc2cccc3ccccc23)cc1)=O)c(ccc1cc(C2=CCCC(C)(C)C2)ccc11)c1-c(c(c(cc1)c2)ccc2C2=CCCC(C)(C)C2)c1NC(Nc1ccc(Cc2cccc3ccccc23)cc1)=O
H30	CC(C)(CC1)CCC1c(cc1cc2)ccc1c(-c(c1ccc(C3CCC(C)(C)CC3)cc1cc1)c1N(CCl)C(Nc1ccc(C(c3cc(C)cc(C)c3)c3cc(C)cc(C)c3)cc1)=O)c2NC(Nc1ccc(C(c2cc(C)cc(C)c2)c2cc(C)cc(C)c2)cc1)=O
H31	O=C(Nc1cc(Cc2cc3ccccc3cc2)cc(Cc2cc3ccccc3cc2)c1)Nc(c(CC(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1CC(F)(F)F)c1N(Cc1cc2ccccc2cc1)C(Nc1cc(Cc2cc3ccccc3cc2)cc(Cc2cc3ccccc3cc2)c1)=O

H32	CC(C)[C@H](C)N(C(Nc1cc([C@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc([C@@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)c1)=O)c1ccc(cc(cc2)-c3cccs3)c2c1-c(c(c(cc1)c2)ccc2-c2cccs2)c1NC(Nc1cc([C@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc([C@@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)c1)=O
H33	O=C(Nc1ccc(C=C2CCCCC2)cc1)Nc(ccc1cc(CC(F)(F)F)ccc11)c1-c1c(ccc(CC(F)(F)F)c2)c2ccc1N(C=C1CCCCC1)C(Nc1ccc(C=C2CCCCC2)cc1)=O
H34	CCCN(C(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O)c(ccc1cc(CC(C)(C)C)ccc11)c1-c(c1ccc(CC(C)(C)C)cc1cc1)c1NC(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O
H35	CCCC(CCC)(CCC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CCC)(CCC)CCC)c2N(C(Nc2cc(CCC)cc(CCC)c2)=O)c2cc(C=CC3)c3cc2)c1NC(Nc1cc(CCC)cc(CCC)c1)=O
H36	CCCN(C(Nc(cc1)ccc1-c1cccs1)=O)c(c(C(c1cc(C(F)(F)F)cc(C(F)(F)F)c1)c1cc(C(F)(F)F)cc(C(F)(F)F)c1)cc1ccccc11)c1-c(c1ccccc1cc1C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c1NC(Nc(cc1)ccc1-c1cccs1)=O
H37	O=C(Nc1cc(C2CCCCC2)cc(C2CCCCC2)c1)Nc(ccc1cc(/C=C\c2ccccc2)ccc11)c1-c(c1ccc(/C=C/c2ccccc2)cc1cc1)c1N(C(Nc1cc(C2CCCCC2)cc(C2CCCCC2)c1)=O)c1cscc1
H38	CCCC(CCC)(CCC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CCC)(CCC)CCC)c2N(C(C)C)C(Nc2ccc(C=C3CCCCC3)cc2)=O)c1NC(Nc1ccc(C=C2CCCCC2)cc1)=O

H39	CC(C)[C@@H](C)c(cc1cc2)ccc1c(-c(c1ccc([C@H](C)C(C)C)cc1cc1)c1N(CC=C)C(Nc1cc(CC3CCCCC3)cc(CC3CCCCC3)c1)=O)c2NC(Nc1cc(CC2CCCCC2)cc(CC2CCCCC2)c1)=O
H40	CC(C)c1cc(NC(Nc(c(C2=CCCC(C)(C)C2)cc2ccccc22)c2-c(c2ccccc2cc2C3=CCCC(C)(C)C3)c2N(CC(C)(C)C)C(Nc2cc(C(C)C)cc(C(C)C)c2)=O)=O)cc(C(C)C)c1
H41	O=C(Nc1cc(/C=C\c2ccccc2)cc(/C=C/c2ccccc2)c1)Nc(ccc1cc(C2CCCC2)ccc11)c1-c(c1ccc(C2CCCC2)cc1cc1)c1N(/C=C/c1ccccc1)C(Nc1cc(/C=C\c2ccccc2)cc(/C=C/c2ccccc2)c1)=O
H42	CCCc(cc1)ccc1NC(Nc(c(-c1c(cccc2)c2cc2ccccc12)cc1ccccc11)c1-c(c1ccccc1cc1-c2c(cccc3)c3cc3ccccc23)c1N(Cc1cccc2ccccc12)C(Nc1ccc(CCC)cc1)=O)=O
H43	CC(C)Cc1cc(NC(Nc(c(C(C)C)cc2ccccc22)c2-c(c2ccccc2cc2C(C)C)c2N(C(C)C)C(Nc2cc(CC(C)C)cc(CC(C)C)c2)=O)=O)cc(CC(C)C)c1
H44	O=C(Nc1ccc(C2CCCCC2)cc1)Nc(c(-c1ccco1)cc1ccccc11)c1-c(c1ccccc1cc1-c2ccco2)c1N(C(Nc1ccc(C2CCCCC2)cc1)=O)c1cc(C=CC2)c2cc1
H45	CC(C)Cc1cc2ccccc2c(-c(c2ccccc2cc2CC(C)C)c2N(CC(C)C)C(Nc2ccc(/C=C\c3ccccc3)cc2)=O)c1NC(Nc1ccc(/C=C\c2ccccc2)cc1)=O
H46	CC(C)c(cc1)ccc1NC(Nc(ccc1cc(C(c2ccc(C(F)(F)F)cc2)c2ccc(C(F)(F)F)cc2)ccc11)c1-c(c1ccc(C(c2ccc(C(F)(F)F)cc2)c2ccc(C(F)(F)F)cc2)cc1cc1)c1N(CCl)C(Nc1ccc(C(C)C)cc1)=O)=O

