NMR Parameter Prediction with Machine Learning

By

WILL GERRARD

Department of Chemistry
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Science.

MARCH 2021

Word count: 28,876

Abstract

The prediction of NMR parameters through machine learning was investigated and several highly accurate prediction algorithms developed. Prediction models sensitive to 3-Dimensional structure in small molecules are presented for chemical shifts and scalar coupling constants, several of which outperform current state-of-the-art algorithms. Several large, high quality DFT datasets were also produced, their construction and composition are detailed in this work. Finally the application of the newly developed prediction algorithms to a realistic diastereomer discrimination task is explored, along with the adaptation of one of the machine learning frameworks to the prediction of binding affinities.

DEDICATION AND ACKNOWLEDGEMENTS

This work is dedicated to everyone who put up with me and supported me during the last four years, both academically and socially; My supervisors, fellow members of the Butts group (especially Lydia, who i hope doesn't still work as late), my parents, brothers, soon-to-be wife Olivia, and Harry.

I am extremely grateful to Prof. Craig Butts for support and guidance over the last four years in both life and chemistry, and for giving me a chance to study for a PhD.

This work was carried out using the computational facilities of the Advanced Computing Research Centre, University of Bristol - http://www.bristol.ac.uk/acrc/.

Finally i thank the EPSRC National Productivity Investment Fund (NPIF) for Doctoral Studentship funding.

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: .....WILL GERRARD..... DATE: .....25TH NOV 2021.....

TABLE OF CONTENTS

Page

List of Tables **xi**

List of Figures **xv**

1 Introduction **1**

1.1 NMR Spectroscopy 1

1.1.1 NMR spectroscopy in structure elucidation 1

1.1.2 NMR parameters and molecular structure 2

1.2 Computational NMR 3

1.2.1 Empirical equations 3

1.2.2 Density Functional Theory 4

1.2.3 Computational NMR and Structure elucidation 7

1.3 Artificial intelligence and Machine learning 8

1.4 Machine learning in NMR 10

1.4.1 Paruzzo et al 2018: Chemical shifts in molecular solids by machine learning 11

1.4.2 Jonas et al 2019: Rapid prediction of NMR spectral properties with quantified uncertainty 13

1.4.3 Gupta et al 2021: Revving up 13C NMR shielding predictions across chemical space: Benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules 14

1.4.4 Shibata et al 2021: Prediction of spin–spin coupling constants with machine learning in NMR 15

1.4.5 Summary 16

1.4.6 Licensed Software: ACD Labs 16

1.4.7 Open Source Software: NMRShiftDB 17
1.5 Aims and objectives 19
1.5.1 Machine learning datasets 20

2 Dataset Production **23**
2.1 Dataset requirements 23
2.1.1 Overfitting 23
2.1.2 Dataset Size 24
2.1.3 Dataset Breadth and Depth 24
2.1.4 Dataset Quality 24
2.1.5 Dataset Credibility and bias 25
2.1.6 Testing Datasets: relevance. 26
2.2 DFT NMR Calculation 27
2.2.1 'Mixed' Keyword Calculation Issue 27
2.2.2 Chemical Shift Scaling 29
2.2.3 Computational Timing 29
2.3 Dataset Workflow 31
2.3.1 Molecule Screening 32
2.4 The Datasets 32
2.4.1 Dataset 1 and 2: Initial random sets 32
2.4.2 Dataset 3: Random Testing Set (DT3) 33
2.4.3 Dataset 4: Adaptive sampling training set (DT4) 33
2.4.4 Dataset 5: ChEMBL (DT5a and DT5b) 35
2.4.5 QM9 Subsets: QM91k and QM960k 36
2.4.6 Experimental Datasets 36
2.5 Dataset Comparison 39

3 IMPRESSION Generation 1 **43**
3.1 Model Architecture and Training 43
3.1.1 Kernel Ridge Regression 43
3.1.2 Chemical Environment Representation 44
3.1.3 Hyper-parameter Optimisation 45

3.1.4 Uncertainty Estimation . . . 45
3.2 Results . . . 48
3.2.1 Model Training and Summary . . . 48
3.2.2 $\delta^1 H$ Prediction . . . 49
3.2.3 $\delta^{13} C$ Prediction . . . 54
3.2.4 $\delta^{15} N$ Prediction . . . 62
3.2.5 $^1J_{CH}$ Prediction . . . 69
3.3 Conclusion . . . 74

4 IMPRESSION Generation 2 **75**
4.1 Model Architecture and Training . . . 75
4.1.1 Kaggle Competition . . . 75
4.1.2 Molecules as Graphs . . . 76
4.1.3 Graph Transformer Network (GTN) . . . 76
4.1.4 Model Training . . . 79
4.2 Results . . . 80
4.2.1 Model Training . . . 80
4.2.2 Model Accuracy Summary . . . 82
4.2.3 $\delta^1 H$ prediction . . . 86
4.2.4 $\delta^{13} C$ prediction . . . 97
4.2.5 $^1J_{CH}$ prediction . . . 107
4.2.6 Further Scalar Coupling Prediction . . . 116
4.3 Comparison to NMRShiftDB . . . 119
4.4 IMPRESSION Generation 1 vs IMPRESSION Generation 2 . . . 122
4.5 QM9 models and overfitting . . . 128
4.6 Conclusion . . . 128

5 Strychnine Prediction Task **131**
5.0.1 Uncertainty estimation . . . 133
5.1 $^1J_{CH}$ comparison . . . 135
5.2 Geometric mean of $\delta^1 H$, $\delta^{13} C$, and $^1J_{CH}$ comparison . . . 137
5.3 Inclusion of further NMR parameters . . . 140

6 IMPRESSION for Binding Affinity Prediction **147**

6.1 Predicting Binding Affinity 147

6.1.1 pChEMBL 148

6.2 Model Architecture 149

6.2.1 ECFP4 neural network reference model 149

6.2.2 IMPRESSION for molecular properties 149

6.3 Active Learning 149

6.3.1 Identification of Binders 150

6.4 Results 154

6.4.1 Training and Testing Datasets 154

6.4.2 Model Training and regression performance 156

6.4.3 Classification performance 160

6.4.4 Active learning molecule selection 161

7 Summary and Future Work **165**

7.1 Training and Testing datasets 165

7.2 Model Architecture 166

7.3 Estimation of uncertainty 167

A Dataset Structures Reference **169**

A.1 CSD and ChEMBL Structure Reference Names 169

A.2 ChEMBL Structures 173

A.3 Full Gaussian Reference 183

Bibliography **185**

List of Tables

Table **Page**

1.1 Summary of RMSE results from Shibata et al. 16

1.2 Results of NMRShiftDB testing. MAE = Mean Absolute Error. 19

2.1 Dataset size and constituent atoms summary. 40

2.2 Number of NMR parameters in each dataset for the NMR parameters of interest in this thesis. 40

3.1 Effect on prediction error of removing environments with pre-prediction variance above a cutoff value for IMPRESSION δ^1H predictions against DFT calculations for dataset 3 (DT3). (Total δ^1H environments in DT3: 5,905) 52

3.2 Effect on prediction error of removing environments with pre-prediction variance above a cutoff value for IMPRESSION δ^1H predictions against DFT calculations for dataset 5b (DT5b). (Total δ^1H environments in DT5b: 11,885) 52

3.3 $\delta^{13}C$ prediction accuracy for sets of molecules in testing dataset 5b containing different sets of nuclei 56

3.4 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 3. (Total $\delta^{13}C$ environments in DT3: 5,262) 59

3.5 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 5b. (Total $\delta^{13}C$ environments in DT5b: 9,912) 59

3.6 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{15}N$ predictions against DFT calculations for dataset 3. (Total $\delta^{15}N$ environments in DT3: 387) 64

3.7 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{15}N$ predictions against DFT calculations for dataset 5b. (Total $\delta^{15}N$ environments in DT5b: 1,285) 65

3.8 Summary of model accuracy in $\delta^{15}N$ prediction for dataset 3 (DT3) and dataset 5b (DT5b) for the IMPRESSION generation 1 models trained using dataset 4 (DT4) and a combination of dataset 4 a 5a (DT45) 67

3.9 Accuracy in NMR prediction for chemical shift, for the experimental datasets DTe1b ($\delta^{1}H$ and $\delta^{13}C$) and DTe2 ($\delta^{15}N$) for DFT, IMPRESSION trained on DT4 and IMPRESSION trained on DT45 for $\delta^{15}N$ 68

3.10 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^{1}J_{CH}$ predictions against DFT calculations for dataset 3. Total number of $^{1}J_{CH}$ environments in DT3: 5,608 72

3.11 Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^{1}J_{CH}$ predictions against DFT calculations for dataset 5b. Total number of $^{1}J_{CH}$ environments in DT5b: 10,641 72

4.1 Accuracy in $\delta^{1}H$ prediction across the three testing datasets, for the DT45 and QM960k trained models. as well as the generation 1, KRR model 87

4.2 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{1}H$ predictions against DFT calculations for dataset 3. Total $\delta^{1}H$ environments in DT3: 5905 89

4.3 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{1}H$ predictions against DFT calculations for dataset 5b. Total $\delta^{1}H$ environments in DT5b: 11,885 90

4.4 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{1}H$ predictions against DFT calculations for dataset 3. Total $\delta^{1}H$ environments in DT3: 5905 90

4.5 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{1}H$ predictions against DFT calculations for dataset 5b. Total $\delta^{1}H$ environments in DT5b: 11,885 90

4.6 Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the $\delta^1 H$ experimental test set (DTe1b). 95

4.7 Accuracy in $\delta^{13} C$ prediction across the three testing datasets, for the DT45 and QM960k trained models. as well as the generation 1, KRR model 99

4.8 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13} C$ predictions against DFT calculations for dataset 3. Total $\delta^{13} C$ environments in DT3: 5,262 100

4.9 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13} C$ predictions against DFT calculations for dataset 5b. Total $\delta^{13} C$ environments in DT5b: 9,912 100

4.10 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13} C$ predictions against DFT calculations for dataset 3. Total $\delta^{13} C$ environments in DT3: 5,262 101

4.11 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13} C$ predictions against DFT calculations for dataset 5b. Total $\delta^{13} C$ environments in DT5b: 9,912 101

4.12 Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the $\delta^{13} C$ experimental test set (DTe1b). 106

4.13 Accuracy in $^1 J_{CH}$ prediction across the three testing datasets 3 (DT3), 5b (DT5b) and QM91k, for the DT45 and QM960k trained models, as well as the generation 1, KRR model. 108

4.14 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1 J_{CH}$ predictions against DFT calculations for dataset 3. Total $^1 J_{CH}$ environments in DT3: 5,608 109

4.15 For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1 J_{CH}$ predictions against DFT calculations for dataset 5b. Total $^1 J_{CH}$ environments in DT5b: 10,641 110

4.16 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 3. Total $^1J_{CH}$ environments in DT3: 5,608 110
4.17 For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 5b. Total $^1J_{CH}$ environments in DT5b: 10,641 110
4.18 QM960k trained model predictions on testing dataset QM91k, split by molecule size. 113
4.19 QM960k trained model predictions on testing dataset 3, split by molecule size. 113
4.20 Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the ^1JCH experimental test set DTe3. 115
4.21 Accuracy comparison between recent published work and the QM960k trained model 117
4.22 Comparison between NMRShiftDB, IMPRESSION generation 1, and Impression generation 2 for $\delta^{13}C$ chemical shift. MAE = Mean Absolute Error. 120
4.23 Comparison between NMRShiftDB, IMPRESSION generation 1, and Impression generation 2 for δ^1H prediction. MAE = Mean Absolute Error. 121
4.24 Caption 122

5.1 Variance cutoff values used for each model for each parameter in the strychnine prediction task. 133

6.1 Model performance for IMPRESSION and the base model in the prediction of pChEMBL for the HSD11 and CDK2 datasets 157
6.2 AUC scores for both models trained using 1000 molecules from each dataset, tested on the remaining molecules. 161

List of Figures

Figure **Page**

1.1 Structure elucidation workflow 7

2.1 a) $^{1}J_{CH}$ Coupling constant distributions for QM9 molecules calculated with and without mixed. b) $^{1}J_{CH}$ Coupling constant distributions for QM9 molecules calculated with and without mixed, scaled according to calculated scaling factors available from reference [1]. 28

2.2 Linear regression fit between magnetic shielding tensors calculated by DFT and experimentally measured chemical shifts. RMSD values for each plot: $\delta^{1}H = 0.26$, $\delta^{13}C = 2.32$, $\delta^{15}N = 12.15$ 30

2.3 Distribution of molecule size, $\delta^{1}H$ Chemical shift and $\delta^{13}C$ Chemical shift values in the experimental dataset 1a. 37

2.4 Distribution of molecule size, $\delta^{1}H$ Chemical shift and $\delta^{13}C$ Chemical shift values in the experimental dataset 1b. 37

2.5 Distribution of molecule size and $\delta^{15}N$ Chemical shift values in the experimental dataset. 38

2.6 Distribution of molecule size and $^{1}J_{CH}$ Coupling constant values in the experimental dataset. 39

2.7 Distribution of NMR Parameter values in the DFT calculated datasets. 41

2.8 Distribution of molecule size in the DFT calculated datasets, size includes H atoms. . 42

3.1 Illustrative plots of different situations and their potential effect on both pre-prediction variance and prediction error. a) Low variance with low error. b) High variance with low error. c) Low variance with high error. d) High variance with high error. 47

3.2 Mean absolute error in NMR parameter prediction, relative to the range of values for that NMR parameter, for the models trained using dataset 4 (DT4), for parameters δ^1H, $\delta^{13}C$, $\delta^{15}N$, and $^1J_{CH}$, for both dataset 3 (DT3) and dataset 5b (DT5b). 49

3.3 a) IMPRESSION predicted and DFT calculated δ^1H for dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 0.24 ppm MAE, 0.39 ppm RMSD and 4.27 ppm MaxE, DT5b fit statistics: 0.34 ppm MAE, 0.54 ppm RMSD, 8.78 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated δ^1H for the DT3 and DT5b testing sets, 12 (DT3) and 50 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4). Structures responsible for the outlying values in (a) also depicted in 2D (IDs: CHEMBL154357, H1 and CHEMBL6320, H5). 50

3.4 a) Error in predicted δ^1H for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated δ^1H for DT3 and DT5b, with variance values highlighted. Results for models trained using dataset 4 (DT4) 51

3.5 a) Error distributions for both IMPRESSION predicted and DFT calculated δ^1H relative to the experimentally measured values for experimental dataset 1b (DTe1b). DTe1b fit statistics: 0.44 ppm MAE, 0.61 ppm RMSD, 2.61 ppm MaxE. b) IMPRESSION predicted and experimentally measured δ^1H for DTe1b with variance highlighted. Results for models trained using dataset 4 (DT4) 53

3.6 a) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{13}C$ for the DT3 and DT5b testing sets, 1 (DT3) and 97 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4) 55

3.7 a) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for molecules containing only H/C/N/O/F atoms in dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 4.62 ppm MAE, 8.97 ppm RMSD, 151.3 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{13}C$ for molecules containing only H/C/N/O/F atoms in the DT3 and DT5b testing sets, 44 (DT3) and 109 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4) 57

3.8 a) Error in predicted $\delta^{13}C$ for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for the DT3 and DT5b testing sets, with variance values highlighted. DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. Models trained using dataset 4 (DT4) 58

3.9 (a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{13}C$ relative to the experimentally measured values in experimental dataset 1b (DTe1b). (b) IMPRESSION predicted and experimentally measured $\delta^{13}C$ for DTe1b with variance highlighted. Fit statistics for DTe1b: 4.76 ppm MAE, 6.82 RMSD, 35.0 ppm MaxE. Models trained using dataset 4 (DT4) 60

3.10 (a) error distributions for IMPRESSION predicted $\delta^{13}C$ relative to DFT calculated and experimentally measured values in experimental dataset 1b (DTe1b). (b) IMPRESSION predicted and DFT calculated $\delta^{13}C$ against experimentally measured $\delta^{13}C$ for DTe1b with variance highlighted. Fit statistics for ML to DTe1b: 4.76 ppm MAE, 6.82 ppm RMSD, 35.0 ppm MaxE. Fit statistics for DFT to DTe1b: 2.18 ppm MAE, 2.80 ppm RMSD, 15.9 ppm MaxE. Models trained using dataset 4 (DT4) 61

3.11 a) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for dataset 3 (DT3) and dataset 5b (DT5b). Fit statistics for DT3: 11.4 ppm MAE, 15.9 ppm RMSD, 67.9 ppm MaxE, Fit statistics for DT5b: 12.1 ppm MAE, 18.5 ppm RMSD, 216.5 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, 2 DT5b values excluded from graph for clarity. Models trained using dataset 4 (DT4) 62

3.12 a) Error in predicted $\delta^{15}N$ for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 11.4 ppm MAE, 15.9 ppm RMSD, 67.9 ppm MaxE, Fit statistics for DT5b: 12.1 ppm MAE, 18.5 ppm RMSD, 216.5 ppm MaxE. Models trained using dataset 4 (DT4) 64

3.13 Model trained on both dataset 4 and 5a. a) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 7.72 ppm MAE, 11.20 ppm RMSD, 77.8 MaxE, fit statistics for DT5b: 5.64 ppm MAE, 9.07 RMSD, 92.6 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets. 66

3.14 Model trained on both dataset 4 and 5a. a) Error in predicted $\delta^{15}N$ for populations with different maximum variance for the DT3 and DT5b testing sets. b) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 7.72 ppm MAE, 11.20 ppm RMSD, 77.8 MaxE, fit statistics for DT5b: 5.64 ppm MAE, 9.07 RMSD, 92.6 ppm MaxE. 66