H47	CCCN(C(Nc1cc(-c2cccc3ccccc23)cc(-c2cccc3ccccc23)c1)=O)c(c(C(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1C(F)(F)F)c1NC(Nc1cc(-c2cccc3ccccc23)cc(-c2cccc3ccccc23)c1)=O
H48	CCC(CC)(CC)c(cc1cc2)ccc1c(-c(c1ccc(C(CC)(CC)CC)cc1cc1)c1N(C)C(Nc1cc(-c3cccs3)cc(-c3cccs3)c1)=O)c2NC(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O
H49	CCCC(CCC)c(cc1)ccc1NC(Nc(c(CCl)cc1ccccc11)c1-c(c1ccccc1cc1CCl)c1N(C(Nc1ccc(C(CCC)CCC)cc1)=O)c1cccc2ccccc12)=O
H50	C#Cc1cc(NC(Nc(c(CC(F)(F)F)cc2ccccc22)c2-c(c2ccccc2cc2CC(F)(F)F)c2N(C(Nc2cc(C#C)cc(C#C)c2)=O)c2c(cccc3)c3cc3ccccc23)=O)cc(C#C)c1
H51	O=C(Nc(cc1)ccc1-c1cccs1)Nc(ccc1cc(/C=C/c2ccccc2)ccc11)c1-c(c1ccc(/C=C\c2ccccc2)cc1cc1)c1N(C(Nc(cc1)ccc1-c1cccs1)=O)c1cocc1
H52	CCCN(C(Nc1ccc(CC(C)(C)C)cc1)=O)c(ccc1cc(-c2cocc2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2cocc2)c1NC(Nc1ccc(CC(C)(C)C)cc1)=O
H53	C=CCc(cc1cc2)ccc1c(-c(c1ccc(CC=C)cc1cc1)c1N(/C=C\c1ccccc1)C(Nc1ccc(C(c3cc(C(F)(F)F)cc(C(F)(F)F)c3)c3cc(C(F)(F)F)cc(C(F)(F)F)c3)cc1)=O)c2NC(Nc1ccc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc1)=O
H54	CC(C)(C)[C@@H](CC1)CC/C1=C/N(C(Nc1ccc(/C=C\c2ccccc2)cc1)=O)c(c(Cc1cccc2ccccc12)cc1ccccc11)c1-c(c1ccccc1cc1Cc2cccc3ccccc23)c1NC(Nc1ccc(/C=C/c2ccccc2)cc1)=O

H55	CCCC(CCC)c1cc(NC(Nc(c(-c2ccco2)cc2ccccc22)c2-c(c2ccccc2cc2-c3ccco3)c2N(C(F)(F)F)C(Nc2cc(C(CCC)CCC)cc(C(CCC)CCC)c2)=O)=O)cc(C(CCC)CCC)c1
H56	O=C(Nc1ccc(C(c2ccccc2)(c2ccccc2)c2ccccc2)cc1)Nc(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1N(C(Nc1ccc(C(c2ccccc2)(c2ccccc2)c2ccccc2)cc1)=O)c1cscc1
H57	CC(C)N(C(Nc1cc([C@@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)cc([C@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)c1)=O)c(ccc1cc([C@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)ccc11)c1-c(c1ccc([C@@H](C/C=C(\C=C/F)/F)c2cc(F)cc(F)c2)cc1cc1)c1NC(Nc1cc([C@@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc([C@@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)c1)=O
H58	O=C(Nc1cc(C(c(cc2)ccc2F)c(cc2)ccc2F)cc(C(c(cc2)ccc2F)c(cc2)ccc2F)c1)Nc(ccc1cc([C@@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)ccc11)c1-c(c1ccc([C@H](C/C=C(\C=C/F)/F)c2cc(F)cc(F)c2)cc1cc1)c1N(CCl)C(Nc1cc(C(c(cc2)ccc2F)c(cc2)ccc2F)cc(C(c(cc2)ccc2F)c(cc2)ccc2F)c1)=O
H59	CC(C)(C)Cc1cc2ccccc2c(-c(c2ccccc2cc2CC(C)(C)C)c2N(C(Nc2cc(C3=CCCC(C)(C)C3)cc(C3=CCCC(C)(C)C3)c2)=O)c2ccco2)c1NC(Nc1cc(C2=CCCC(C)(C)C2)cc(C2=CCCC(C)(C)C2)c1)=O

H60	CN(C(Nc1cc(C2=CCCCC2)cc(C2=CCCCC2)c1)=O)c(c(C1CCCC1)cc1ccccc11)c1-c(c1ccccc1cc1C2CCCC2)c1NC(Nc1cc(C2=CCCCC2)cc(C2=CCCCC2)c1)=O
H61	CCCC(CCC)(CCC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CCC)(CCC)CCC)c2N(C)C(Nc(cc2)ccc2-c2cccs2)=O)c1NC(Nc(cc1)ccc1-c1cccs1)=O
H62	O=C(Nc1cc(-c2cc(C=CC3)c3cc2)cc(-c2cc(C=CC3)c3cc2)c1)Nc(ccc1cc(-c2cccs2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2cccs2)c1N(CCl)C(Nc1cc(-c2cc(C=CC3)c3cc2)cc(-c2cc(C=CC3)c3cc2)c1)=O
H64	CCCc(cc1)ccc1NC(Nc(ccc1cc(-c2ccco2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccco2)c1N([C@@H](C)C(C)C)C(Nc1ccc(CCC)cc1)=O)=O
H65	C#CN(C(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O)c(ccc1cc(C(c2ccccc2)c2ccccc2)ccc11)c1-c(c1ccc(C(c2ccccc2)c2ccccc2)cc1cc1)c1NC(Nc1cc(-c2ccco2)cc(-c2ccco2)c1)=O
H66	CC(C)(C)[C@H](CC1)CC/C1=C/N(C(Nc1cc(Cc2cccc3ccccc23)cc(Cc2cccc3ccccc23)c1)=O)c1ccc(cc([C@H](C/C=C(\C=C/F)/F)c2cc(F)cc(F)c2)cc2)c2c1-c(c1ccc([C@@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)cc1cc1)c1NC(Nc1cc(Cc2cccc3ccccc23)cc(Cc2cccc3ccccc23)c1)=O
H67	CCN(C(Nc1ccc(/C=C2/CC[C@H](C)CC2)cc1)=O)c(c(CC(C)C)cc1ccccc11)c1-c(c1ccccc1cc1CC(C)C)c1NC(Nc1ccc(/C=C2/CC[C@H](C)CC2)cc1)=O

H68	CC(C)(C)CN(C(Nc1ccc([C@@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)cc1)=O)c(ccc1cc(-c2cccc3ccccc23)ccc11)c1-c(c(c(cc1)c2)ccc2-c2cccc3ccccc23)c1NC(Nc1ccc([C@@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc1)=O
H69	CC(C)c(cc1)ccc1NC(Nc(ccc1cc(-c2ccccc2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccccc2)c1N(C(Nc1ccc(C(C)C)cc1)=O)c1ccco1)=O
H71	O=C(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)Nc(c(CC(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1CC(F)(F)F)c1N(C(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O)c1cocc1
H72	C/C=C/N(C(Nc1ccc(C2CCCCC2)cc1)=O)c(c(C)cc1ccccc11)c1-c(c1ccccc1cc1C)c1NC(Nc1ccc(C2CCCCC2)cc1)=O
H73	CCN(C(Nc1cc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)c1)=O)c(ccc1cc(-c2ccco2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccco2)c1NC(Nc1cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)cc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)c1)=O
H74	CCCc1cc(NC(Nc(ccc2cc(-c3ccco3)ccc22)c2-c(c(c(cc2)c3)ccc3-c3ccco3)c2N(CC=C)C(Nc2cc(CCC)cc(CCC)c2)=O)=O)cc(CCC)c1
H75	CC(C)(C)Cc1cc(NC(Nc(ccc2cc(C3CCC(C)(C)CC3)ccc22)c2-c(c2ccc(C3CCC(C)(C)CC3)cc2cc2)c2N(Cc2cc3ccccc3cc2)C(Nc2cc(CC(C)(C)C)cc(CC(C)(C)C)c2)=O)=O)cc(CC(C)(C)C)c1
H76	O=C(Nc1cc(C(c2ccccc2)c2ccccc2)cc(C(c2ccccc2)c2ccccc2)c1)Nc(ccc1cc(Cc2ccccc2)ccc11)c1-c(c1ccc(Cc2ccccc2)cc1cc1)c1N(C(Nc1cc(C(c2ccccc2)c2ccccc2)cc(C(c2ccccc2)c2ccccc2)c1)=O)c1cocc1
H77	CC(C)c(cc1cc2)ccc1c(-c(c1ccc(C(C)C)cc1cc1)c1N(/C=C\c1ccccc1)C(Nc1ccc(CC(F)(F)F)cc1)=O)c2NC(Nc1ccc(CC(F)(F)F)cc1)=O