3.15 For the model trained using dataset 4 only. a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{15}N$ relative to the experimentally measured values in experimental dataset 2 (DTe2). b) IMPRESSION predicted and experimentally measured $\delta^{15}N$ with variance highlighted. Fit statistics for DTe2: 33.04 ppm MAE, 46.43 ppm RMSD, 141.3 ppm MaxE. 68

3.16 For the model trained on both dataset 4 and 5a. a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{15}N$ relative to the experimentally measured values in experimental dataset 2 (DTe2). b) IMPRESSION predicted and experimentally measured $\delta^{15}N$ with variance highlighted. Fit statistics for DTe2: 27.02 ppm MAE, 37.3 ppm RMSD, 110.7 ppm MaxE 69

3.17 a) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, 1 (DT3) and 129 (DT5b) values excluded from graph for clarity. Models trained using dataset 4 (DT4) 70

3.18 Molecules containing H,C,N,O,F atoms only. a) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.43 Hz MAE, 2.26 Hz RMSD, 19.3 Hz MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, 2 (DT3) and 159 (DT5b) values excluded from graph for clarity. Models trained using dataset 4 (DT4) 70

3.19 Linear correction applied to DT5b DFT values. a) Error in predicted $^1J_{CH}$ for populations with different maximum variance for the DT3 and DT5b testing sets. b) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. Models trained using dataset 4 (DT4) 71

3.20 (a) error distributions for both IMPRESSION predicted and DFT calculated $^1J_{CH}$ relative to the experimentally measured $^1J_{CH}$ values for experimental dataset 2 (DTe2). (b) IMPRESSION predicted and experimentally measured $^1J_{CH}$ for DTe2 with variance highlighted. Fit statistics for DTe2: 6.01 Hz MAE, 11.18 Hz RMSD, 54.3 Hz MaxE. Models trained using dataset 4 (DT4) 74

4.1 Simplified graph transformer network diagram. 77

4.2 Out of sample for dataset 3 (DT3) and in sample loss during training for models trained on datasets 4 (DT4: a), 4 and 5a combined (DT45: b), and QM960k (c) 81

4.3 Out of sample for dataset 3 (DT3) loss split by target NMR parameter during training for models trained on datasets 4 (DT4: a), 4 and 5a combined (DT45: b), and QM960k (c) 81

4.4 Comparison in model accuracy for testing datasets 3(DT3, a), 5b(DT5b, b), and QM91k(c). Bar height represents the mean absolute error as a percentage of the full range of values for that NMR parameter, each bar is annotated with the raw MAE values. The ${}^{1}J_{CH}$ and ${}^{2}J_{CC}$ bars for the model trained using QM9 in (b) are cut off for clarity, the relative MAE values are 42% and 28% respectively. 84

4.5 Mean absolute error for the worst predicted 100 environments for selected NMR parameters: $\delta^{1}H$, $\delta^{13}C$, ${}^{1}J_{CH}$, ${}^{3}J_{HH}$. Errors presented for models trained on datasets 4 and 5 combined (DT45) and QM960k, tested on datasets 3 (DT3), 5b (DT5b), and QM91k. 86

4.6 For the model trained using DT45: IMPRESSION predicted and DFT calculated $\delta^{1}H$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. DT3 fit statistics: 0.22 ppm MAE, 0.36 ppm RMSD, 8.0 ppm MaxE, DT5b fit statistics: 0.25 ppm MAE, 0.36 ppm RMSD, 6.0 ppm MaxE, QM91k fit statistics: 0.64 ppm MAE, 1.00 ppm RMSD, 6.30 ppm MaxE. 91

4.7 For the model trained using QM960k: IMPRESSION predicted and DFT calculated $\delta^{1}H$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. DT3 fit statistics: 0.67 ppm MAE, 1.00 ppm RMSD, 9.05 ppm MaxE, DT5b fit statistics: 1.65 ppm MAE, 1.97 ppm RMSD, 9.68 ppm MaxE, QM91k fit statistics: 0.06 ppm MAE, 0.09 ppm RMSD, 1.56 ppm MaxE. 91

4.8 For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{1}H$, for the DT3, DT5b (a) and the QM91k (b) testing datasets. DT3 fit statistics: 0.22 ppm MAE, 0.36 ppm RMSD, 8.0 ppm MaxE, DT5b fit statistics: 0.25 ppm MAE, 0.36 ppm RMSD, 6.0 ppm MaxE, QM91k fit statistics: 0.64 ppm MAE, 1.00 ppm RMSD, 6.30 ppm MaxE. 92

4.9 For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{1}H$, for the DT3, DT5b (a) and the QM91k (b) testing datasets. DT3 fit statistics: 0.67 ppm MAE, 1.00 ppm RMSD, 9.05 ppm MaxE, DT5b fit statistics: 1.65 ppm MAE, 1.97 ppm RMSD, 9.68 ppm MaxE, QM91k fit statistics: 0.06 ppm MAE, 0.09 ppm RMSD, 1.56 ppm MaxE. 92

4.10 Accuracy in δ^1H prediction across the three testing datasets (DT3, DT5b, QM91k) for subsets of molecules with different size. For the model trained using DT45 (a) and the model trained using QM960k (b) 94

4.11 For the DT45 trained model predictions on the δ^1H experimental testing dataset DTe1b. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured δ^1H, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 0.39 ppm MAE, 0.58 ppm RMSD, 2.86 ppm MaxE. 96

4.12 For the QM960k trained model predictions on the δ^1H experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured δ^1H, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 0.78 ppm MAE, 1.10 ppm RMSD, 3.88 ppm MaxE. The two structures are representative of the structures which cause the similar set of errors around 7.3ppm in (b). 97

4.13 For the model trained using DT45: IMPRESSION predicted and DFT calculated $\delta^{13}C$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 4.41 ppm MAE, 6.71 RMSD, 90.82 MaxE, fit statistics for DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.13 ppm MaxE, fit statistics for QM91k: 14.5 ppm MAE, 21.9 ppm RMSD, 97.8 ppm MaxE. 102

4.14 For the model trained using QM960k: IMPRESSION predicted and DFT calculated $\delta^{13}C$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 11.4 ppm MAE, 19.6 RMSD, 120.4 MaxE, fit statistics for DT5b: 32.3 ppm MAE, 45.9 ppm RMSD, 150.5 ppm MaxE, fit statistics for QM91k: 0.89 ppm MAE, 1.29 ppm RMSD, 25.4 ppm MaxE. 102

4.15 For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{13}C$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 4.41 ppm MAE, 6.71 RMSD, 90.82 MaxE, fit statistics for DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.13 ppm MaxE, fit statistics for QM91k: 14.5 ppm MAE, 21.9 ppm RMSD, 97.8 ppm MaxE. 103

4.16 For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{13}C$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 11.4 ppm MAE, 19.6 RMSD, 120.4 MaxE, fit statistics for DT5b: 32.3 ppm MAE, 45.9 ppm RMSD, 150.5 ppm MaxE, fit statistics for QM91k: 0.89 ppm MAE, 1.29 ppm RMSD, 25.4 ppm MaxE. 103

4.17 Accuracy in $\delta^{13}C$ prediction across the three testing datasets for subsets of molecules with different size, for the models trained using DT45 (a), and QM960k (b). 104

4.18 For the DT45 trained model predictions on the $\delta^{13}C$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $\delta^{13}C$, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 3.76 ppm MAE, 5.25 ppm RMSD, 25.45 ppm MaxE. 106

4.19 For the DT45 trained model predictions on the $\delta^{13}C$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $\delta^{13}C$, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 7.15 ppm MAE, 11.9 ppm RMSD, 57.3 ppm MaxE 107

4.20 For the model trained using DT45: IMPRESSION predicted and DFT calculated $^1J_{CH}$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, fit statistics for DT5b: 2.92 Hz MAE, 3.82 Hz RMSD, 23.4 Hz MaxE, fit statistics for QM91k: 7.01 Hz MAE, 8.69 Hz RMSD, 45.2 Hz MaxE 111

4.21 For the model trained using QM960k: IMPRESSION predicted and DFT calculated $^1J_{CH}$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 7.51 Hz MAE, 10.5 Hz RMSD, 43.0 Hz MaxE, fit statistics for DT5b: 26.7 Hz MAE, 31.6 Hz RMSD, 69.7 Hz MaxE, fit statistics for QM91k: 0.54 Hz MAE, 0.77 Hz RMSD, 8.72 Hz MaxE 111

4.22 For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated $^1J_{CH}$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, fit statistics for DT5b: 2.92 Hz MAE, 3.82 Hz RMSD, 23.4 Hz MaxE, fit statistics for QM91k: 7.01 Hz MAE, 8.69 Hz RMSD, 45.2 Hz MaxE 112

4.23 For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated $^1J_{CH}$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 7.51 Hz MAE, 10.5 Hz RMSD, 43.0 Hz MaxE, fit statistics for DT5b: 26.7 Hz MAE, 31.6 Hz RMSD, 69.7 Hz MaxE, fit statistics for QM91k: 0.54 Hz MAE, 0.77 Hz RMSD, 8.72 Hz MaxE 112

4.24 Accuracy in $^1J_{CH}$ prediction across the three testing datasets for subsets of molecules with different size, for the model trained using DT45 (a) and QM960k (b). 114

4.25 For the DT45 trained model predictions on the $^1J_{CH}$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $^1J_{CH}$, with pre-prediction variance highlighted (b). Fit statistics for DTe3: 6.69 Hz MAE, 10.6 Hz RMSD, 73.2 Hz MaxE 115

4.26 For the QM960k trained model predictions on the $^1J_{CH}$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $^1J_{CH}$, with pre-prediction variance highlighted (b). Fit statistics for DTe3: 6.50 Hz MAE, 10.1 Hz RMSD, 60.8 Hz MaxE. 116

4.27 Comparison in model accuracy for testing datasets 3(a), 5b(b), QM91k(c). Bar height represents the mean absolute error as a percentage of the full range of values for that NMR parameter, each bar is annotated with the raw MAE values. 118

4.28 Comparison between IMPRESSION generation 1 trained using DT4 and IMPRESSION generation 2 trained using DT45, in terms of the correlation between the mean absolute error and the pre-prediction variance. For three NMR parameters against both testing datasets: δ^1H for DT3 (a), δ^1H for DT5b (b), $\delta^{13}C$ for DT3 (c), $\delta^{13}C$ for DT5b (d), $^1J_{CH}$ for DT3 (e), $^1J_{CH}$ for DT5b (f). 125

4.29 Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for $\delta^1 H$ prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 0.24 ppm MAE, 0.39 ppm RMSD, 4.27 ppm MaxE, DT5b: 0.34 ppm MAE, 0.54 ppm RMSD, 8.78 ppm MaxE. Fit statistics for Generation 2, DT3: 0.22 ppm MAE, 0.36 ppm RMSD, 8.01 ppm MaxE, DT5b: 0.27 ppm MAE, 0.36 ppm RMSD, 5.96 ppm MaxE. 126

4.30 Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for $\delta^{13} C$ prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. Fit statistics for Generation 2, DT3: 4.41 ppm MAE, 6.71 ppm RMSD, 90.12 ppm MaxE, DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.1 ppm MaxE. 127

4.31 Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for ${}^1 J_{CH}$ prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. Fit statistics for Generation 2, DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, DT5b: 2.92 Hz MAE, 3.8 Hz RMSD, 23.4 Hz MaxE. 127

5.1 The structure of the natural occurring structure of strychnine (1), along with 12 energetically viable diastereomers (2-13). 134

5.2 Mean absolute error between experimentally measured ${}^1 J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by mean absolute error in DFT prediction 136

5.3 Mean absolute error, adjusted using a softmin function, between experimentally measured ${}^1 J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by softmin of the mean absolute error in DFT prediction 137

5.4 Geometric mean across the mean absolute error between experimentally measured δ^1H, $\delta^{13}C$, and $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by mean absolute error in DFT prediction 139

5.5 Geometric mean across the mean absolute error, adjusted using a softmin function, between experimentally measured δ^1H, $\delta^{13}C$, and $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by softmin of the mean absolute error in DFT prediction 139

5.6 For the NMR parameters calculated by DFT. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures. 141

5.7 For the NMR parameters predicted by the generation 1 model trained using dataset 4. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures. 142

5.8 For the NMR parameters predicted by the generation 2 model trained using datasets 4 and 5. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures. 142

5.9 For the NMR parameters predicted by the generation 2 model trained using dataset QM960k. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures. 143

5.10 Score metric between experimentally measured NMR parameters for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green, using the MAE in $\delta^{13}C$ prediction), impression generation 2 trained on DT45 (labeled DT45, pink, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{3}J_{HH}$), impression generation 2 trained on QM960k (labeled QM960k, yellow, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{2}J_{HH}$), and DFT (labeled DFT, black, using the geometric mean across $\delta^{13}C$ and $^{1}J_{CH}$) for all structures. Structures ordered by mean absolute error in DFT prediction 145

5.11 Relative softmin populations between experimentally measured NMR parameters for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green, using the MAE in $\delta^{13}C$ prediction), impression generation 2 trained on DT45 (labeled DT45, pink, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{3}J_{HH}$), impression generation 2 trained on QM960k (labeled QM960k, yellow, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{2}J_{HH}$), and DFT (labeled DFT, black, using the geometric mean across $\delta^{13}C$ and $^{1}J_{CH}$) for all structures. Structures ordered by mean absolute error in DFT prediction 146

6.1 The 5 best binding molecules for the HSD11 target, as ranked by pChEMBL value. Molecule IDs: CHEMBL1098145 CHEMBL1096451 CHEMBL1098130 CHEMBL1096870 CHEMBL1098131 152

6.2 The 5 best binding molecules for the CDK2 target, as ranked by pChEMBL value. Molecule IDs: CHEMBL462385 CHEMBL191003 CHEMBL364370 CHEMBL184510 CHEMBL317703 153

6.3 Plots of the tanimoto similarity for HSD11, CDK2 and random datasets from CHEMBL155

6.4 Distribution of pChEMBL values (a), and distribution of molecules sizes (b) for the HSD11 and CDK2 datasets 155

6.5 Out of sample learning curves for the IMPRESSION model (a) and ECFP4 neural network (b), for datasets HSD11 and CDK2. The out of sample error is the mean absolute error in prediction of pChEMBL for molecules not in the training dataset (1000 molecules). 156

6.6 Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the HSD11 dataset. Errors displayed as error distributions (a) and scatter plots (b). Fit statistics for IMPRESSION: 0.78 MAE, 0.99 RMSD, 3.38 MaxE, fit statistics for ECFP4: 1.05 MAE, 1.29 RMSD, 4.39 MaxE. 158
6.7 Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the HSD11 dataset. Errors displayed as 2D Histograms. Fit statistics for IMPRESSION: 0.78 MAE, 0.99 RMSD, 3.38 MaxE, fit statistics for ECFP4: 1.05 MAE, 1.29 RMSD, 4.39 MaxE. 158
6.8 Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the CDK2 dataset. Errors displayed as error distributions (a) and scatter plots (b). Fit statistics for IMPRESSION: 0.75 MAE, 1.29 RMSD, 4.39 MaxE, fit statistics for ECFP4: 1.13 MAE, 1.41 RMSD, 4.29 MaxE. 159
6.9 Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the CDK2 dataset. Errors displayed as 2D Histograms. Fit statistics for IMPRESSION: 0.75 MAE, 1.29 RMSD, 4.39 MaxE, fit statistics for ECFP4: 1.13 MAE, 1.41 RMSD, 4.29 MaxE. 159
6.10 Receiver operating characteristic (ROC) plots for the IMPRESSION and ECFP4 models, for the HSD11 (a) and CDK2 (b) datasets. 160
6.11 AUC at each selection round for the IMPRESSION (a) and ECFP4 (b) models, for each selection scheme, for the HSD11 dataset. Select schemes: F1(low), F2(high), F3(range), F4(distribution), F5(inverse) 162
6.12 AUC at each selection round for the IMPRESSION (a) and ECFP4 (b) models, for each selection scheme, for the CDK2 dataset. Select schemes: F1(low), F2(high), F3(range), F4(distribution), F5(inverse) 163

CHAPTER 1

INTRODUCTION

1.1 NMR Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy is used extensively in multiple scientific fields, medicine (primarily through magnetic resonance imagery; MRI) and some industrial processes.

Likely the best known use of NMR to the general public is the MRI scan, a common diagnostic tool which generates a map of the water and fat in the body, distinguishable due to the relative difference hydrogen NMR signals which arise from the difference in water content in various parts of the body. One of the key benefits of the MRI scan is the fact that it is non-invasive, this aspect of NMR spectroscopy in general has made it a popular technique in several industrial applications, including the analysis of flow in oil pipelines [2], imaging of solid rocket fuel without disturbing the packing [3] and the imagine of internal features in wood [4]. [5]

Specifically in chemistry, the applications of NMR spectroscopy are still wide-ranging. NMR has been used to improve detection Fentanyl in cocaine samples [6], in reaction monitoring [7–9], and protein binding in drug discovery [10–12].

1.1.1 NMR spectroscopy in structure elucidation

NMR spectroscopy is central to the elucidation of molecular structures in solution[13–15], and the accurate prediction of NMR parameters plays a key role in modern structure elucidation techniques [16–18]. Predicted NMR parameters allow the construction of multiple theoretical

NMR spectra (or derived information from such a spectrum) from known structures. Matching the NMR parameters from the real NMR spectrum to one of the constructed sets of parameters identifies the unknown structure as that which was used to produced the predicted NMR parameters.

The measurement of NMR spectra yields important atom and atom-pair based properties: chemical shifts and scalar coupling constants. Chemical shift is the resonant frequency of a nucleus relative to a standard in a constant magnetic field. More specifically it is related to the Larmor procession frequency of the magnetic moment of a nucleus in a static magnetic field, the chemical shift being the difference between the measured frequency and a reference frequency, divided by the frequency of the instrument in which the measurement was made. For Carbon and Hydrogen nuclei the most commonly used reference frequency are those measured for tetramethylsilane (TMS). Chemical shift is given in parts per million for convenience (ppm , δ), as the frequency differences have units of Hz, whereas the spectrometer frequency will be of the order of MHz. The chemical shift of a particular nucleus is affected by the arrangement of electrons in the molecule, which provide a shielding effect acting against the external magnetic field, and as such is highly sensitive to the 3-dimensional arrangement of the molecule.