H78	CCN(C(Nc1cc(/C=C2/CC[C@@H](C)CC2)cc(/C=C2/CC[C@H](C)CC2)c1)=O)c(ccc1cc(-c2cocc2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2cocc2)c1NC(Nc1cc(/C=C2/CC[C@@H](C)CC2)cc(/C=C2/CC[C@@H](C)CC2)c1)=O
H79	CN(C(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O)c(c(C1CCCC1)cc1ccccc11)c1-c(c1ccccc1cc1C2CCCC2)c1NC(Nc1cc(-c2cccs2)cc(-c2cccs2)c1)=O
H81	O=C(Nc1cc(CC2CCCCC2)cc(CC2CCCCC2)c1)Nc(c(-c1cccs1)cc1ccccc11)c1-c(c1ccccc1cc1-c2cccs2)c1N(/C=C\c1ccccc1)C(Nc1cc(CC2CCCCC2)cc(CC2CCCCC2)c1)=O
H82	CCCN(C(Nc1cc(Cc2ccccc2)cc(Cc2ccccc2)c1)=O)c(c(C=C)cc1ccccc11)c1-c(c1ccccc1cc1C=C)c1NC(Nc1cc(Cc2ccccc2)cc(Cc2ccccc2)c1)=O
H83	C/C=C/c1cc2ccccc2c(-c(c2ccccc2cc2/C=C/C)c2N(CC(F)(F)F)C(Nc2ccc(CC=C)cc2)=O)c1NC(Nc1ccc(CC=C)cc1)=O
H84	C#Cc(cc1)ccc1NC(Nc(ccc1cc([C@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)ccc11)c1-c1c(ccc([C@H](C/C=C(/C=C/F)\F)c2cc(F)cc(F)c2)c2)c2ccc1N(Cc1cc2ccccc2cc1)C(Nc(cc1)ccc1C#C)=O)=O
H85	CCCN(C(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O)c1ccc(cc([C@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)cc2)c2c1-c(c1ccc([C@@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc1cc1)c1NC(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O
H86	C=CCc1cc(NC(Nc(c(CC(F)(F)F)cc2ccccc22)c2-c(c2ccccc2cc2CC(F)(F)F)c2N(C=C2CCCCC2)C(Nc2cc(CC=C)cc(CC=C)c2)=O)=O)cc(CC=C)c1

H87	CC(C)c(cc1cc2)ccc1c(-c(c1ccc(C(C)C)cc1cc1)c1N(C(Nc(cc1)ccc1C1=CCCCC1)=O)c1cccs1)c2NC(Nc(cc1)ccc1C1=CCCCC1)=O
H88	C=CCc(cc1)ccc1NC(Nc(ccc1cc(-c2cccs2)ccc11)c1-c1c(ccc(-c2cccs2)c2)c2ccc1N(C=C)C(Nc1ccc(CC=C)cc1)=O)=O
H89	CCCC(CCC)c1cc(NC(Nc(c(C(C)C)cc2ccccc22)c2-c(c2ccccc2cc2C(C)C)c2N(C(Nc2cc(C(CCC)CCC)cc(C(CCC)CCC)c2)=O)c2cccs2)=O)cc(C(CCC)CCC)c1
H90	O=C(Nc(cc1)ccc1-c1cccc2ccccc12)Nc(ccc1cc(C(F)(F)F)ccc11)c1-c(c1ccc(C(F)(F)F)cc1cc1)c1N(C(Nc(cc1)ccc1-c1cccc2ccccc12)=O)c1cccs1
H91	CCN(C(Nc1cc(C)cc(C)c1)=O)c(c(CC(C)(C)C)cc1ccccc11)c1-c(c1ccccc1cc1CC(C)(C)C)c1NC(Nc1cc(C)cc(C)c1)=O
H92	CC(C)[C@H](C)c(cc1)ccc1NC(Nc(c(CCl)cc1ccccc11)c1-c(c1ccccc1cc1CCl)c1N(/C=C/c1ccccc1)C(Nc1ccc([C@H](C)C(C)C)cc1)=O)=O
H93	CCN(C(Nc1ccc(C2CCC(C)(C)CC2)cc1)=O)c(c(CC(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1CC(F)(F)F)c1NC(Nc1ccc(C2CCC(C)(C)CC2)cc1)=O
H94	C=CCN(C(Nc(cc1)ccc1-c1cscc1)=O)c(ccc1cc(-c2cocc2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2cocc2)c1NC(Nc(cc1)ccc1-c1cscc1)=O
H95	C[C@@H](CC1)CC/C1=C/c(cc1cc2)ccc1c(-c(c1ccc(/C=C3/CC[C@H](C)CC3)cc1cc1)c1N(/C=C/c1ccccc1)C(Nc1ccc(CC3CCCCC3)cc1)=O)c2NC(Nc1ccc(CC2CCCCC2)cc1)=O

H96	C[C@H](CC1)CC/C1=C/N(C(Nc1cc(/C=C\C)cc(/C=C/C)c1)=O)c(c(C(c1cc(C)cc(C)c1)c1cc(C)cc(C)c1)cc1ccccc11)c1-c(c1ccccc1cc1C(c2cc(C)cc(C)c2)c2cc(C)cc(C)c2)c1NC(Nc1cc(/C=C/C)cc(/C=C/C)c1)=O
H97	CCC(CC)(CC)c(cc1cc2)ccc1c(-c(c1ccc(C(CC)(CC)CC)cc1cc1)c1N(C(Nc1ccc(C(C)C)cc1)=O)c1ccccc1)c2NC(Nc1ccc(C(C)C)cc1)=O
H98	CC(C)(C)CN(C(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O)c1ccc(cc(cc2)-c3cccs3)c2c1-c(c(c(cc1)c2)ccc2-c2cccs2)c1NC(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O
H99	CCCC(CCC)(CCC)c1cc2ccccc2c(-c(c2ccccc2cc2C(CCC)(CCC)CCC)c2N([C@@H](C)C(C)C)C(Nc2ccc(/C=C3/CC[C@@H](C)CC3)cc2)=O)c1NC(Nc1ccc(/C=C2/CC[C@H](C)CC2)cc1)=O
H100	CCN(C(Nc1ccc(/C=C\c2ccccc2)cc1)=O)c(c([C@H](C)C(C)C)cc1ccccc11)c1-c(c1ccccc1cc1[C@H](C)C(C)C)c1NC(Nc1ccc(/C=C/c2ccccc2)cc1)=O
H101	CN(C(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O)c(c(C(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1C(F)(F)F)c1NC(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O
H102	O=C(Nc1ccc(CC(F)(F)F)cc1)Nc(c(-c1ccco1)cc1ccccc11)c1-c(c1ccccc1cc1-c2ccco2)c1N(/C=C/c1ccccc1)C(Nc1ccc(CC(F)(F)F)cc1)=O

H104	CC(C)(C)[C@@H](CC1)CC/C1=C/c1cc(NC(Nc(ccc2cc(Cc3cccc4ccccc34)ccc22)c2-c(c2ccc(Cc3cccc4ccccc34)cc2cc2)c2N(Cc2cccc3ccccc23)C(Nc2cc(/C=C(\CC3)/CC[C@H]3C(C)(C)C)cc(/C=C(\CC3)/CC[C@H]3C(C)(C)C)c2)=O)=O)cc(/C=C(\CC2)/CC[C@@H]2C(C)(C)C)c1
H105	CC(C)(C)Cc1cc(NC(Nc(c(CCl)cc2ccccc22)c2-c(c2ccccc2cc2CCl)c2N(/C=C\C)C(Nc2cc(CC(C)(C)C)cc(CC(C)(C)C)c2)=O)=O)cc(CC(C)(C)C)c1
H106	O=C(Nc1cc(Cc2ccccc2)cc(Cc2ccccc2)c1)Nc(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1N(C=C1CCCCC1)C(Nc1cc(Cc2ccccc2)cc(Cc2ccccc2)c1)=O
H107	CCCC(CCC)(CCC)c(cc1)ccc1NC(Nc(c(Cc1ccccc1)cc1ccccc11)c1-c(c1ccccc1cc1Cc2ccccc2)c1N(C(F)(F)F)C(Nc1ccc(C(CCC)(CCC)CCC)cc1)=O)=O
H108	O=C(Nc(cc1)ccc1-c1ccco1)Nc(c(-c1cccs1)cc1ccccc11)c1-c(c1ccccc1cc1-c2cccs2)c1N(C(Nc(cc1)ccc1-c1ccco1)=O)c1cccc2ccccc12
H109	CC(C)N(C(Nc1cc(/C=C\C)cc(/C=C\C)c1)=O)c(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1NC(Nc1cc(/C=C\C)cc(/C=C/C)c1)=O
H110	CC1(C)CC(N(C(Nc2cc(-c3c(C=CC4)c4ccc3)cc(-c3c(C=CC4)c4ccc3)c2)=O)c(ccc2cc(-c3c(C=CC4)c4ccc3)ccc22)c2-c(c(c(cc2)c3)ccc3-c3c(C=CC4)c4ccc3)c2NC(Nc2cc(-c3c(C=CC4)c4ccc3)cc(-c3c(C=CC4)c4ccc3)c2)=O)=CCC1
H111	CC(C)Cc1cc2ccccc2c(-c(c2ccccc2cc2CC(C)C)c2N(C(Nc(cc2)ccc2-c2cccs2)=O)C#C)c1NC(Nc(cc1)ccc1-c1cccs1)=O

H112	O=C(Nc1cc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c1)Nc(ccc1cc(CC2CCCCC2)ccc11)c1-c(c1ccc(CC2CCCCC2)cc1cc1)c1N(C(Nc1cc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c1)=O)c1cocc1
H113	C/C=C\N(C(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O)c(ccc1cc(CC2CCCCC2)ccc11)c1-c(c1ccc(CC2CCCCC2)cc1cc1)c1NC(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O
H114	CCc(cc1)ccc1NC(Nc(c(Cc1cccc2ccccc12)cc1ccccc11)c1-c(c1ccccc1cc1Cc2cccc3ccccc23)c1N(CC)C(Nc1ccc(CC)cc1)=O)=O
H115	O=C(Nc1ccc(C=C2CCCCC2)cc1)Nc(c(CC(F)(F)F)cc1ccccc11)c1-c(c1ccccc1cc1CC(F)(F)F)c1N(C(F)(F)F)C(Nc1ccc(C=C2CCCCC2)cc1)=O
H116	CC(C)[C@H](C)c(cc1)ccc1NC(Nc(ccc1cc(Cc2cccc3ccccc23)ccc11)c1-c(c1ccc(Cc2cccc3ccccc23)cc1cc1)c1N(C(Nc1ccc([C@H](C)C(C)C)cc1)=O)c1cc(C=CC2)c2cc1)=O
H117	CCCC(CCC)c(cc1)ccc1NC(Nc(ccc1cc(C=C2CCCCC2)ccc11)c1-c(c1ccc(C=C2CCCCC2)cc1cc1)c1N(C(Nc1ccc(C(CCC)CCC)cc1)=O)c1cscc1)=O