Indirect or scalar coupling between nuclear spins of nuclei connected by bonds leads to the appearance of multiple resonant frequencies for a single nucleus. This is the result of limited combinations of accessible spin states between the two nuclei. The magnitude of this splitting, visible in certain measured NMR spectra, is referred to as a scalar coupling constant, or J coupling constant, measured in Hz. Scalar coupling constants can be measured for nuclei connected by any number of bonds, but 1, 2 and 3 bonds couplings are the most commonly used. Coupling constants can be measured for the same nuclei (homonuclear) or different nuclei (heteronuclear), and the common notation ${}^{n}J_{XY}$ is used where n is the number of bonds connecting the two nuclei X and Y. E.g. a 3 bond proton-carbon coupling would be described as a ${}^{3}J_{CH}$ coupling. Like chemical shifts, coupling constants are sensitive to the 3-dimensional arrangement of electrons in the surrounding molecule.

1.1.2 NMR parameters and molecular structure

NMR parameters of nuclei are intrinsically linked to the composition and conformation of the surrounding molecule. Chemical shifts take different values in different functional groups [19],

for example methyl protons on carbon atoms typically have chemical shifts around 0.9 ppm, whereas aromatic protons typically have chemical shifts between 6.5 ppm and 8.2 ppm. ^{13}C chemical shifts in methyl carbons, as part of an alkyl group, typically have values around 10-30 ppm, whereas in aromatic carbons the chemical shift is typically around 100-160 ppm.

Scalar coupling constants display similar relationships with structural features, with aromatic or alkenyl $^{1}J_{CH}$ coupling constants typically taking values of 155-170 Hz, compared with 155-170 Hz for alkyl $^{1}J_{CH}$ coupling constants, and 240-250 Hz for alkyne $^{1}J_{CH}$ coupling constants. Similarly aromatic $^{3}J_{HH}$ coupling constants typically take values of 6-10 Hz, Alkene $^{3}J_{HH}$ take values of 6-12 Hz (cis) or 12-18 Hz (trans).

Due to the variety and complexity of chemical structures the identification of functional groups from single chemical shifts or coupling constants is rarely possible, but the important point is that chemical shift has a fixed relationship with chemical structure; two ^{13}C atoms in identical molecules measured under reasonably similar conditions will give the same chemical shift value. The complex but reliable relationship between NMR parameters and the 3-dimensional structure of molecules creates the ability to predict the NMR parameter of a given atom (chemical shift) or pair of atoms (scalar coupling constant) given the relevant structural information.

1.2 Computational NMR

The methods used to compute NMR parameters from structural information can be as simple as a linear equation (the Karplus equation[20]), or as complex as the evaluation of the Schrodinger equation in density functional theory calculations [21]. A chemical structure can be described mathematically through different features such as Cartesian coordinates, inter-atomic distances, or atomic numbers. These features can then be used to calculate some target value, such as the chemical shift or scalar coupling constant for a given atom or pair of atoms.

1.2.1 Empirical equations

The Karplus equation [20] is one of the most popular and successful NMR-based empirical equations, which relates the value of the $^{3}J_{HH}$ coupling in vicinal protons to the dihedral angle between them.

$$^3J_{HH} = A + B\cos\phi + \cos 2\phi \tag{1.1}$$

Where A, B, and C are constants given values 4.22 Hz, -0.5 Hz, and 4.5 Hz respectively in the original work. ϕ is the dihedral angle. These constants are reported for a bond length of 1.543Åbetween sp^3 hybridised carbons, and an average energy (δE) of 9 e.v. In this sense the original karplus relation is not very general, however extensions to the original equation have successfully expanded its generality to substituted ethanes [22, 23], and with greater accuracy [24–26]. Similar relationships have been published for $^1J_{CH}$ [27–29], $^2J_{CH}$[30] and $^3J_{CH}$[30, 31] couplings.

Similar methods for chemical shifts come in the form of additivity rules for δ^1H [32] and $\delta^{13}C$ [33]. These work on the principle of assigning a base value to a chemical shift in a given substructure, then applying a series of additive rules for substitutions within that substructure. The HOSE (Hierarchically Ordered Spherical Description of Environment) code and associated algorithms [34] can be seen as either the furthest expansion of empirical NMR prediction, or the most simplified version of a machine learning prediction model. Designed for the prediction of $\delta^{13}C$, HOSE codes describe a chemical environment through a series of concentric spheres. ^{13}C environments can be matched to environments with known chemical shift values through these HOSE codes.

Empirical equations provide fast and accurate estimations of NMR parameters, however even in the most complex equations the accuracy is restricted to a very limited region of chemical space. Efforts to expand the generality of empirical equations invariably tend towards the production of large numbers of equations, one for each new type of environment.

1.2.2 Density Functional Theory

NMR calculation algorithms which utilise density functional theory (DFT) are far more accurate and more general than any empirical equation, though this accuracy comes at a much higher computational cost. DFT calculations go far beyond the scope of empirical equations in terms of atomic environment features to work with the density of electrons at each point in the molecule, though this is still derived from the 3D atom coordinates and types. The DFT calculation primarily consists of a functional to approximate the true electron density function, and a basis set of wavefunctions to approximate the true molecular orbitals.

There is an approximate hierarchy of functionals in terms of their chemical accuracy [35], with less accurate functionals utilising more approximations in an attempt to reduce computational cost. The relationship between functional accuracy and cost is complicated by the vast and complex cancellation of errors which occurs through the the use of multiple approximations in the same calculation. Therefore choice of functional for a given calculation can be difficult, and often made on subjective grounds. Ideally benchmark work can be performed to test a range of functionals on a similar problem with known target results, such as was available in this work [1]. Functionals are named arbitrarily, with names arising variously from abbreviations of author names, or titles of papers, or features of the functionals themselves. For example the name of one of the most popular DFT functionals, B3LYP [36][37], is short for Becke, 3-parameter, Lee–Yang–Parr, where Becke, Lee, Yang and Parr are the surnames of authors of functional, and 3-parameter refers to the fact that three fitted parameters are used in the functional. The PBE functional [38] is similarly named after the three authors of the functional Perdew, Burke, and Ernzerhof.

The hierarchy of basis sets is more straightforward; larger basis sets provide a more accurate approximation of the molecular orbitals, but are more computationally expensive. The target in selecting a basis set for a calculation is to use the smallest basis set possible which still provides enough complexity to accurately model the molecular orbitals, where the required accuracy is dependent on the purpose of the calculations. Smaller basis sets can be used for geometry optimisations than would be used for energy calculations, for example. Once again benchmarking work is incredibly useful in selecting basis sets for calculations.

There are several different types of basis set, ranging from the minimal basis sets which use Slater-Type Orbitals (STO) and n Gaussian primitive functions to describe each orbital (labelled STO-nG [39]) which are the fastest but least accurate basis sets, to the correlation consistent basis sets [40] designed for Post-Hartree Fock methods (and the similar polarisation consistent basis sets for DFT [41]) which are highly accurate but extremely computationally expensive as they converge towards the complete basis set limit. Existing in a cost-benefit region between these two types of basis set are the split-valence or Pople basis sets[42], identified in benchmarking work as providing an optimal trade-off between computational expense and accuracy for NMR prediction across several parameters [1].

Standard notation for the naming of Pople basis sets is of the form X-YZg[42] where Z is

the number of primitive gaussian functions comprising each core atomic orbital basis function, Y and Z give the number of primitive gaussian functions which form the basis functions for the valence orbitals. If there are two numbers, each valence orbital is comprised of two basis functions each, termed a double-zeta basis set. Triple-zeta and quadruple-zeta basis sets are also common. Additional polarisation and diffuse functions can be specified in brackets after the X-YZg notation, for example the notation 6-31G(d,p) would be the 6-31G basis set supplemented by one set of d functions on heavy atoms, and one set of p functions on hydrogens.

NMR magnetic shielding tensors are calculated through DFT via the gauge-independent atomic orbital framework (GIAO) [43–45], which calculates the components (9 components accounting for interactions between x,y, and z components of the magnetic field and magnetic moment of the nucleus) of the shielding tensor from the electronic energy of the molecule, external magnetic field and magnetic moment of the nucleus. The isotropic shielding, commonly used to calculate chemical shift, is defined as one-third of the trace of the shielding tensor.

The scalar coupling constant is calculated as the sum of several components: the Fermi contact term, the paramagnetic spin-orbit term, the diamagnetic spin-orbit term (DSO), and the spin-dipolar term. In general, the Fermi contact term dominates, and so often this term is used alone as the scalar coupling constant as this saves computational expense. All terms were computed and used for DFT calculations in this project. The only notable exception to this is couplings involving ^{19}F nuclei, which are rarely of interest. [46, 47]

One of the most common computational chemistry software packages available is the Gaussian (09 [48] or 16 [49]) software package. Most NMR DFT data available in literature, and all data produced for this report, are calculated using this software. Several keywords can be used in the NMR command line for gaussian NMR calculations. 'Tight' or 'VeryTight' refer to the optimisation threshold for the SCF calculation, 'Tight' sets the convergence threshold at 1×10^{-6} Hartree, 'Verytight' sets this as 2×10^{-9} Hartree, often the stricter 'VeryTight' criteria is used to ensure the structure is fully optimised. These energy values are approximations, the convergence is assessed in terms of both force and displacement. 'Tight' is the default value [50]. 'Fine' or 'Ultrafine' refer to the density of the integration grids in the optimisation, increasing the density through the 'Ultrafine' keyword allows further optimisation of the structure as under the 'Fine' integral grids the actual minimum may lie between points on the grid. The default grid for Gaussian09 calculations was 'Fine' which uses a grid with 75,302 points, in Gaussian16 the

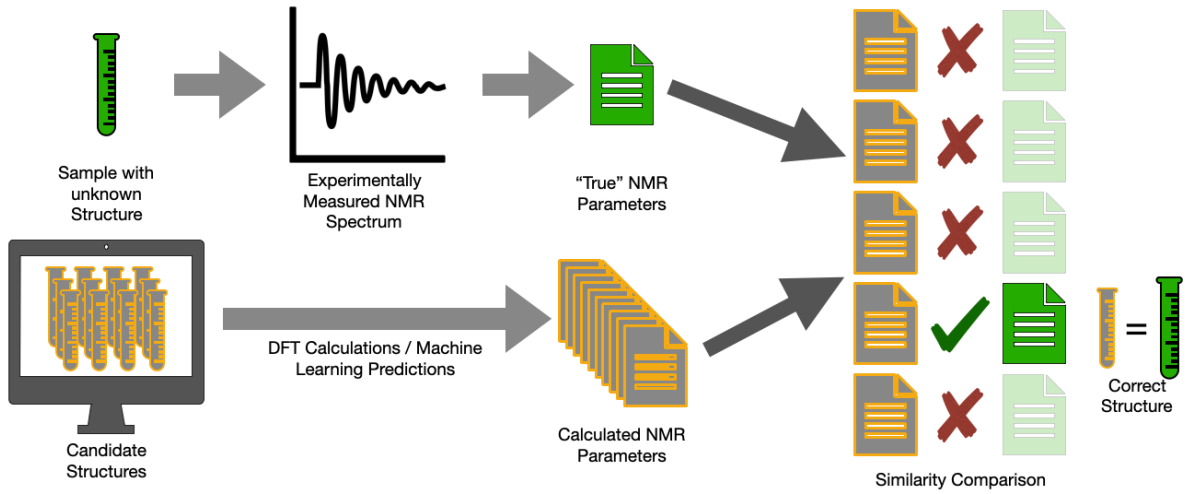

Figure 1.1: Structure elucidation workflow

default is 'Ultrafine' which raises this number to 99,590.

1.2.3 Computational NMR and Structure elucidation

The use of calculated NMR parameters in structural elucidation problems takes many forms, but for the purposes of this work it is useful to lay out one particular example. A compound for which some element of the 3D structure is unknown is synthesised or obtained, and the necessary NMR spectra measured to obtain a set of NMR parameters for the compound in solution. A set of candidate structures is produced which, in the ideal case, contains the correct structure. A Computational method is used to generate the NMR parameters for the set of candidate structures, and these are compared against the experimentally obtained values. Molecules with computationally derived NMR parameters which closely match those obtained from the real NMR spectra are more likely to be the true structure of the molecule. The methods for generating candidate structures (such as torsional angle searching and molecular mechanics [51, 52], predicting NMR parameters (discussed in this section), and comparing them to the experimental values [53, 54] are numerous and the optimal set of methods to use varies by application.

Limiting factors in this process are the accuracy of the predicted NMR parameters relative to the difference in those NMR parameters between each candidate structure, and the computational cost of obtaining those predictions. Inaccurate NMR parameters can lead to an incorrectly chosen

structure or an inconclusive result. High computational cost can lead to the use of less accurate methods, or limiting the candidate pool of structures upon which the calculations are carried out. Limiting the set of candidate structures can result in an incorrectly assigned structure, as the true structure may be inadvertently excluded from the comparison.

1.3 Artificial intelligence and Machine learning

Artificial intelligence (AI) is an extremely broad term used to refer to the recreation of human cognitive process in machines. Machine learning is a technique used to achieve outcomes desirable in AI systems, such as the detection of properties in datasets and prediction of outcomes based on input information. Machine learning is broadly split into two categories: supervised and unsupervised learning. Supervised is a term used to refer to the presence (or absence) of labelled data. In supervised learning (the type of learning used in this work) the training data is provided with labels, which are the information the machine will be expected to predict in the testing phase. Supervised learning is common in developing machines to predict values or properties.

The applications of supervised machine learning are wide-ranging, and have been used to perform tasks as diverse as the prediction of house prices [55], the analysis of tweets [56] and the classification of flowers [57].

Further to the distinction between supervised and unsupervised, most machine learning models can be described as being either a classification or a regression machine. Classification machines assign one of a discrete number of labels in each case, whereas regression machines assign a floating point value. The prediction of NMR parameters is a supervised regression machine learning task.

Even within supervised regression tasks there are as many different machine learning algorithms as there are applications, however two of the most popular categories are kernel methods and neural networks. All of the machine learning models discussed and generated in this work fall into one of these two categories.

Kernel methods define a distance between all input data points in the training dataset, two of the most popular kernel distances are laplacian and gaussian kernels. This matrix of distances is used to define a mapping between the distance from an input point to all other input points and the correct value assigned to the focus input point. New predictions are then made by calculating the distance from the new input point to all existing points, then via regression the predicted

value of the new point is determined. The key advantage of kernel methods is in dimensionality reduction, as an input vector of arbitrarily many feature dimensions is converted to a 1-D vector of distances. This drastically simplifies the regression task, making it much faster to perform. The disadvantage here is that as the size of the training dataset increases, so does the size of the model, which limits the improvement of the model through new training data.

Neural network is a general term for machine learning algorithms which use connected networks of nodes. Data is passed through a network, which can contain any number of nodes, and they can be configured in nearly any way in terms of their connections with each other. In one of the most simple cases, an input point could consist of several numbers, which are passed to different input nodes in the network, these are combined via addition into a single number, which is the output value. For the case where there are 3 input features, this network would have 3 input nodes, and a single output node. In order for this network to learn, the connections between the input nodes and output node will have variable weights, which affect the value of the output node. This could also very simple be described by the equations:

$$\text{Output} = I_0 W_0 + I_1 W_1 + I_2 W_2 \tag{1.2}$$

Where I_n are the input features, and W_n are the weights. In the training process for this network, a series of inputs, each with 3 features, would be used to calculate the predicted value using the current values of the weights. In each iteration, the difference between the predicted value and the true value will be calculated, this is called the loss in the model. There are many different loss functions used in machine learning but one of the most simple is the L_1 loss or absolute loss, calculated by:

$$\text{loss} = |\,\text{Output} - \text{Truth}\,| \tag{1.3}$$

Finally in order for the model to learn, the weights must be adjusted such that this loss value decreases. This is done via an optimisation algorithm, again of which there are many. Gradient descent is one of the most popular and easy to understand optimisation algorithms. Gradient descent works by adjusting the weights in the direction of steepest negative gradient across the set of weights.

If there is a simple relationship between the three input values which will yield the required output value, then this solution should be found in a reasonably short number of training steps, given appropriate training data. To solve more complex problems which have more complex

solutions, larger networks are used, with varying architectures and mathematical functions in the nodes. To obtain the correct weights in these more complex networks usually requires many more input examples. This increase in complexity and training data can however yield incredibly accurate predictions. The term deep learning has been used to describe networks with many layers of nodes, rather than just the two layers described here.

1.4 Machine learning in NMR

The issues apparent in highly expensive DFT calculations or inaccurate and restricted empirical equations, alongside the enormous popularity and increasing availability of machine learning (ML) methods, has given rise to a growing field in ML NMR prediction.

Machine learning is a sub-genre of artificial intelligence which refers to algorithms which improve through experience or example data. Machine learning can be supervised or unsupervised, this refers to whether the correct answer to a given problem is shown to the algorithm during training. All of the ML methods discussed and presented in this work are supervised algorithms, in which training examples (chemical structures) are shown to the algorithm alongside the correct answer (target value: chemical shift, etc).