H119	CC(C)Cc(cc1cc2)ccc1c(-c(c1ccc(CC(C)C)cc1cc1)c1N(C(Nc1ccc(C(c3cc(C(F)(F)F)cc(C(F)(F)F)c3)c3cc(C(F)(F)F)cc(C(F)(F)F)c3)cc1)=O)c1ccco1)c2NC(Nc1ccc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc1)=O
H120	CC(C)c(cc1)ccc1NC(Nc(c(/C=C/C)cc1ccccc11)c1-c(c1ccccc1cc1/C=C\C)c1N(C(F)(F)F)C(Nc1ccc(C(C)C)cc1)=O)=O
H121	CC(C)(C)[C@@H](CC1)CC/C1=C/N(C(Nc1cc(Cc2cccc3ccccc23)cc(Cc2cccc3ccccc23)c1)=O)c(ccc1cc(-c2ccco2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccco2)c1NC(Nc1cc(Cc2cccc3ccccc23)cc(Cc2cccc3ccccc23)c1)=O
H122	CCCC(CCC)c(cc1cc2)ccc1c(-c(c1ccc(C(CCC)CCC)cc1cc1)c1N(C1CCCCC1)C(Nc1cc(Cc3ccccc3)cc(Cc3ccccc3)c1)=O)c2NC(Nc1cc(Cc2ccccc2)cc(Cc2ccccc2)c1)=O
H123	C[C@@H](CC1)CC/C1=C/c(cc1)ccc1NC(Nc(c(-c1ccco1)cc1ccccc11)c1-c(c1ccccc1cc1-c2ccco2)c1N(Cc1cc2ccccc2cc1)C(Nc1ccc(/C=C2/CC[C@@H](C)CC2)cc1)=O)=O
H124	O=C(Nc1ccc(C(c2ccccc2)c2ccccc2)cc1)Nc(c(Cc1cc2ccccc2cc1)cc1ccccc11)c1-c(c1ccccc1cc1Cc2cc3ccccc3cc2)c1N(CCl)C(Nc1ccc(C(c2ccccc2)c2ccccc2)cc1)=O
H125	CC(C)(CC1)CCC1c1cc(NC(Nc(ccc2cc(Cc3cc4ccccc4cc3)ccc22)c2-c(c2ccc(Cc3cc4ccccc4cc3)cc2cc2)c2N(C(Nc2cc(C3CCC(C)(C)CC3)cc(C3CCC(C)(C)CC3)c2)=O)c2ccco2)=O)cc(C2CCC(C)(C)CC2)c1

H126	CCN(C(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O)c(ccc1cc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)ccc11)c1-c(c1ccc(C(c2cc(C(F)(F)F)cc(C(F)(F)F)c2)c2cc(C(F)(F)F)cc(C(F)(F)F)c2)cc1cc1)c1NC(Nc1cc(C(F)(F)F)cc(C(F)(F)F)c1)=O
H127	CCCN(C(Nc1ccc(C(F)(F)F)cc1)=O)c(c(C1=CCCC(C)(C)C1)cc1ccccc11)c1-c(c1ccccc1cc1C2=CCCC(C)(C)C2)c1NC(Nc1ccc(C(F)(F)F)cc1)=O
H128	O=C(Nc1ccc(C(c2ccccc2)(c2ccccc2)c2ccccc2)cc1)Nc(ccc1cc(-c2cccs2)ccc11)c1-c1c(ccc(-c2cccs2)c2)c2ccc1N(C(F)(F)F)C(Nc1ccc(C(c2ccccc2)(c2ccccc2)c2ccccc2)cc1)=O
H129	Cc1ccc(C(c2ccc(C)cc2)c(cc2)ccc2NC(Nc(c(C2CCCCC2)cc2ccccc22)c2-c(c2ccccc2cc2C3CCCCC3)c2N(C(c(cc2)ccc2F)c(cc2)ccc2F)C(Nc2ccc(C(c3ccc(C)cc3)c3ccc(C)cc3)cc2)=O)=O)cc1
H130	CCN(C(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O)c(ccc1cc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)ccc11)c1-c(c1ccc(/C=C(\CC2)/CC[C@H]2C(C)(C)C)cc1cc1)c1NC(Nc1cc(-c2ccccc2)cc(-c2ccccc2)c1)=O
H131	CC(C)(C)[C@H](CC1)CC/C1=C/N(C(Nc1ccc(CC2CCCCC2)cc1)=O)c(c(C(c1ccccc1)(c1ccccc1)c1ccccc1)cc1ccccc11)c1-c(c1ccccc1cc1C(c2ccccc2)(c2ccccc2)c2ccccc2)c1NC(Nc1ccc(CC2CCCCC2)cc1)=O
H132	CCCC(CCC)N(C(Nc(cc1)ccc1-c1ccco1)=O)c(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1NC(Nc(cc1)ccc1-c1ccco1)=O
H133	C=CCc(cc1cc2)ccc1c(-c(c1ccc(CC=C)cc1cc1)c1N(C(Nc1cc(C=C)cc(C=C)c1)=O)c1cccc3ccccc13)c2NC(Nc1cc(C=C)cc(C=C)c1)=O

H134	CC(C)CN(C(Nc1cc(CC(F)(F)F)cc(CC(F)(F)F)c1)=O)c(ccc1cc(-c2ccco2)ccc11)c1-c(c(c(cc1)c2)ccc2-c2ccco2)c1NC(Nc1cc(CC(F)(F)F)cc(CC(F)(F)F)c1)=O
H135	O=C(Nc1cc(C(c2ccccc2)c2ccccc2)cc(C(c2ccccc2)c2ccccc2)c1)Nc(ccc1cc(C(c2ccc(C(F)(F)F)cc2)c2ccc(C(F)(F)F)cc2)ccc11)c1-c1c(ccc(C(c2ccc(C(F)(F)F)cc2)c2ccc(C(F)(F)F)cc2)c2)c2ccc1N(C=C1CCCCC1)C(Nc1cc(C(c2ccccc2)c2ccccc2)cc(C(c2ccccc2)c2ccccc2)c1)=O
H136	O=C(Nc1cc([C@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)cc([C@@H](C/C=C(\C=C\F)/F)c2cc(F)cc(F)c2)c1)Nc(c(CC1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1CC2CCCCC2)c1N(C(Nc1cc([C@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)cc([C@@H](C/C=C(/C=C\F)\F)c2cc(F)cc(F)c2)c1)=O)c1cccs1
H137	C=CCc1cc2ccccc2c(-c(c2ccccc2cc2CC=C)c2N(Cc2cc3ccccc3cc2)C(Nc2ccc(C3CCCCC3)cc2)=O)c1NC(Nc1ccc(C2CCCCC2)cc1)=O
H138	CC(C)c(cc1)ccc1NC(Nc(c(C=C1CCCCC1)cc1ccccc11)c1-c(c1ccccc1cc1C=C2CCCCC2)c1N(C(Nc1ccc(C(C)C)cc1)=O)c1c(C=CC2)c2ccc1)=O

Appendix 3 Dataset for Chapter 5

The final dataset used ^{19}F NMR data obtained from the following papers:

(1) Rosenau, C. P.; Jelier, B. J.; Gossert, A. D.; Togni, A. Exposing the Origins of Irreproducibility in Fluorine NMR Spectroscopy. Angew Chem Int Ed Engl 2018, 57 (30), 9528-9533. DOI: 10.1002/anie.201802620.

(2) Bacauanu, V.; Cardinal, S.; Yamauchi, M.; Kondo, M.; Fernandez, D. F.; Remy, R.; MacMillan, D. W. C. Metallaphotoredox Difluoromethylation of Aryl Bromides. Angew Chem Int Ed Engl 2018, 57 (38), 12543-12548. DOI: 10.1002/anie.201807629.

(3) Beeson, T. D.; Macmillan, D. W. Enantioselective organocatalytic alpha-fluorination of aldehydes. J Am Chem Soc 2005, 127 (24), 8826-8828. DOI: 10.1021/ja051805f.

(4) Ventre, S.; Petronijevic, F. R.; MacMillan, D. W. Decarboxylative Fluorination of Aliphatic Carboxylic Acids via Photoredox Catalysis. J Am Chem Soc 2015, 137 (17), 5654-5657. DOI: 10.1021/jacs.5b02244.

(5) Phipps, R. J.; Hiramatsu, K.; Toste, F. D. Asymmetric fluorination of enamides: access to alpha-fluoroimines using an anionic chiral phase-transfer catalyst. J Am Chem Soc 2012, 134 (20), 8376-8379. DOI: 10.1021/ja303959p.

(6) Yang, X.; Wu, T.; Phipps, R. J.; Toste, F. D. Advances in Catalytic Enantioselective Fluorination, Mono - , Di - , and Tri fl uoromethylation , and Tri fl uoromethylthiolation Reactions. 2015.

(7) Meyer, C. F.; Hell, S. M.; Misale, A.; Trabanco, A. A.; Gouverneur, V. Hydrodifluoromethylation of Alkenes with Difluoroacetic Acid. Angew Chem Int Ed Engl 2019, 58 (26), 8829-8833. DOI: 10.1002/anie.201903801.

(8) Meyer, C. F.; Hell, S. M.; Sap, J. B. I.; Misale, A.; Peschiulli, A.; Oehlrich, D.; Trabanco, A. A.; Gouverneur, V. Hydrochlorofluoromethylation of unactivated alkenes with chlorofluoroacetic acid. Tetrahedron 2019, 75 (47), 130679-130679. DOI: ARTN 130679

(9) Laudadio, G.; Bartolomeu, A. A.; Verwijlen, L.; Cao, Y.; de Oliveira, K. T.; Noel, T. Sulfonyl Fluoride Synthesis through Electrochemical Oxidative Coupling of Thiols and Potassium Fluoride. J Am Chem Soc 2019, 141 (30), 11832-11836. DOI: 10.1021/jacs.9b06126.