The implementation, success and limitations of recent publications in ML NMR prediction are discussed here, covering a range of ML techniques. Machine learning techniques for NMR prediction have been developed for decades, however a model capable of predicting properties for 3-dimensional structures (as opposed to 2D structures, or smiles strings) had not been published until 2018, (Paruzzo et al. [58]). NMR parameters depend on the 3D arrangement of electrons, and so models which take into account 3-dimensional data will be more accurate, and such models would theoretically be able to distinguish between environments which differ in ways only apparent in 3 dimensions, such as diastereomers. Recent advances in neural network architectures, most notably the advent of graph based neural networks, have also provided improvements in accuracy. One of the most recent machine learning models for chemical shift prediction (Jonas et al. [59]) still relies upon 2 dimensional structural data, but through the use of a graphical neural network (GNN) and a large experimental training set, outperforms density functional theory calculations for a limited set of test structures.

The QM9 dataset is a very popular dataset in machine learning applications [60–63] as it contains a complete exploration of chemical space for H/C/N/O/F atoms in configurations up to

9 heavy atoms (non-H). The 133,885 molecule QM9 dataset is based on smiles strings taken from the GDB17 chemical universe [64], a construction of smiles strings representing all 166 billion organic small molecules with up to 17 atoms containing only C/N/O/F/S/Cl/Br/I/At atoms. QM9 has become the standard benchmark for machine learning in chemistry, and validation scores against QM9 are regarded generally as rigorous and transferable. One of the first NMR prediction model trained using QM9 is presented by Gupta et al ([65] in their 2021 publication using a kernel ridge regression model to predict $\delta^{13}C$.

Machine learning models which predict scalar coupling constants are rare, and in fact the publication which arose as a part of this thesis included one of the first ML models to predict coupling constant in 3D-molecules, in that case $^{1}J_{CH}$ [66]. More recently Shibata et al [67] have presented a set of machine learning models using the popular LightGBM framework [68] and the QM9 dataset which are capable of predicting 8 different scalar coupling constants.

These key publications are discussed in further detail below, and together present the advent of 3D NMR prediction [58], the most recent developments in 2D NMR prediction [59], the application of QM9 to train and validate models for chemical shift [65] and scalar coupling[67] prediction. Furthermore these publications present the benchmark and target accuracy against which the work in this thesis can be compared.

1.4.1 Paruzzo et al 2018: Chemical shifts in molecular solids by machine learning

The work by Paruzzo et al [58] in 2018 presented a machine learning model to predict chemical shifts for solid state NMR. Due to the very limited amount of published experimental data, training a model to directly predict experimental values is not practical. Techniques to accurately calculate chemical shifts in solid-state NMR through DFT calculations however allow the development of a training set with at least good agreement with the experimental values. They also note in this work that the availability of experimental data would bring with it further challenges as these reported values also depend on the dynamics and conditions of a real system, introducing ambiguity into the structure to target value relationship. The DFT calculations used differ from those described in this thesis, as they relate specifically to the calculation of solid-state NMR chemical shifts. To calculate the chemical shifts the Gauge Including Projector Augmented Waves method (GIPAW) [69] is used, as opposed to the gauge-independent atomic orbital framework

(GIAO) [43–45] used for solution state chemical shifts.

For this study, a set of 61,000 structures were obtained from the Cambridge Structural Database of X-ray crystal structures, comprising structures matching the following criteria:

- Fewer than 200 atoms.
- Containing C and H atoms.
- Possibly containing N and/or O atoms.

A random subset of 500 structures is selected from this set to act as the testing dataset. A set of 2000 structures is selected by farthest point sampling to act as the training dataset. Farthest point sampling (FPS) algorithms [70, 71] define a distance between objects (in this case a smooth overlap of atomic positions (SOAP) Kernel [72]) and select the set of objects which are least similar to each other. In this way the training set for this work was selected to evenly cover as broad a range of chemical environments as possible. Environments were removed from the training set using a cross-validation procedure: predictions were made using 40 models trained on random subsets of the full training set (in each case excluding the environments to be predicted), if the average deviation in predictions to the reference DFT calculation was greater than three times the variance across the predictions themselves, then the environment was discarded. Finally all symmetrically equivalent environments were removed from the training set. No such pruning procedures were carried out for the testing set. DFT NMR calculations were carried out using the program Quantum ESPRESSO [73], using the functional PBE [38], after a DFT geometry relaxation step. In terms of computational time cost, an estimation is given of 62-150 CPU hours for a DFT chemical shift calculation for a structure containing 86 atoms

The machine learning model is based on a gaussian process regression framework using the smooth overlap of atomic positions (SOAP) Kernel [72]. Each chemical environment is represented as a 3-dimensional superposition of gaussian functions centered on the surrounding atoms which fall within a cutoff radius. This approach is similar to the Kernel Ridge Regression framework reported in this work in Chapter 3.

The models achieve an accuracy of 0.49 ppm for ^{1}H, 4.3 ppm for ^{13}C, and 13.3 ppm for ^{15}N relative to the DFT calculated values. The reported DFT chemical shift accuracy to experiment for their method is around 0.4 ppm for ^{1}H, 2.0 ppm for ^{13}C, and 5.4 ppm for ^{15}N. This work is comparable to elements of this report in terms of the ML framework (Chapter 3) and the

training/testing structures (Chapters 3 and 4), and so suggests a reasonable benchmark accuracy of between 20% and 250% worse than the accuracy of the underlying DFT method.

1.4.2 Jonas et al 2019: Rapid prediction of NMR spectral properties with quantified uncertainty

The work by Jonas et al [59] in 2019 reports a Graphical Neural Network (GNN) [74] approach to predicting solution state ^{1}H and ^{13}C chemical shifts. Training and testing data were obtained from the nmrshiftdb2 [75] database of 43,468 molecules with experimental NMR data. Molecules were included which match the following criteria:

- Molecule contains no more than 64 atoms.
- Molecule contains only H/C/N/O/F/P/S/Cl atoms.
- Molecule must pass the 'sanitize' process in RDKit [51]

This resulted in a dataset of 32,538 molecules with an average size of 29 atoms. They note that several thousand nuclei had multiple measurements and, taking the average of the measurements as the 'True' value, they calculate a mean absolute error of 0.51 ppm for $\delta^{13}C$ and 0.09 ppm for $\delta^{1}H$. They suggest that this reflects the intrinsic error in the experimental measurement itself. Multiple values for a single nucleus were included in the training and testing datasets, however it was ensured that no molecule in the test set had the same SMILES string as any molecule in the training set. 80% of this data was used for training and 20% for testing.

In order to compare DFT calculated chemical shifts as well, they calculated these values for a set of 177 molecules which had the greatest number of independent spectral measurements in nmrshiftdb. For each structure, a conformational search was carried out using macromodel [52] to identify the most probably conformers by molecular mechanics calculated energy. Each conformer then underwent a DFT geometry optimisation using the functional B3LYP [36] and the basis set 6-31+G(d,p). The isotropic shielding tensors were calculated using the functional mPW1PW91 and basis set 6-311+G(2d,p), using a PCM solvent model with chloroform solvent. All DFT calculations were carried out using the Gaussian16 software [49]. The resulting shielding values were Boltzman weighted using the calculated molecular energies from the DFT NMR calculation. The isotropic shielding values were then converted to chemical shift through a linear

fit to experimental data. They note that calculating the fit parameters using the intended test data, rather than an independent set, may result in an over-estimation of the accuracy of their DFT method.

The reported mean absolute error of the machine learning model in predicting experimental chemical shift values is 1.43 ppm/0.97 ppm for $\delta^{13}C$ and 0.28 ppm/0.29 ppm for $\delta^{1}H$ for the 20% test set and 177 molecule subset respectively. The reported accuracy of their DFT method is 1.92 ppm MAE for $\delta^{13}C$ and 0.37 ppm MAE for $\delta^{1}H$ for the 177 molecule subset. This work is most comparable to the Graph transformer network reported in Chapter 4, although the molecules used for testing here are on average 10-20 atoms smaller and the largest molecules more than 50 atoms smaller than those used in the main test sets in this work. The accuracy presented the work by Jonas et al therefore represents a good target for the accuracy in chemical shift prediction in this work, especially if it can be achieved for a much larger variety of chemical environments.

1.4.3 Gupta et al 2021: Revving up 13C NMR shielding predictions across chemical space: Benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules

Gupta et al [65] present a Kernel Ridge Regression framework (KRR) [76] trained using the QM9 dataset [77] to predict $\delta^{13}C$. A 50,000 molecule test set was obtained at random from QM9, and up to 100,000 molecules were used for training, also randomly selected. A further validation dataset was obtained by taking 8 subsets of 25 molecules containing 10-17 heavy atoms from the GDB17 dataset, a total of 200 molecules.

Minimum energy geometries were obtained for all 134k molecules using the functional B3LYP and basis set 6-31G(2df,p). Structures were excluded which fragment during the optimisation, 3,054 in total. NMR shielding tensors were calculated using the functional mPW1PW91 and basis set 6-311+G(2d,p). Calculations were carried out using the Gaussian16 software package [49]. For all DFT calculations the 'ultrafine' integration grid was used, along with a 'VeryTight' SCF threshold (discussed above). This procedure was also followed for the 200 molecule validation dataset. Further NMR shielding values were calculated at a lower level of theory; functional B3LYP and basis set STO-3G, with geomtries optimised at the PM7 level using the MOPAC software package. The calculated ^{13}C isotropic shielding tensors were converted to chemical

shifts using a reference value calculated for tetramethylsilane (TMS).

Models were trained using Kernel Ridge Regression and one of three kernels; CM [78], SOAP [72], FCHL [79]. The model which used the FCHL kernel performed best, achieving a mean absolute error of 1.88 ppm against the 50k test set. Further accuracy was obtained by learning the difference between the lower theory DFT calculated values and the target, higher theory values, with a mean absolute error of 1.36 ppm. On the larger set of molecules this model performed much worse, with an MAE of around 3 ppm. This highlights a potentially key issue in models trained using QM9, that generalising prediction accuracy to larger molecules is not straightforward, and results in Chapter 4 will support this conclusion.

1.4.4 Shibata et al 2021: Prediction of spin–spin coupling constants with machine learning in NMR

Shibata et al [67], present a set of machine learning models to predict eight types of scalar coupling: $^{1}J_{NH}$, $^{1}J_{CH}$, $^{2}J_{HH}$, $^{2}J_{NH}$, $^{2}J_{CH}$, $^{3}J_{HH}$, $^{3}J_{CH}$, $^{3}J_{NH}$. The models use LightGBM[68], a decision tree algorithm, with a set of molecular descriptors either calculated directly or obtained through RDKit. The model uses the QM9 dataset for both training and testing, making a single 70/30 split in the dataset. This resulted in a training dataset of 59,502 molecules and a testing dataset of 25,501 molecules. The DFT calculated values were obtained from work by Bratholm et al[80], in which the structures were optimised using the functional B3LYP [36] and basis set 6-31g(2df,p). The DFT NMR coupling constants were calculated using the same functional and basis set.

The model achieved a root mean squared deviation of 1.82 Hz in the prediction of the DFT calculated $^{1}J_{CH}$, and 0.67 Hz for $^{3}J_{HH}$ on the 25k molecule test set, full results in Table 1.1[67]. Although utilising a ML architecture not used in this work, these prediction errors give a useful benchmark for the coupling constant prediction described in Chapters 3 and 4. The prediction accuracy reported here is for molecules from QM9 which are limited to 9 heavy atoms and, as observed in the work by Gupta et al [65], it would be expected that the model performance would deteriorate when tested against larger molecules.

NMR Parameter	RMSD [Hz]
$^1J_{NH}$	0.98
$^1J_{CH}$	1.82
$^2J_{HH}$	0.48
$^2J_{NH}$	0.51
$^2J_{CH}$	0.82
$^3J_{HH}$	0.67
$^3J_{CH}$	1.07
$^3J_{NH}$	0.37

Table 1.1: Summary of RMSE results from Shibata et al.

1.4.5 Summary

The publications presented above form an overview of recent work in the machine learning prediction of NMR parameters. Kernel based and neural network based methods both demonstrate the potential to provide highly accurate NMR parameters in a fraction of the computational time their underlying DFT methods take to run. Direct comparisons between the accuracy of different methods is not straightforward due to the lack of a universally accepted benchmark dataset, and the limited chemical space covered by the current most popular testing dataset (QM9). Despite this, reported accuracy in these publications provides some target and benchmark values for the prediction models presented in this report.

1.4.6 Licensed Software: ACD Labs

Several companies have also developed software to predict NMR parameters using machine learning models, especially proton and carbon chemical shifts for structures based on 2D coordinates. The most popular of these are the ACD labs NMR prediction tools for carbon and proton chemical shifts [81], published accuracy data is limited, and further testing could not be performed due to the lack of a license for the software, however the carbon NMR predictions are reported as having an accuracy of 2.9 ppm standard deviation [82]. ACD labs themselves report the accuracy of the carbon chemical shift predictor on a subset of shifts from NMRShiftDB [83] as 1.79 ppm absolute deviation and 3.22 ppm standard deviation [84]. The reported accuracy for the proton chemical shift prediction tool is 0.22 ppm standard deviation, though this has since been removed from the ACD labs website, the value is reported in a 2008 paper by Kuhn et al [85]. There are issues with the reported accuracies for these tools due to the limit variety of compounds used

in the comparison made by Meiler et al, and the fact that the prediction accuracy is based on making predictions from molecules drawn in 2D. Whilst this can provide accurate chemical shifts for many compounds, the focus of this work is on 3D prediction, and so comparisons between the methods are difficult. It is expected that models based on 3D prediction will likely forfeit some accuracy on simple molecules which are accurately depicted in 2D, in exchange for much greater accuracy on more complex molecules. Without the ability to directly input coordinates into a given model it is not possible to test where and how this affects accuracy for different subsets of molecules, but should this become feasible it would represent a very useful piece of analysis.

Furthermore, as is noted by Meiler at al. in a later published addendum to their original work referenced above [86], without knowing the compounds which form the training dataset it is not possible to evaluate the suitability of any validation dataset which may be used to test the accuracy of a prediction tool. It is of course understandable that commercial companies cannot make this information public, but this severely limits the ability to draw comparisons between the accuracy of their tools and other published work. This same issue extends to other vital factors in assessing the value of a machine learning solution to NMR prediction such as the cost of obtaining the training parameters, cost of training (and crucially retraining) the machine learning model, and the cost and speed of making predictions. All of these factors are readily comparable in the fully published models discussed in this section, but not commercial software.

1.4.7 Open Source Software: NMRShiftDB

The final NMR prediction algorithm worth noting is that provided by NMRShiftDB [83, 87]. Primarily a database of exprerimental NMR spectra and assignments, the website also offers an NMR prediction tool based on a neural network algorithm. This is the most readily available NMR prediction tool, accessible via a simple google search and drawing a molecule, or uploading a structure. As such it is a good point of reference for the outcomes in this work, due to its ease of use and availability, any method developed as a part of this work must be substantially more accurate in order to provide a benefit to the average user.

One important factor to note is the errors present in the experimental data used to produce the NMRShiftDB predictions, it is estimated that the experimental database contains around 8% errors through mis-assignments, transcription errors and incorrect structures [84]. Using DFT calculated data as in this work avoids this issue entirely. This also however makes the

comparison below and in Chapter 4 difficult, as the comparison is being made between a model trained to predict experimental values and a model trained to predict DFT values. This means there will be some inaccuracy to this comparison, however given the accuracy of the DFT method to experimental values, it is minor enough to mean the comparison is still useful.

The performance of NMRShiftDB was investigated by selecting 20 molecules at random from the CHEMBL test dataset (described in later sections) and attempting to make Carbon and Proton chemical shift predictions using the NMRShiftDB web server. As can been seen in the results in table 1.2, many of the compounds uploaded to NMRShiftDB returned an error, in some cases the specific reason was given that the molecule contained atoms the system considered invalid, it is assumed the neural network was only trained for a certain subsection of nuclei. In many other cases however simply a generic error was reported, the predictions for these molecules were attempted multiple times, and on multiple different days to rule out genuine server issues, so there must be some further criteria being enforced, or bug in the prediction code, which means these predictions are not available or possible.

The error in chemical shift prediction is generally good and comparable to the methods outlined above, however there are some larger errors, especially for molecule CHEMBL6889 which has a mean absolute error in carbon chemical shift prediction of 9.89 ppm. The performance of the NMRShiftDB prediction tool is discussed in more detail in Chapter 4, where it is also compared to the performance of the models produced as part of this work.

Molecule ID	Result	Carbon Chemical Shift MAE [ppm]	Proton Chemical Shift MAE [ppm]
CHEMBL1075841	Site Error	N/A	N/A
CHEMBL1084953	Success	2.85	0.22
CHEMBL1086530	Invalid atom(s)	N/A	N/A
CHEMBL1094672	Site Error	N/A	N/A
CHEMBL1096781	Success	4.04	0.75
CHEMBL1213982	Success	3.71	0.53
CHEMBL174668	Site Error	N/A	N/A
CHEMBL4116108	Success	1.88	1.76
CHEMBL4116148	Success	2.49	0.79
CHEMBL437851	Site Error	N/A	N/A
CHEMBL501943	Site Error	N/A	N/A
CHEMBL507540	Site Error	N/A	N/A
CHEMBL538928	Success	3.27	0.33
CHEMBL573427	Site Error	N/A	N/A
CHEMBL574221	Invalid Atom(s)	N/A	N/A
CHEMBL579584	Success	2.73	0.87
CHEMBL595793	Success	5.36	1.01
CHEMBL608847	Success	3.01	1.05
CHEMBL6225	Success	5.46	0.71
CHEMBL6889	Success	9.86	0.77

Table 1.2: Results of NMRShiftDB testing. MAE = Mean Absolute Error.