(10) Lou, T. S.; Bagley, S. W.; Willis, M. C. Cyclic Alkenylsulfonyl Fluorides: Palladium-Catalyzed Synthesis and Functionalization of Compact Multifunctional Reagents. Angew Chem Int Ed Engl 2019, 58 (52), 18859-18863. DOI: 10.1002/anie.201910871.

(11) Ponra, S.; Yang, J.; Kerdphon, S.; Andersson, P. G. Asymmetric Synthesis of Alkyl Fluorides: Hydrogenation of Fluorinated Olefins. Angew Chem Int Ed Engl 2019, 58 (27), 9282-9287. DOI: 10.1002/anie.201903954.

(12) Parisi, G.; Colella, M.; Monticelli, S.; Romanazzi, G.; Holzer, W.; Langer, T.; Degennaro, L.; Pace, V.; Luisi, R. Exploiting a "Beast" in Carbenoid Chemistry: Development of a Straightforward Direct Nucleophilic Fluoromethylation Strategy. J Am Chem Soc 2017, 139 (39), 13648-13651. DOI: 10.1021/jacs.7b07891.

(13) Saielli, G.; Bini, R.; Bagno, A. Computational 19F NMR. 1. General features. Theoretical Chemistry Accounts 2012, 131 (3), 1-11. DOI: 10.1007/s00214-012-1140-z.

(14) Bagutski, V.; Ros, A.; Aggarwal, V. K. Improved method for the conversion of pinacolboronic esters into trifluoroborate salts: facile synthesis of chiral secondary and tertiary trifluoroborates. Tetrahedron 2009, 65 (48), 9956-9960. DOI: https://doi.org/10.1016/j.tet.2009.10.002.

(15) Fier, P. S.; Hartwig, J. F. Selective C-H Fluorination of Pyridines and Diazines Inspired by a Classic Amination Reaction. Science 2013, 342 (6161), 956-960. DOI: 10.1126/science.1243759 (acccessed 2023/03/21).

(16) Taylor, N. J.; Emer, E.; Preshlock, S.; Schedler, M.; Tredwell, M.; Verhoog, S.; Mercier, J.; Genicot, C.; Gouverneur, V. Derisking the Cu-Mediated ¹⁸F-Fluorination of Heterocyclic Positron Emission Tomography Radioligands. Journal of the American Chemical Society 2017, 139 (24), 8267-8276. DOI: 10.1021/jacs.7b03131 (acccessed 2023-03-21T22:46:03).

(17) Nairoukh, Z.; Wollenburg, M.; Schlepphorst, C.; Bergander, K.; Glorius, F. The formation of all-cis-(multi)fluorinated piperidines by a dearomatization–hydrogenation process. Nature Chemistry 2019, 11 (3), 264-270. DOI: 10.1038/s41557-018-0197-2 (acccessed 2023-03-21T22:48:50).

(18) Liu, W.; Huang, X.; Cheng, M. J.; Nielsen, R. J.; Goddard, W. A., 3rd; Groves, J. T. Oxidative aliphatic C-H fluorination with fluoride ion catalyzed by a manganese porphyrin. Science 2012, 337 (6100), 1322-1325. DOI: 10.1126/science.1222327.

(19) Wiesenfeldt, M. P.; Nairoukh, Z.; Li, W.; Glorius, F. Hydrogenation of fluoroarenes: Direct access to all-cis-(multi)fluorinated cycloalkanes. Science 2017, 357 (6354), 908-912. DOI: 10.1126/science.aao0270 (acccessed 2023/03/21).

(20) Li, G.; Dilger, A. K.; Cheng, P. T.; Ewing, W. R.; Groves, J. T. Selective C−H Halogenation with a Highly Fluorinated Manganese Porphyrin. Angewandte Chemie International Edition 2018, 57 (5), 1251-1255. DOI: 10.1002/anie.201710676 (acccessed 2023-03-21T22:56:39).

(21) Denavit, V.; Laine, D.; St-Gelais, J.; Johnson, P. A.; Giguere, D. A Chiron approach towards the stereoselective synthesis of polyfluorinated carbohydrates. Nat Commun 2018, 9 (1), 4721. DOI: 10.1038/s41467-018-06901-y.

(22) Neumann, C. N.; Hooker, J. M.; Ritter, T. Concerted nucleophilic aromatic substitution with 19F− and 18F−. Nature 2016, 534 (7607), 369-373. DOI: 10.1038/nature17667 (acccessed 2023-03-21T23:01:23).

(23) Kim, S.; Khomutnyk, Y.; Bannykh, A.; Nagorny, P. Synthesis of Glycosyl Fluorides by Photochemical Fluorination with Sulfur(VI) Hexafluoride. Org Lett 2021, 23 (1), 190-194. DOI: 10.1021/acs.orglett.0c03915.

(24) Chen, P.; Wang, P.; Long, Q.; Ding, H.; Cheng, G.; Li, T.; Li, M. Synthesis of Reverse Glycosyl Fluorides and Rare Glycosyl Fluorides Enabled by Radical Decarboxylative Fluorination of Uronic Acids. Org Lett 2020, 22 (23), 9325-9330. DOI: 10.1021/acs.orglett.0c03514.