1.5 Aims and objectives

The purpose of this work is to investigate the ability of machine learning methods to replicate the accuracy of DFT calculations in the prediction of NMR parameters for small organic molecules. The underlying hypothesis of the project is that this is possible, and that the loss in accuracy in using a machine learning model over a DFT calculation is more than compensated for by a significant decrease in computational cost. Accuracy is treated in relative terms throughout; the accuracy of machine learning predicted NMR parameters are judged relative to the parameters given by the DFT method used to train the machine learning model. The accuracy of the underlying DFT method relative to experimentally measured values is also discussed, as well as the accuracy of the machine learning models to the experimental values, but this discussion is secondary to the core aims of the research. In an ideal case the DFT method chosen to calculate the NMR parameters in this work would already be exceptionally accurate and provide values

nearly indistinguishable from the experimental values, however the computational cost of such DFT calculations was not feasible as part of this research. The assumption here is that if a method could be identified which enabled the training of machine learning models which can accurately reproduce DFT calculations, such a model could be trained on any DFT method, and so improvements in affordable and highly accurate DFT calculations will inevitable filter down into the results of work such as this, improving the accuracy of the machine learning models predictions relative to experiment, but crucially having no impact on the accuracy of the model with respect to the DFT calculated values, this is expected to remain relatively consistent.

There are several smaller objectives which make up the project, not all of which were apparent at the start. The first of which is to obtain high quality training and testing datasets of DFT calculated NMR parameters, this is discussed in further detail in Chapter 2 and below. Secondly several models needed to be designed, developed and tested in order to identify an improvement on existing methods, this is again discussed further in Chapters 3 and 4. A further objective which developed as a result of the prevalence of QM9 trained models in the literature was to investigate the accuracy of QM9 trained models using the machine learning frameworks already developed as a part of this research, and evaluate the accuracy of these models on molecules outside of the QM9 dataset. Finally an extension to this research was developed in partnership with Astrazeneca, in which one of the machine learning models was adapted for the prediction of binding affinity, the objective here was to determine how easily the successful machine learning frameworks identified could be adapted to perform tasks outside of NMR parameter prediction.

1.5.1 Machine learning datasets

The cornerstone of this work, and indeed all machine learning applications, is the underlying data used for training and evaluating the model. Training datasets define the performance limit for the model in terms of the accuracy of predictions, and the space in which that accuracy will hold. The testing and validation datasets define what properties of the model can be proved, they define the space for which one can claim the model to be accurate, and to what degree.

The core purpose of this work is to demonstrate the ability to predict experimental NMR parameters using machine learning techniques. This presents a major challenge however, as high quality, reliable experimental NMR data is scarce. Obtaining a dataset on the scale required for this project was therefore not feasible, and unless the issues around reliability and consistency

could be resolved, not desirable.

The accurate prediction of experimental NMR parameters using DFT [54, 88–91] is a commonly used tool [92–95] and the accuracy of such calculations will only improve in the future. If the relationship between DFT and experimental values can therefore be assumed to be readily solvable, if not solved in some cases, then the primary purpose of the models developed in this work should be to predict the NMR parameters of a reasonable DFT method, where reasonable will be defined in section 2.2. This better isolates the scientific question being posed thus making it easier to solve, and the quality of the proposed solution easier to evaluate. The strategy in this work therefore is to develop models capable of accurately predicting the output of a reasonable DFT method, and then introduce some limited experimental data as validation of this approach.

CHAPTER 2

DATASET PRODUCTION

2.1 Dataset requirements

For the purposes of this work several datasets were developed. The development of new datasets was driven by limitations of the existing datasets or by a desired expansion of the prediction models applicability. The requirements for datasets in this work can be easily divided into the following desired characteristics: Size, Breadth, Depth, Quality, and Credibility.

2.1.1 Overfitting

An important issue in machine learning which directly impacts dataset development is the potential for overfitting. Often treated as an issue in model training, it is in reality a combination of bad dataset and model design. Overfitting is when a model is trained to make highly accurate predictions for a certain, limited set of input cases, however this prediction accuracy does not generalise beyond this set. Avoiding overfitting can be thought of as an exercise in balancing the resolution of the information a model is capable of extracting with the size and breadth of the dataset. A model which is capable of identifying only relatively simple relationships between the input features and desired outputs will not require a very large dataset. More complex models, such as those used in most modern machine learning applications, are capable of identifying extremely complex patterns in data as a result often of the large number of tunable parameters in the model itself. Such a model is naturally more likely to overfit than a simple model, and so

care must be taken to select a dataset of sufficient size and breadth so as to avoid overfitting and retain good generalisation/

2.1.2 Dataset Size

The size of the training set of is one of the biggest limiting factors in the accuracy of any model [58, 79, 96, 97]. More accurate DFT calculations are more expensive [98], and molecule size (which further increases calculation time, commonly proportional to the square of the increase in number of electrons) is an important factor in obtaining breadth in the dataset. Increasing the size of a high quality dataset can therefore take weeks to months of time. Well trained machine learning models improve their accuracy by an order of magnitude if the size of the dataset is also improved by the same order of magnitude [79], increasing the size of the dataset therefore becomes an increasingly less important factor in a models performance, the bigger the dataset has become.

2.1.3 Dataset Breadth and Depth

The selection of molecules for both the training and testing datasets dictate the accuracy of the model, the range of structures for which that accuracy holds true, and importantly the extent to which that accuracy can be demonstrated [99]. A deep dataset, i.e. one with a huge amount of information for a limited region of chemical space, will be accurate for molecules similar to those in the training set, but a broad dataset, in terms of variety in chemical composition and conformation, will be less accurate but will hold that accuracy for a much larger variety of molecules. A testing set with either of the same weaknesses will unnecessarily devalue the trained model by failing to show specificity of prediction over small changes (dataset not sufficiently deep) or the range over which the model holds accuracy (dataset not sufficiently broad). The balance between these two properties is primarily limited by the selection pool from which structures are obtained, and the method of selection from the pool.

2.1.4 Dataset Quality

Quality in this context refers to the presence of errors in the dataset. This could be an incorrectly measured or transcribed value, a missing value, atom or bond, a mistake in the structure reported alongside the data, or any of a vast array of problems which impede the ability of a model to learn

information from the dataset. Many of these issues can be avoided by calculating data from initial starting structures ourselves, rather than relying on external sources, however this shifts the burden onto the computational workflow which produces the final optimised structure coordinates and NMR parameters. Measures taken to address sources of error in the processing workflow will be discussed in section 2.3. Producing DFT datasets reduces the reliance on external data to the source of the initial structures, however this is still important. High quality sources of initial structures include public repositories [100, 101] which have their own advantages and disadvantages.

2.1.5 Dataset Credibility and bias

The credibility of a model is an often overlooked but important factor. This primarily relates to statements made about a given model in a publication, and has several contributory factors. If a given accuracy is reported for a specific model on a specific set of test molecules, the implication is that this accuracy holds for some molecules not included in this test set. The extent to which this accuracy generalises is often not discussed explicitly, however it depends on how the training and testing data were selected, and how the model was trained.

Care must be taken to avoid limiting the training and testing dataset to too narrow a chemical space (where 'narrow' can be seen in terms of the size of molecules, variety of constituent elements, complexity of structure, or any other feature), which can allow a model to train to a relatively high accuracy, which will not hold outside of this space. For example a training set consisting exclusively of molecules with Carbon and Hydrogen atoms only will potentially achieve a good accuracy for similar molecules, but would not be expected to predict to a similar accuracy the chemical shift of a carbon atom bonded (or nearby to) an oxygen or nitrogen atom. Typically this can be avoided by selecting structures at random from a suitably large and diverse source relative to the diversity of molecules in the intended application. In terms of credibility it is more important to scrutinise the breadth of the testing set, as an overly narrow training set will reveal itself given a suitable validation procedure . It is also the case that models do generalise beyond the scope of the training data provided, therefore the breadth of the training dataset is not necessarily an issue in all cases.

It is also of vital importance that correct procedures are followed to avoid data leakage during training; data leakage refers to the model gaining access during training to information in the

testing set by any means, which may produce an unrealistically accurate model (when tested on the same testing set). Whilst simple mistakes such as including test molecules in the training dataset are rarely found in published work, more subtle examples can be harder to avoid. For example, if multiple model architectures are being developed in parallel (such as in this work), it can be easy to select the best model based on its performance on the testing set. Even if best practices have been followed during the training of each model, selection based on the testing dataset performance undermines the credibility of the model, unless a further independent dataset reinforces this selection.

Bias in models is related to credibility but refers to issues with the training set. Bias in this context refers to a model being more or less accurate for a specific class of molecule or chemical environment. All models are fundamentally biased in some way, as no training set can claim to cover all of known chemical space, however bias becomes an issue if not identified. A biased training set combined with a similarly biased testing set can produce a model which appears highly general and accurate, but which will perform poorly in further application. An example of this would be a model trained on molecules with a maximum of 7 atoms, which may perform well given a testing set containing only similarly small molecules, but in a dataset containing a wider range of molecule sizes, the model will be biased towards the environments that are common in the smaller molecules, and so will potentially perform poorly on environments in the larger molecules.

2.1.6 Testing Datasets: relevance.

Most of the above criteria apply to all data gathered for the testing and training of models. There are however, differences in the aims of training and testing data which are important. Whilst it may appear desirable to select a testing set which covers the largest area of chemical space possible, this can be counter-productive to demonstrating the efficacy of a given model. Covering more chemical space requires inclusion of structures, nuclei, and parameter values which are rarely seen in the real world. If the aim of a testing set is to prove the predictive ability of a model in application, then the testing set should be as representative of the real world tasks the model is likely to be asked to perform. The relevance of the testing set to a given problem is an important quality, and one which need not always be considered for training data.

2.2 DFT NMR Calculation

In order to demonstrate the ability of a model to predict the outcome of DFT NMR calculations, the calculated DFT NMR values must be sufficiently linked to chemical space to enable suitably complex relationships to be learned by the model. Even if these relationships are incorrect in terms of the ultimately desired experimental value, a model which can learn relationships of the necessary complexity will be able to learn the correct relationship given the necessary data.

To that end a DFT method was selected based on the ability to predict several NMR parameters of interest (initially δ^1H, $\delta^{13}C$ and $^3J_{HH}$ scalar couplings) to a sufficient accuracy, within a reasonable time-scale. The chosen method was identified by (unpublished) benchmarking work performed by Claire Dickson [1]. For optimising the structures the DFT functional mPW1PW91 was used with the basis set 6-311g(d,p). For calculating the NMR parameters the functional ωb97xd was used with the same basis set [102–106]. The 'tight' optimisation criteria and ultrafine integral grids were used in the optimisations (details of which are discussed in section 1.2.2). No solvent model was used in the calculations. For machine learning prediction and DFT calculations the lack of a solvent model should make little or no difference to the qualitative outcome of the comparison. In the case of the comparison to experimental values, the choice to not use a solvent model will decrease the accuracy of the DFT calculated and machine learning predicted values relative to experiment, whilst this is not ideal, the focus of this research is on the re-production of DFT calculated NMR parameters, and so it was of equal interest whether the machine learning models could match the DFT predicted values for the compounds where experimental values were available. In order for this to be evaluated the DFT calculated values needed to be calculated in the same way as for the other datasets used for comparison. For this reason the DFT calculations for the molecules for which experimental NMR parameters were obtained were also performed without any solvent model.

The DFT calculations were performed on one of several high performance computing clusters available at the University of Bristol. The calculations were run using 8 Intel CPUs with 26GB of available RAM, the exact CPU used for each calculation varied by the cluster used.

2.2.1 'Mixed' Keyword Calculation Issue

The intention was to use the 'mixed' option to calculate the scalar coupling constants in all calculations, which improve the accuracy of these calculations by using an uncontracted basis

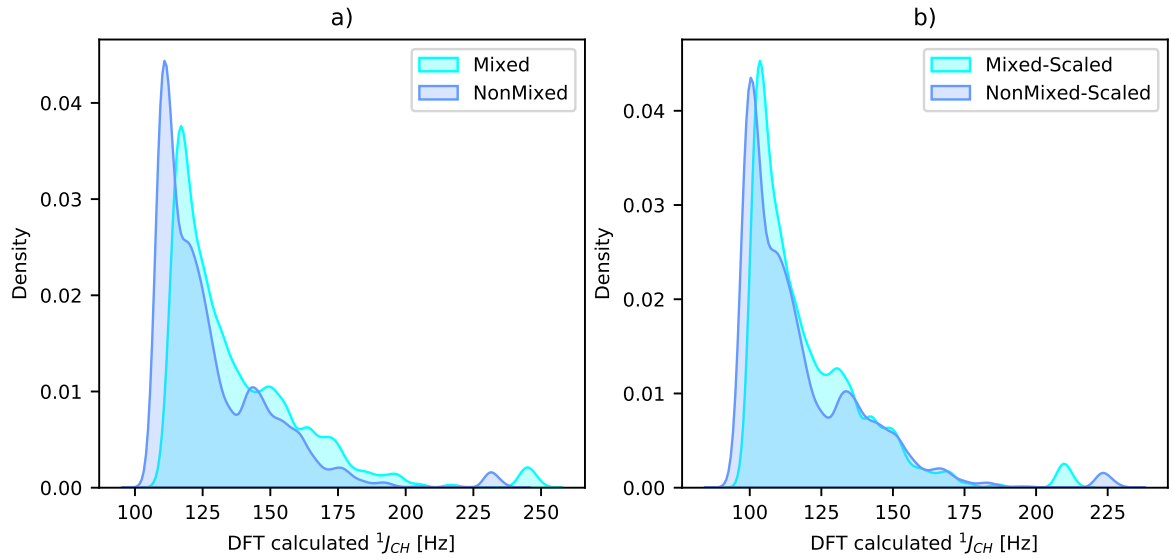

Figure 2.1: a) $^1J_{CH}$ Coupling constant distributions for QM9 molecules calculated with and without mixed. b) $^1J_{CH}$ Coupling constant distributions for QM9 molecules calculated with and without mixed, scaled according to calculated scaling factors available from reference [1].

set and adding tight polarisation functions for the core orbitals in the calculation of the Fermi contact term [106]. Unfortunately it was discovered during the writing of this thesis that this option has not been used in the NMR calculations for some calculations in QM9 (discussed below), in benchmarking work on strychnine the use of the mixed option improved the mean absolute error in experimental ^1JCH prediction from 11.7 Hz to 3.20 Hz [1], and so not using this option may have had a significant impact on the quality of the calculated coupling constants.

Scaling factors were available between DFT calculated and experimentally measured values for DFT calculations with and without mixed for strychnine [1], these were applied to the values calculated for the QM9 molecules, but made only a marginal improvement to the accuracy of the coupling constants in all cases, as is shown for $^1J_{CH}$ in Figure 2.1. The coupling constants calculated without the mixed option were therefore left as calculated, and the effects of this on the accuracy of predictions is discussed where appropriate.

The most desirable solution is to re-calculate the coupling constants in all cases where mixed was not used, and re-train each model using this data. This represents several months of real time for the calculations to be performed and the models retrained, and so is not feasible within the scope of this work. Considering the only use of the QM9 data is for benchmarking this work relative to other publications, and that the only likely effect of this error is to reduce the apparent

quality of the models and predictions, using the dataset as it has been constructed appears reasonable.

2.2.2 Chemical Shift Scaling

The magnetic shielding tensors calculated through DFT were converted to chemical shifts using reference calculations, using the linear scaling method reported by Tantillo et al [93]. For ^{13}C and ^{1}H chemical shifts, reference compounds available from the CHESHIRE Chemical shift repository [107] were used, for ^{15}N experimental data was obtained from separate published work [108]. The scaling method requires DFT calculations to be carried out according to the proposed method, then a linear regression fit made between the calculated magnetic shielding tensors and the reported experimental values. The regression parameters can then be used to calculate chemical shifts from the shielding tensors produced by that DFT method for any molecule, according to the following equation.

$$\text{Chemical Shift} = \frac{\text{Intercept} - \text{Isotropic Shielding}}{-\text{Slope}} \tag{2.1}$$

The linear regression fits (shown in Fig. 2.2) are as follows:

$$\begin{aligned} \delta^{1}H: &\quad y = -1.0209x + 31.9947 \\ \delta^{13}C: &\quad y = -1.0401x + 187.9351 \\ \delta^{15}N: &\quad y = -1.0876x - 161.7067 \end{aligned} \tag{2.2}$$

where y is the DFT calculated shielding value, and x is the chemical shift in ppm.

2.2.3 Computational Timing

The entire purpose of developing machine learning models to predict NMR parameters is to replace, in certain circumstances, the more expensive DFT calculations in order to obtain the same scientific outcome in a vastly reduced time-frame.

To that end it is important to have an understanding of the time scales involved in the different calculations, the time taken for a calculation is highly dependent on the size of the molecule involved, however general ranges are still useful. For the geometry optimisation method used in this thesis, the calculations took between 1 and 100 CPU hours, with a mean time of 15 CPU hours for the 772 molecules in dataset 4 (discussed below). The NMR calculations took between 1 and 200 CPU hours, with a mean time of 42 CPU hours for the molecules in dataset 4.

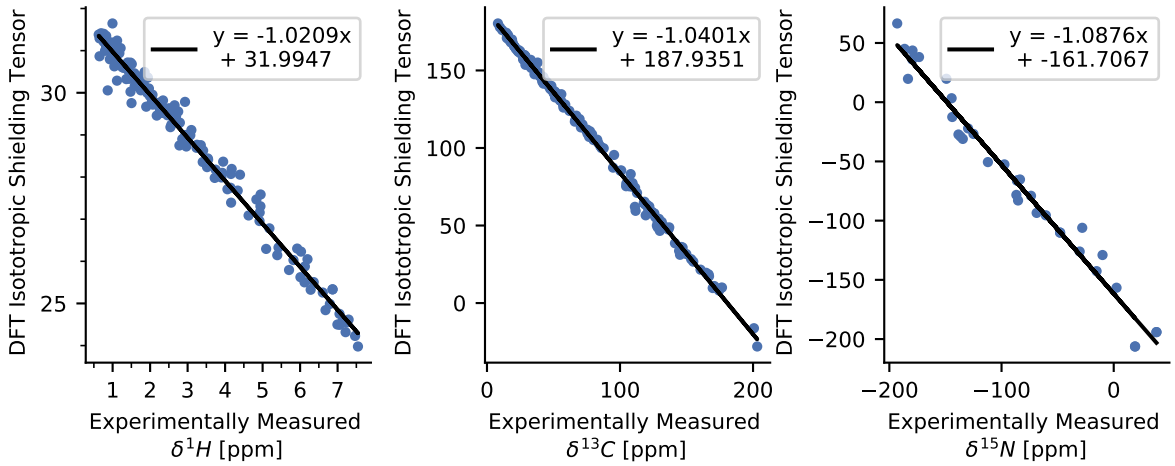

Figure 2.2: Linear regression fit between magnetic shielding tensors calculated by DFT and experimentally measured chemical shifts. RMSD values for each plot: $\delta^1H = 0.26$, $\delta^{13}C = 2.32$, $\delta^{15}N = 12.15$

The CPU time cost for all of the machine learning algorithms is less than 1 minute per molecule, and often less than 1 second when calculating NMR parameters for multiple molecules at once. As such the specific timings are less relevant than the general statement that the machine learning predictions presented in this thesis are obtainable in a few minutes at most, relative to the several hours required for the DFT calculation.

The geometry optimisation is required prior to both the DFT and machine learning NMR calculations, however the level of optimisation required to calculate experimentally relevant NMR parameters is an open scientific question, and beyond the scope of this thesis. It is clear that geometry affects the NMR parameters which are calculated by a given DFT NMR calculation, and that poorly optimised geometries are capable or providing highly inaccurate NMR parameters. It is less straightforward to evaluate the improvement obtained in terms of the accuracy relative to experiment of DFT calculated NMR parameters when the quality of geometry optimisation is significantly improved, and this theoretical improvement could potentially only be obtained through a significantly more expensive calculation. It is also worth noting that the accuracy of the chosen DFT NMR method has a bearing on the relative worth of using a more expensive geometry optimisation, and in order to see an improvement in terms of NMR parameter accuracy, the DFT NMR method would have to be sufficiently accurate. "Sufficiently accurate" here is a vague term, and deliberately so, it would not be possible to evaluate the required accuracy until tens of thousands of CPU hours had been spent calculating datasets to test the combination of

methods.

The geometry optimisation calculation here was chosen as it was known to provide, in combination with the chosen NMR calculation method, experimentally accurate NMR parameters. The method is then also used for the molecules prior to machine learning prediction in order to give a fair comparison of the NMR prediction methods.

2.3 Dataset Workflow

For all of the DFT calculated NMR parameters used in this project the same process was followed:

1. (for non-experimental data) The candidate pool of starting structures is obtained from the external data source, applying some selection criteria to reduce the size of the pool.

2. (for non-experimental data) Structures are chosen from the selection pool according to the relevant selection criteria and sampling algorithm.

3. (for non-experimental data) Checks are performed on the chosen structures to identify mistakes or structures unlikely to optimise.

4. The 3D atom coordinates of the initial structures are optimised using the DFT method described above.

5. Successfully optimised structures are passed to DFT NMR calculations using DFT method described above.

6. The NMR calculated shielding tensors are converted to Chemical shifts using reference calculations.

The reliable execution of this workflow relies upon scripts and packages either publicly available or written specifically for this project. The open-source package 'mol_translator' [109] was written to handle conversion between different chemical structure file formats, set up DFT calculations, and prepare the datasets for use with the machine learning algorithms. The mol_translator package was derived from the autoenrich set of scripts and modules initially written to perform this function which are referred to in the IMPRESSION generation 1 publication [66]. The mol_translator package is written in python 3, and makes extensive use of functions available through the numpy[110], rdkit , openbabel[111] and pybel[112] python packages.

2.3.1 Molecule Screening

A key part of the dataset production workflow is the scripts which screen structures for potential mistakes or undesirable properties. Some properties are straightforward to avoid: checking atom types are within the allowed values (the CSD search algorithm does not do this 100% accurately), and checking molecule size can be performed by loading the molecule into rdkit or pybel and checking the relevant molecule object properties.

A more complex issue is avoiding mistakes where there are missing atoms from the structure, or molecules that are charged. The primary method of detecting these issues is to iterate through a molecule and count the number of bonds connected to each atom, simple rules can then be used to calculate if each atom has the correct number of bonds, and therefore detect if the molecule may be charged or have an atom missing. The issue with this approach is it relies upon the bonds between atoms being a fixed property, however in several cases the CSD record, rdkit, and pybel all disagreed upon the correct set of bonds in a molecule. Also in some cases a molecule was not recorded as charged because of a counterion which would subsequently be lost due to the workflow removing disconnected parts of molecules. Furthermore when dealing with 2D structures and converting them to 3D, the optimisation step is relatively expensive and so ideally mistakes would be screened out prior to this, however the rdkit package can change bonds and bond order in its optimisation routine.

The approach taken in this work was to perform checks for missing atoms through counting bonds as described, check for disconnected molecules through recursive path-searching based on pybel determined bonds, and to assess datasets at the post-DFT stage to look for unusual NMR values in order to manually remove bad structures. Far more attention was paid to molecules in testing sets in this regard, as models may gain useful chemical-space information from unrealistic structures, but in testing sets they will only devalue the performance of the model.

2.4 The Datasets

2.4.1 Dataset 1 and 2: Initial random sets

Datasets 1 and 2 were used for very early initial testing, and were inherited from previous unpublished work [1]. No calculations or data from these datasets have been used for the models referenced in this work, however it is necessary to explain their existence to justify the naming

conventions for the remaining datasets.

2.4.2 Dataset 3: Random Testing Set (DT3)

Referring to the criteria set out in section 2.1, a suitable testing set was desired to evaluate models' ability to predict NMR parameters. Published work on Machine-learning solid-state NMR predictions [58] included a set of 500 compounds chosen at random from the Cambridge Structural Database (CSD) [100]. Using the same structures from Reference [58] would allow a direct comparison with this work, and the accuracy reported in Reference [58] appeared to generalise well.

The set of 500 structures were obtained from the CSD as X-ray crystal structures. The atomic coordinates were then optimised using the DFT method in section 2.2. 70 structures failed to optimise (as a result of mistakes in the obtained structure such as missing atoms or physically unrealistic geometries) and so were discarded from the set. The NMR parameters were calculated for the remaining structures and the shielding tensors converted to chemical shifts according to the method above. A further 84 structures were found to contain two separate molecules, and 20 were found with missing protons, these were also removed from the dataset.

The resulting dataset 3 (DT3) contains 326 molecules, consisting of 6236 ^{1}H, 5569 ^{13}C, 450 ^{15}N, and 1,012 ^{17}O environments. The distribution of molecule sizes is shown in Figure 2.8, the distribution of chemical shift values for $\delta^{1}H$, $\delta^{13}C$, and $\delta^{15}N$, as well as $^{1}J_{CH}$, and $^{3}J_{HH}$ coupling constants are shown in Figure 2.7.

This dataset represents the core testing set against which all models will be evaluated. One of the advantages of this set is in its relevance to the tasks the IMPRESSION models are likely to be used in. Considering the CSD is comprised of X-ray crystal structures submitted by research scientists to the database, it should be biased towards the types of structures which form solids and are commonly the subject of scientific research. Whilst still being a very broad range of structures, this is a useful bias for a testing set in this case.

2.4.3 Dataset 4: Adaptive sampling training set (DT4)

Referring to the criteria set out in section 2.1, a suitable training set was required for the first attempt at developing NMR prediction models. Due to the testing set having already been determined (Dataset 3), the training set was also obtained from the Cambridge structural

database. Whilst selecting structures from different sources could produce a more general model, at the initial stage it was thought this would introduce a unnecessary additional variable. In order to develop a training set which optimised the model performance across the parameters of interest at this stage (Generation 1: δ^1H, $\delta^{13}C$, $^1J_{CH}$), an adaptive sampling scheme was used.

A superset of organic structures which matched the following criteria was obtained from the CSD.

- H/C/N/O/F atoms only.
- 3D Coordinates available.
- Molecule is not charged.
- No reported Errors in the structure.

Charged molecules were excluded from all datasets in this work, on the basis that DFT NMR calculations can be less accurate on charged molecules, especially without explicit solvent, and that this would introduce an unnecessary additional variable to the intended analysis.

A total of 75,382 were downloaded in this superset. An initial set of 100 structures was taken at random from dataset 3 and used to train 5 kernel ridge regression models (details in section 3) each using 80% of the dataset. These models were used to make predictions on the entire 75,382 molecule superset, and the variance calculated across the 5 predictions. The 100 molecules with the highest variance in each of the three parameters of interest (δ^1H, $\delta^{13}C$, $^1J_{CH}$) were selected, and processed according to the workflow in section 2.2. The initial set of 100 structures was discarded after the first round of selection, and in all rounds molecules any from Dataset 3 were not eligible for selection. The set of 300 molecules was then used to train another 5 models, and the process repeats. In total 4 rounds were performed, for a total of 1200 molecules selected. Due to this process deliberately selecting molecules with unusual structures, a higher than normal amount of structures failed to optimise, 428 in total. As in the calculation of dataset 3, these molecules which were discarded failed to optimise due to a combination of missing atoms, physically unrealistic structures, or other structure defects.

The resulting dataset 4 (DT4) contains 772 molecules, consisting of 16,187 ^1H, 14,984 ^{13}C, 1,284 ^{15}N, 2,733 ^{17}O, and 213 ^{19}F environments. Several molecules in this dataset were found to contain multiple disconnected fragments, however these were retained in the dataset in this case.

The distribution of molecule sizes is shown in Figure 2.8, the distribution of chemical shift values for δ^1H, $\delta^{13}C$, and $\delta^{15}N$, as well as $^1J_{CH}$, and $^3J_{HH}$ coupling constants are shown in Figure 2.7.

2.4.4 Dataset 5: ChEMBL (DT5a and DT5b)

In an attempt to improve both the generality of models produced, and the relevance of data available for validation, a set of molecules was obtained from the ChEMBL database of drug-like molecules. The molecules were chosen which matched the following criteria:

- H/C/N/O/F/Si/P/S/Cl/Br atoms only.
- Number of heavy (non-H) atoms greater than 9 and less than 70.
- Molecule is not charged.
- Molecule contains at least 1 H and at least 1 C atom.

Molecules were chosen at random from the set of 1,941,404 small molecules available from ChEMBL. The molecules were available as 2D structures, therefore it was necessary to generate a 3D conformer for each structure using RDKit. The generated 3D structures were submitted to the same workflow described above in section 2.2. 2001 structures were selected initially, one of which failed to optimise. The resulting set was split into a training set (Dataset 5a, 1600 molecules) and a testing set (Dataset 5b, 400 molecules). It is noted here that significantly fewer molecules failed to optimise for this dataset than for datasets 3 and 4, due to the fact that the compounds were obtained as 2D structures, and then converted into 3D using RDKit, this almost entirely removes the presence of physically unrealistic structures.

The resulting training dataset 5a (DT5a) contains 1600 molecules, consisting of 50,618 ^1H and 41,365 ^{13}C environments as well as others (full data in Tables 2.1, 2.2). The testing set 5b (DT5b) contains 400 molecules, consisting of 11,885 ^1H and 9,912 ^{13}C environments.

The distribution of molecule sizes is shown in Figure 2.8, the distribution of chemical shift values for δ^1H, $\delta^{13}C$, and $\delta^{15}N$, as well as $^1J_{CH}$, and $^3J_{HH}$ coupling constants are shown in Figure 2.7.

2.4.5 QM9 Subsets: QM91k and QM960k

The QM9 dataset[77] is a popular dataset used to benchmark machine learning algorithms in chemistry [60, 79, 80, 113, 114]. The dataset production workflow was also performed on 74,391 molecules chosen at random from the 133,885 molecule QM9 dataset. A testing dataset of 1,000 molecules was selected at random from the 74,391 molecules to act as a comparative test set to datasets 3 and 5b. The testing dataset is referred to as dataset QM91k. The remaining 63,391 molecules form the QM960k training dataset.

The resulting training dataset (QM960k) contains 63,391 molecules, consisting of 565,420 ^{1}H and 404,484 ^{13}C environments as well as N, O, and F Nuclei (full data in Tables 2.1, 2.2).

As mentioned previously, the calculations for some molecules QM9 were run without the mixed option for the coupling constants, which may have had an effect on the accuracy of the coupling constant calculations. The majority of the calculations in this case were run without the mixed option, and so to simplify the analysis of the model performance, those (248) molecules in the testing dataset QM91k for which coupling constants were calculated using mixed were removed.

The resulting testing set (still referred to as QM91k) contains 752 molecules, consisting of 6,949 ^{1}H and 4,751 ^{13}C environments. The distribution of molecule sizes is shown in Figure 2.8, the distribution of chemical shift values for $\delta^{1}H$, $\delta^{13}C$, and $\delta^{15}N$, as well as $^{1}J_{CH}$, and $^{3}J_{HH}$ coupling constants are shown in Figure 2.7.

2.4.6 Experimental Datasets

2.4.6.1 ^{1}H and ^{13}C Data (Experimental Dataset 1: DTe1a and DTe1b)

Experimental data was available from previous (unpublished) work [1], for a set of 12 compounds for which 154 ^{1}H and 216 ^{13}C chemical shifts had been experimentally measured. This data was used to calculate the necessary linear scaling factors to convert magnetic shielding tensors from DFT into chemical shifts. This dataset is referred to as experimental dataset 1a.

Further experimental data was obtained from the work by Smith and Goodman [53], a set of 46 structures containing 906 ^{1}H and 654 ^{13}C chemical shifts, this dataset functions as an experimental validation set for $\delta^{1}H$ and $\delta^{13}C$ prediction. This dataset is referred to as experimental dataset 1b. The structures from both datasets were also processed according to the

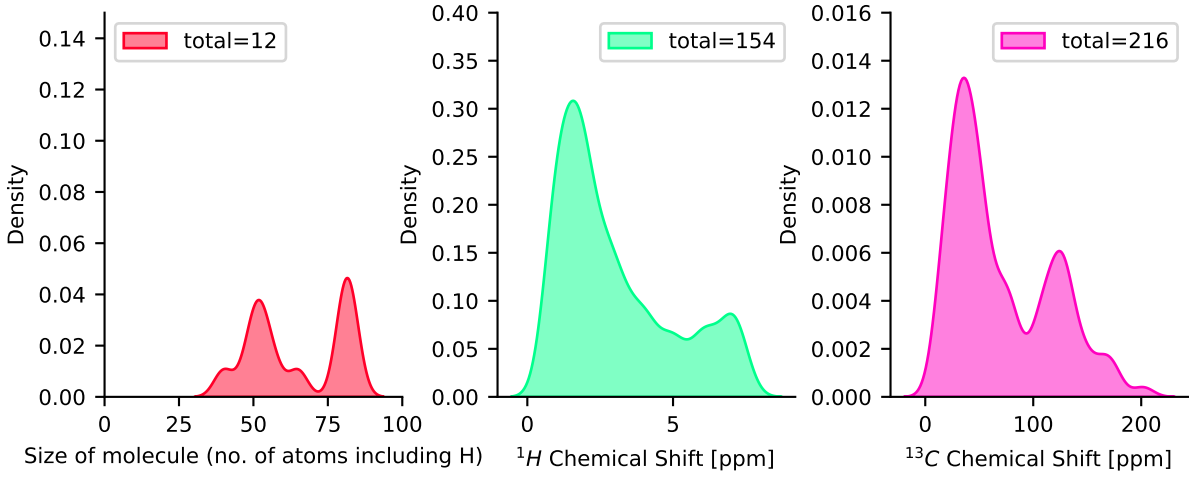

Figure 2.3: Distribution of molecule size, $\delta^{1}H$ Chemical shift and $\delta^{13}C$ Chemical shift values in the experimental dataset 1a.

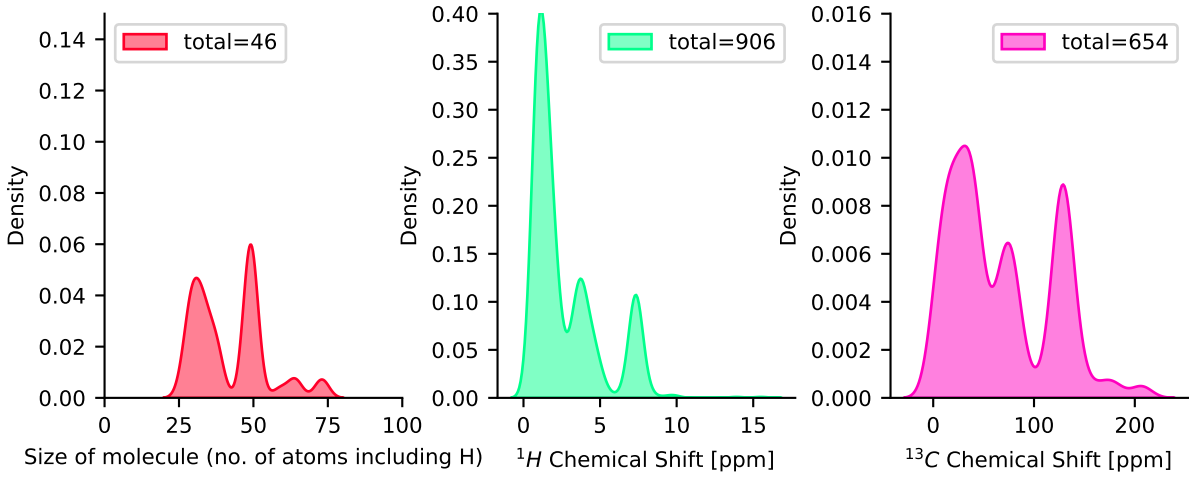

Figure 2.4: Distribution of molecule size, $\delta^{1}H$ Chemical shift and $\delta^{13}C$ Chemical shift values in the experimental dataset 1b.

processing workflow 2.3, in order to obtain DFT calculated NMR parameters.

The size distributions in Figures 2.3 and 2.4 indicate a range of structure sizes, covering the majority of the size range of the DFT based training and testing sets (Figure 2.8). Both $\delta^{13}C$ distributions appear similar to each other, and the DFT based distributions in Figure 2.3. The $\delta^{1}H$ chemical shift distributions are similar, however the linear scaling dataset covers a more limited range, with no values above 8 ppm. Overall therefore the linear scaling dataset (1a) should enable the calculation of good scaling factors for both $\delta^{13}C$ and $\delta^{1}H$ chemical shifts, although more data in the 8-12 ppm range would be advantageous. Furthermore the validation

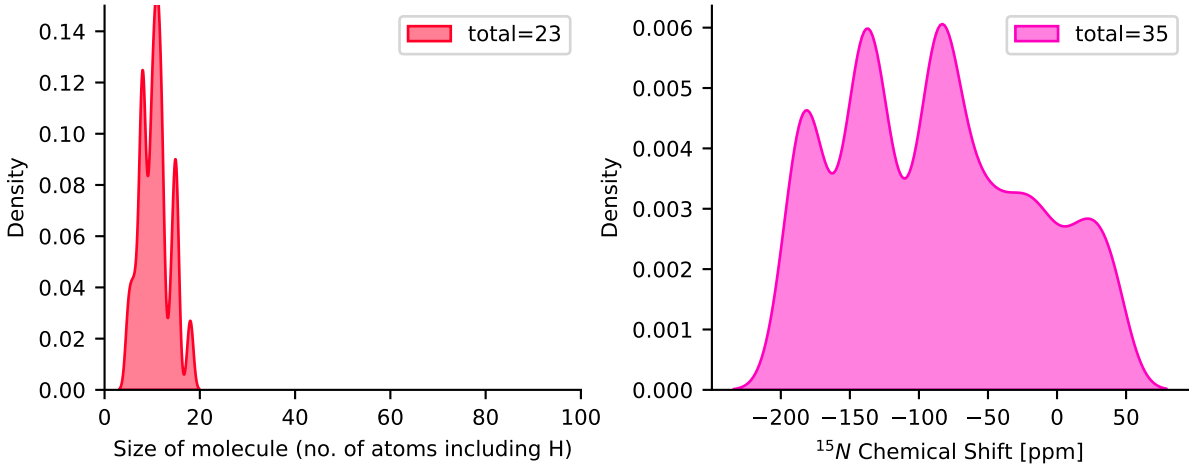

Figure 2.5: Distribution of molecule size and $\delta^{15}N$ Chemical shift values in the experimental dataset.

dataset (1b) should provide a reasonable and fair test of the machine learning models ability to predict NMR parameters for structures from a different source, and how accurate both DFT and machine learning are to the experimental data for $\delta^{13}C$ and $\delta^{1}H$.

2.4.6.2 ^{15}N Data (Experimental dataset 2: DTe2)

Experimental data for a set of 23 compounds with 35 measured ^{15}N chemical shifts was obtained from published work [108].

2.4.6.3 $^{1}J_{CH}$ Data (Experimental dataset 3: DTe3)

Experimental data for a set of 131 compounds with 721 measured $^{1}J_{CH}$ Scalar coupling constants was obtained from published work [115]. In comparison to the DFT methods used in this work, the experimental data obtained here was found to have a consistent 10.91 Hz offset, in published work the correction was made to the DFT data [66], and so this has also been applied in this thesis.

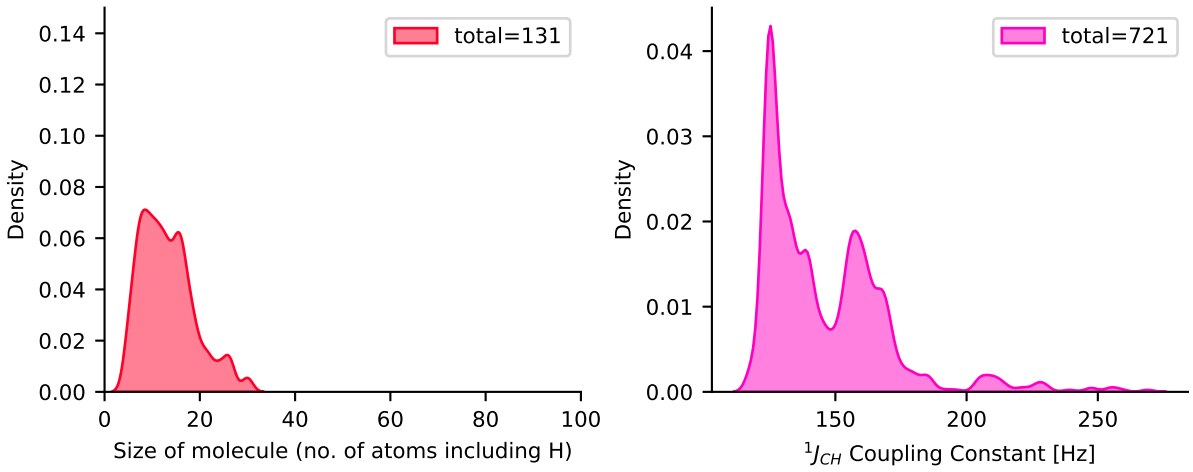

Figure 2.6: Distribution of molecule size and $^1J_{CH}$ Coupling constant values in the experimental dataset.

2.5 Dataset Comparison

There are clear similarities between datasets 3, 4, and 5 in terms of size and chemical shift distribution. The two QM9 subsets are composed of much smaller molecules as expected, however there is a stark difference in the distribution of NMR parameters between the QM9 and non-QM9 datasets. In the δ^1H distribution (Figure 2.7a) there are few values above 5 ppm relative to the other datasets, despite following a similar pattern in the rest of the distribution. Similarly in the $\delta^{13}C$ distribution (Figure 2.7b) there are few values above 100 ppm, whereas the peak of the distribution for all other datasets is between 100 and 150 ppm. The differences in the $\delta^{15}N$ distribution (Figure 2.7c) appear more subtle as here the CSD and ChEMBL datasets diverge from one another as well, however there is still a clear reduction in higher ppm values relative to the other datasets. Whilst some change in the distributions is to be excepted with such a reduction in molecule size, the extent of the difference suggests models trained on QM9 data should struggle to accurately predict chemical shifts with values around these differences. Further comparison to the distribution of experimental data obtained for $\delta^{13}C$ highlights this issue further, as there is a substantial number of values in the 100-150 ppm range.

The variations between the CSD derived datasets (3 and 4) and the ChEMBL derived datasets (5a, 5b) are much smaller, though the significant difference in $\delta^{15}N$ distribution, especially in the -50 to 0 ppm range suggests there is type of environment which is common in the ChEMBL datasets, but relatively rare in the CSD. This highlights the potential benefits of using data

	Dataset 3	Dataset 4	Dataset 5a	Dataset 5b	Dataset QM91K	Dataset QM960K
Size	306	772	1600	400	752	63391
Nuclei						
H	5,905	16,187	50,618	11,885	6,949	565,420
C	5,262	14,984	41,365	9,912	4,751	404,484
N	387	1,284	5,029	1,285	761	60,241
O	960	2,733	8,009	1,943	1,085	86,374
F	0	213	477	125	4	321
Si	0	0	2	0	0	0
P	0	0	69	20	0	0
S	0	0	481	136	0	0
Cl	0	0	308	79	0	0
Br	0	0	67	20	0	0

Table 2.1: Dataset size and constituent atoms summary.

Parameter	Dataset 3	Dataset 4	Dataset 5a	Dataset 5b	Dataset QM91K	Dataset QM960K
^{1}H	5,905	16,187	50,618	11,885	6,949	565,418
^{13}C	5,262	14,984	41,365	9,912	4,751	404,484
^{15}N	387	1,284	5,029	1,285	761	60,241
$^{1}J_{CH}$	5,608	30,324	91,112	10,641	6,284	1,022,650
$^{3}J_{HH}$	3,954	20,714	75,964	8,727	5,111	817,740

Table 2.2: Number of NMR parameters in each dataset for the NMR parameters of interest in this thesis.

from multiple sources, as a model trained purely on CSD data is likely to perform worse on such environments.

As mentioned previously, the NMR calculations for dataset 5 and the QM9 were missing the 'mixed' option which improves the accuracy of the coupling constant calculations. In the distributions of coupling constants for both $^{1}J_{CH}$ and $^{3}J_{HH}$ clear differences between the dataset 5 datasets, the QM9 subsets and the CSD derived datasets (dataset 3 and 4) can be seen, indicating that although there may be a difference due to the missing 'mixed' option, there are also significant differences in the distributions due underlying differences in the sources of the structures in each case.

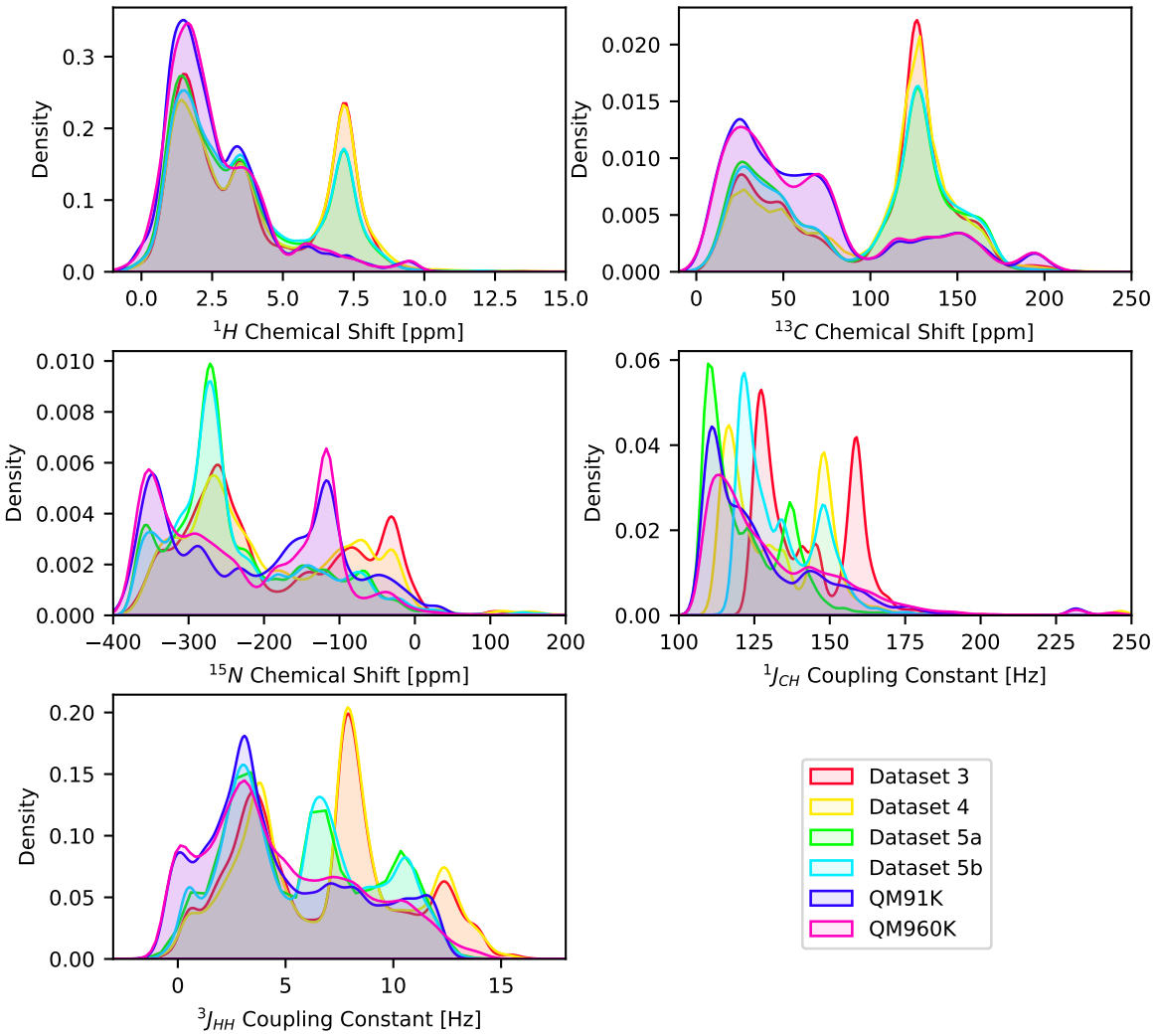

Figure 2.7: Distribution of NMR Parameter values in the DFT calculated datasets.

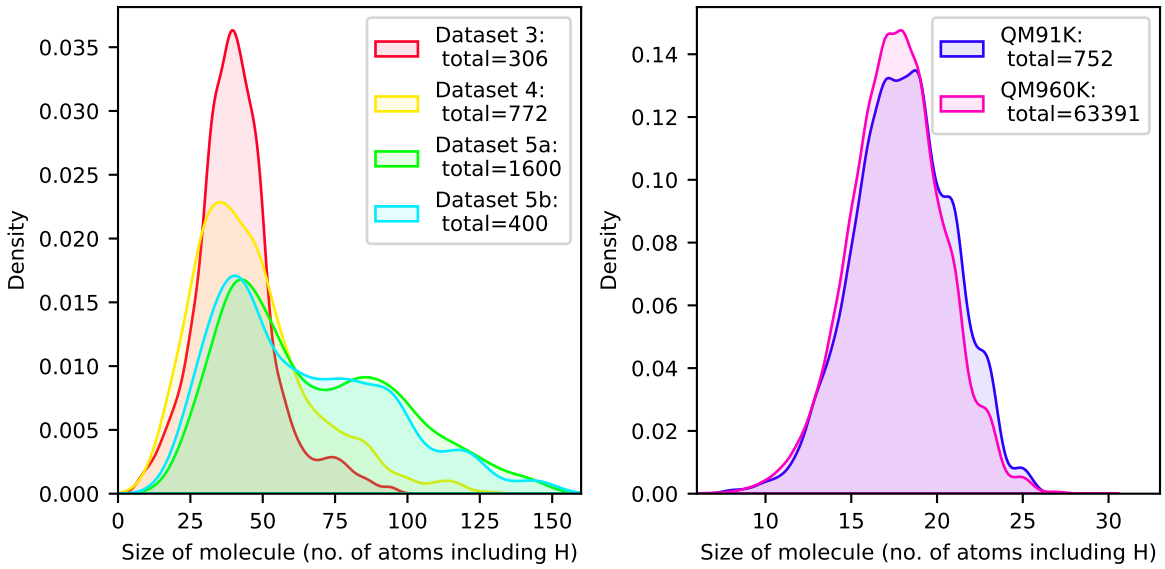

Figure 2.8: Distribution of molecule size in the DFT calculated datasets, size includes H atoms.

CHAPTER 3

IMPRESSION GENERATION 1

3.1 Model Architecture and Training

3.1.1 Kernel Ridge Regression

The first generation NMR prediction model [66] is based on Kernel Ridge Regression (KRR), a popular and straightforward machine learning architecture [76]. KRR algorithms can map molecular information onto a target space for a given observable, in this case NMR parameters. The model consists of an input representation of the molecule, which can take several forms [116–121] and a kernel function to calculate a similarity between these representations, which again can take several forms [122, 123].

Taking the example of $\delta^1 H$ chemical shift prediction, the chemical shift (y_i) for a proton in a given chemical environment ($\mathbf{E}_i$) is estimated as a linear combination of the chemical environments' similarity to all other (observed) chemical environments for which the chemical shift is known:

$$y_i^{\text{pred}} = \sum_j^N \alpha_j k\left(\mathbf{E}_i, \mathbf{E}_j\right) \tag{3.1}$$

Where N is the number of chemical environments in the training dataset, k is a kernel function, α are the regression parameters calculated to map the training environments to

chemical shift values. The regression parameters are calculated by regularised least-squares optimisation:

$$\textbf{minimise} \quad \alpha; \qquad \sum_{i}^{N}\left(y_i^{\text{exp}} - y_i^{\text{pred}}\right)^2 + \lambda \sum_{i}^{N} \alpha_i^2 \tag{3.2}$$

where y_i^{pred} is given by equation 3.1. λ is the regularisation coefficient, which controls the strength of the regularisation penalty on the α parameters to prevent overfitting. The l_2 regularisation (where l_2 refers to the use of the squared term in the penalty) penalises solutions where α contains large, less uniform values.

3.1.2 Chemical Environment Representation

The kernel function, and atomic representation, used for the first generation models are based on the work by Faber et al [118]. Their 'FCHL' representation and kernel (acronym derived from authors initials) divides the terms usually included in a atomic representation into M-body terms: 1-body terms account for chemical composition, the 2-body terms account for interatomic distances, the 3-body terms introduce the angles between pairs of atoms. Each of these terms is constructed by Gaussian functions with tunable width. In a more common atomic representation such as a coulomb matrix[116], terms such as these would be calculated and flattened into a vector, a kernel function would then calculate the distance between each pair of vectors. In contrast, the FCHL approach keeps them separate, and calculates the kernel distance on each term separately before combining the separate distances into the final kernel distance:

$$k(E_i, E_j) = \int_N d\chi_1 ... d\chi_N (E_i, E_j) \tag{3.3}$$

Where each $d\chi_n$ is a term such as those discussed above.

In order to predict scalar coupling constants, a pair-wise property, this approach to atomic representation and kernel distance was augmented to take into account pairs of atoms linked to a single NMR parameter. For the prediction of $^1J_{CH}$ values, the kernel distance between two $^1J_{CH}$ environments was evaluated as the product of the kernel distances between the ^1H and ^{13}C environments:

$$k\left(\mathbf{E}_i^{\text{CH}}, \mathbf{E}_j^{\text{CH}}\right) = k\left(\mathbf{E}_i^{\text{H}}, \mathbf{E}_j^{\text{H}}\right) k\left(\mathbf{E}_i^{\text{C}}, \mathbf{E}_j^{\text{C}}\right) \tag{3.4}$$

3.1.3 Hyper-parameter Optimisation

For the architecture defined, three hyper-parameters were altered in order to achieve an optimised model: The cutoff distance C, the kernel width σ, and the regularisation coefficient λ. For all of the terms beyond 1-body interactions in the FCHL representation, the cutoff defines at what distance from the central atom other atoms are included in the representation. Effectively terms involving atoms beyond this cutoff are reduced to zero. The kernel width in a traditional kernel defines how quickly the similarity between two vectors falls to zero, in the FCHL formulation this kernel width is present across multiple terms, but has the same effect. The regularisation coefficient, discussed above, reduces overfitting in the model by penalising large or non-uniform regression parameters in α.

There are many strategies for hyper-parameter optimisation, the primary theme among them being to test multiple sets of values to find the optimal combination. Basic strategies such as random and grid search retain popularity due their ease of implementation, but more complex search methods can provide better optimised hyper-parameters in shorter time [124]. A Bayesian optimisation algorithm was used to optimise the hyper-parameters for the first generation models.

Bayesian hyper-parameter optimisation involves creating a surrogate model which maps the hyper-parameters onto the desired optimisation criteria, in this case the mean absolute error (MAE) over cross-validation. This is calculated by training 5 separate models on 80% subsets of the training dataset, for each subset model the target values are predicted for the remaining 20%. In this way the mean absolute error is calculated across the entire training dataset using predictions from models where the test environment was not part of the training dataset. The surrogate model is trained on each point evaluated in the optimisation. An acquisition function then searches the remaining hyper-parameter space to identify new points to evaluate, using a tunable balance of exploration and exploitation. The most common surrogate model is a gaussian process, which was also used in this work. The python package BayesianOptimisation was used to perform the hyper-parameter searching for the generation 1 models [125].

3.1.4 Uncertainty Estimation

The adaptive sampling algorithm used to generate Dataset 4 (Section 2.4.3) relies upon using the variance in predictions made across several drop-out models on the same environment. This exact same methodology can be used to provide an estimation of the uncertainty in any given

prediction. For each model, 5 more models are trained using 80% random subsets of the training dataset, predictions are made using each drop-out model, and the variance calculated.

This pre-prediction variance correlates with the prediction error (see below), however many predictions with low variance will still have high error, and high variance predictions can still have seemingly correct values. To illustrate how these two situations might occur a vastly simplified situation can be envisaged (Figure 3.1) where the environments can be placed on just 2 dimensions in chemical space. The training dataset consists of three subsets (1-3) in all cases, three drop-out models are trained to calculate the pre-prediction variance, with each model having one of the three subsets removed from the training set.

The ideal situation is shown in Figure 3.1a, in which the target environment exists in a well mapped region of chemical space, and so different subsets of the training environments should produce models which predict the value for this target environment equally well. Figure 3.1c illustrates the same training data, but a target environment that exists in a non-mapped region of chemical space, here the drop-out models would be expected to provide similar predictions for this target environment, but for them all to be relatively inaccurate compared to Figure 3.1a. A situation which yields high variance but low error relies on the target environment existing in a region of chemical space which is well mapped but only by a very small number of structures. The result of this, visualised in Figure 3.1b, is that the predictions from one of the dropout models (in this case the model which has subset 3 removed) will be highly inaccurate, giving a large pre-prediction variance but the model trained on the entire dataset will still provide an accurate prediction. The situations visualised in Figures 3.1a and 3.1b are the worst case scenarios, as in these cases the pre-prediction variance provides no useful information. Using the same distribution of training environments as 3.1b, the equivalent high error situation is shown in Figure 3.1d. In this case the target environment is equally far from the regions of chemical space mapped by the different subsets, but the subsets themselves occupy very different regions. Therefore the predictions from each drop-out model will be very different, and the predictions from the full model highly inaccurate.

Using the the pre-prediction variance it is possible identify environments which fit into the situations described in 3.1a and 3.1d, and so highlight predictions which may be less accurate. This provides a benefit in the application of prediction models to real-world tasks, where the true value is not known. The pre-prediction variance will however be unable to discriminate between

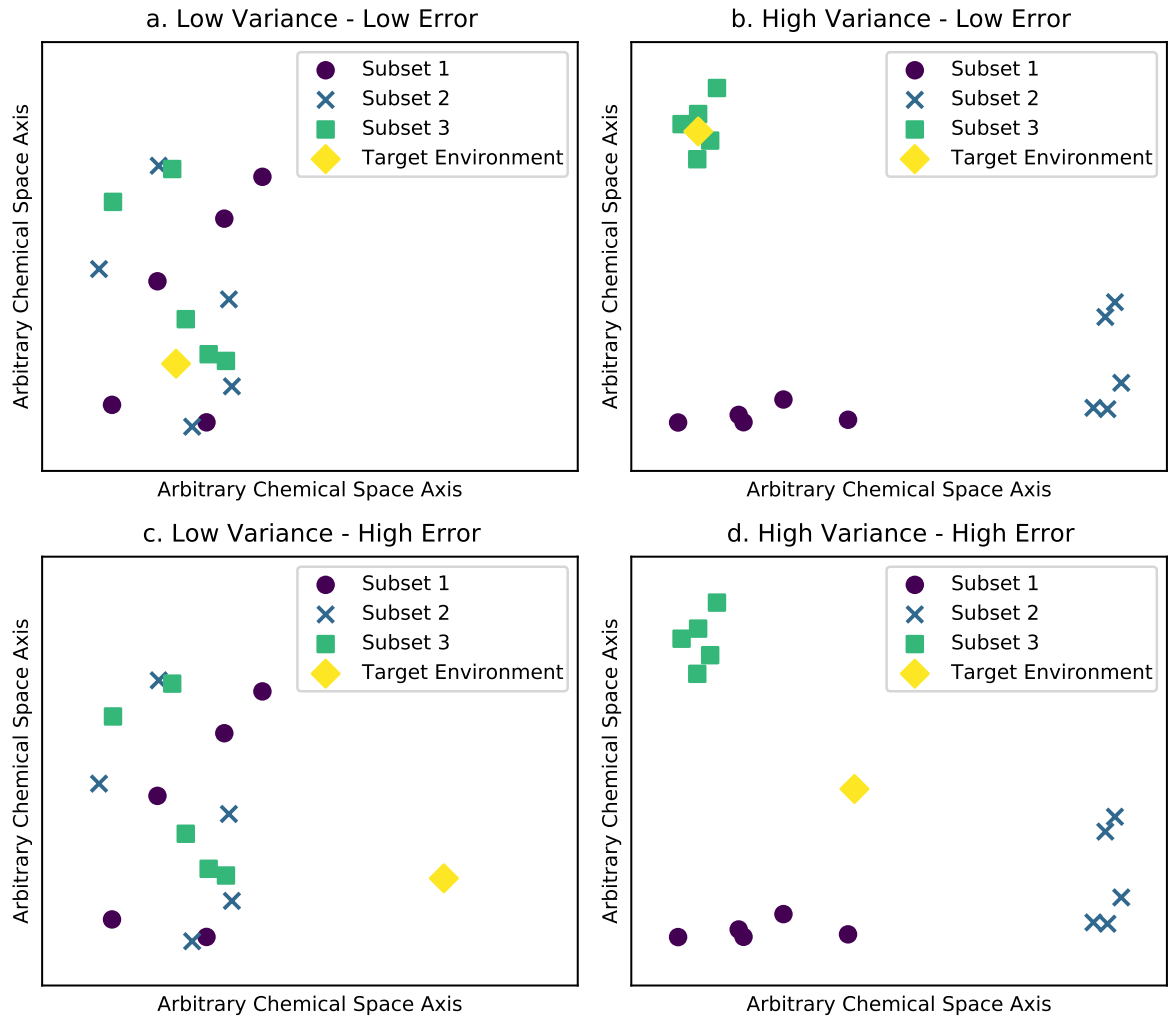

Figure 3.1: Illustrative plots of different situations and their potential effect on both pre-prediction variance and prediction error. a) Low variance with low error. b) High variance with low error. c) Low variance with high error. d) High variance with high error.

the situations described in 3.1a and 3.1c, or between situations 3.1b and 3.1d. The prevalence of environments which match the conditions described in 3.1b and 3.1c will therefore reduce the effectiveness of the pre-prediction variance as a predictor of prediction error.

3.2 Results

3.2.1 Model Training and Summary

Initial models were trained using dataset 4 (DT4) to predict δ^1H, $\delta^{13}C$, and $\delta^{15}N$ Chemical shifts, as well as ${}^1J_{CH}$ Coupling constants. The models were optimised through Bayesian hyper-parameter optimisation, for a minimum of 40 epochs, though most models converged to an optimal set of hyper-parameters after roughly 10 epochs. The optimised model in each case was selected based on the minimisation of the cross-validation loss. The cross-validation loss is the mean absolute error in prediction of NMR parameters for the entire training dataset, where through the cross-validation drop-out procedure, predictions for each molecules are made using a model that did not have that molecule in its training dataset.

The models accurately predict each NMR parameter with a mean absolute error (MAE) of between 1% and 3% of the total range of values for testing dataset 3 (DT3). The performance on testing dataset 5b (DT5b) is comparatively worse for each NMR parameter, with MAEs between 1.5% and 6% of the total range of values. The initial accuracy values here indicate that DT5b presents a more difficult test of the machine learning model trained using DT4, rather than DT3. This is likely due to DT3 and DT4 sharing the Cambridge Structural Database (CSD) as their source repository, whilst DT5b was obtained from ChEMBL. Furthermore DT5b contains molecules with nuclei not found in DT3 and DT4, though this will be looked at in further detail below. The performance of the models across the four NMR parameters is shown in Figure 3.2.

3.2.1.1 Computational Timing

Training each model on the DT4 dataset takes approximately 20-40 minutes, giving a total training time of 100 hours for the 40 epoch, 5-fold cross-validated hyper-parameter optimisation. The time taken to predict one NMR parameter for the entire DT3 testing dataset is about 40 minutes, including making multiple predictions for variance calculation. This increases significantly for coupling parameters to around 2 hours.

This means that the NMR parameter prediction for all predicted NMR parameters for each molecule takes less than a minute.

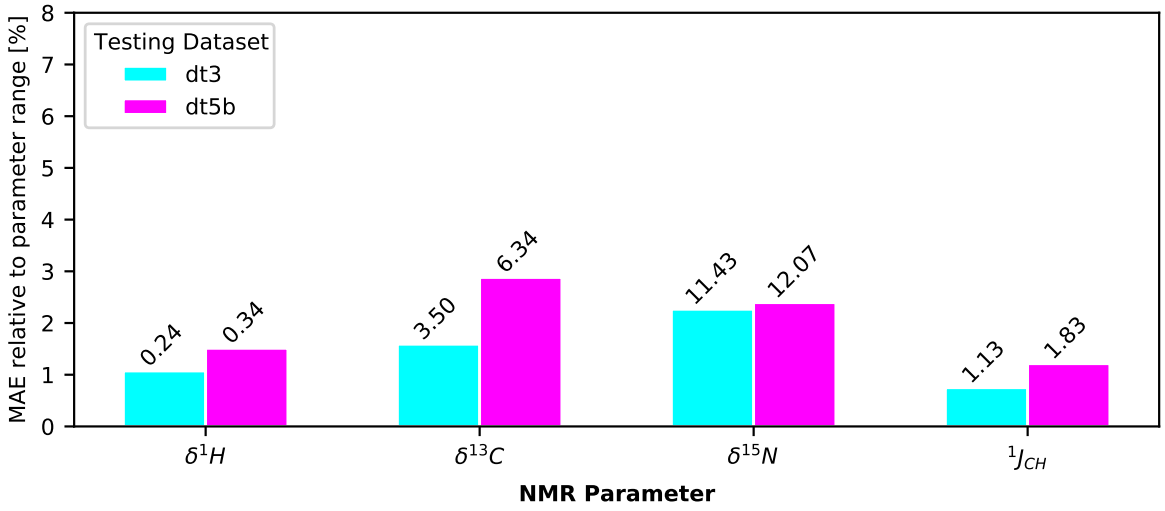

Figure 3.2: Mean absolute error in NMR parameter prediction, relative to the range of values for that NMR parameter, for the models trained using dataset 4 (DT4), for parameters δ^1H, $\delta^{13}C$, $\delta^{15}N$, and $^1J_{CH}$, for both dataset 3 (DT3) and dataset 5b (DT5b).

3.2.2 δ^1H Prediction

3.2.2.1 Performance relative to DFT

The IMPRESSION generation 1 model for δ^1H prediction, trained using DT4 (772 molecules, 16,187 ^1H environments, Section 2.4.3) achieved a mean absolute error (MAE) of 0.24 ppm and a root mean squared deviation (RMSD) of 0.39 ppm against the CSD derived test set DT3 (306 molecules, 5905 ^1H environments, Section 2.4.2). The maximum error (MaxE) in this set of predictions was 4.27 ppm. On the ChEMBL derived test set DT5b (400 molecules, 11885 ^1H environments, Section 2.4.4) the model achieved an accuracy of 0.34 ppm MAE, 0.54 ppm RMSD with a maximum error of 8.78 ppm.

The two molecules with absolute errors greater than 8 ppm in the DT5b dataset, responsible for the outlying values in Figure 3.3a, are shown in the same Figure. The four protons which cause these very large errors are attached to sulphur atoms, a type of nuclei which is poorly represented in any of the available training datasets. This lack of training data readily explains the poor prediction of these chemical shift values. The other significant outlying values on this plot from DT5b are caused by the same issue.

The IMPRESSION model provides reasonably accurate predictions for both the testing sets relative to the range of values contained in the testing dataset. The model performed better on

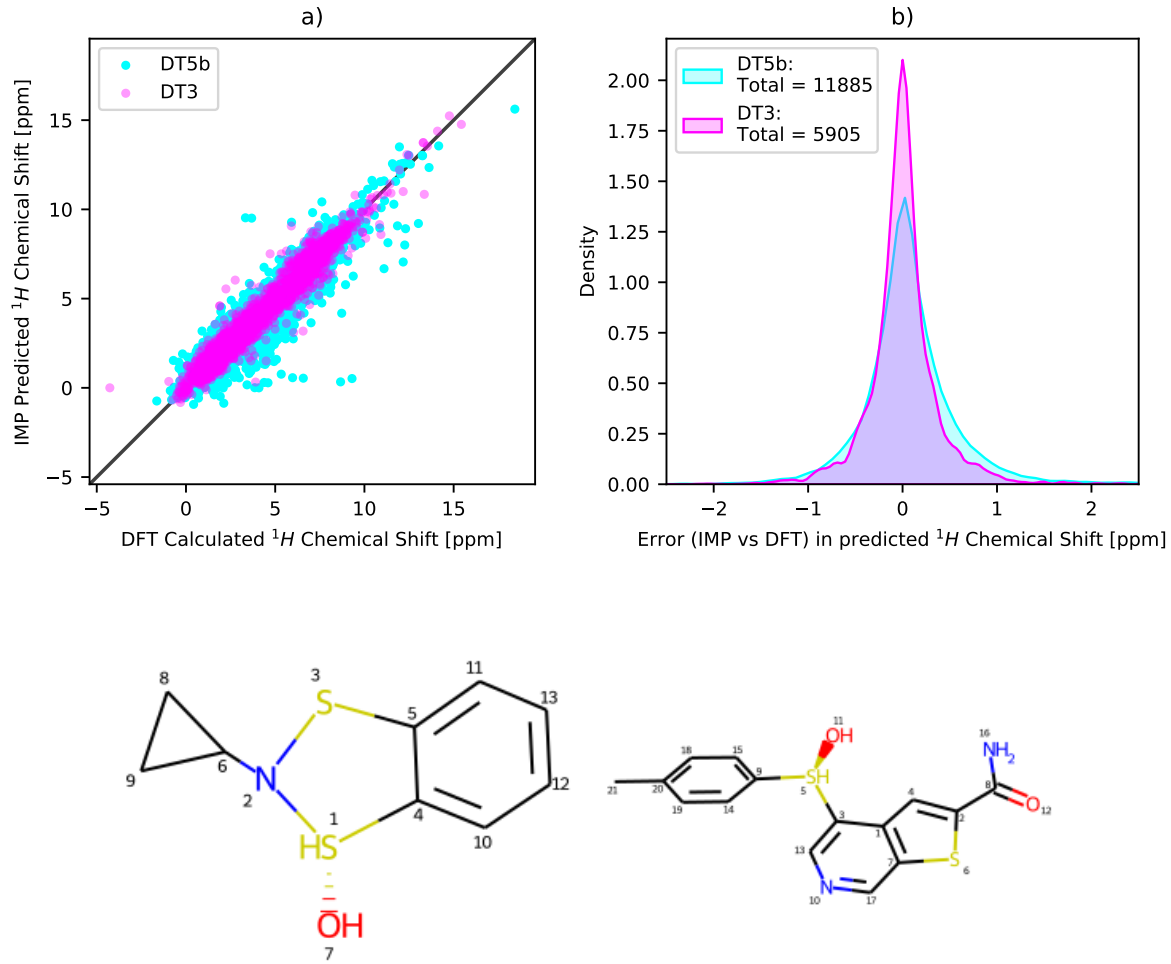

Figure 3.3: a) IMPRESSION predicted and DFT calculated δ^1H for dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 0.24 ppm MAE, 0.39 ppm RMSD and 4.27 ppm MaxE, DT5b fit statistics: 0.34 ppm MAE, 0.54 ppm RMSD, 8.78 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated δ^1H for the DT3 and DT5b testing sets, 12 (DT3) and 50 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4). Structures responsible for the outlying values in (a) also depicted in 2D (IDs: CHEMBL154357, H1 and CHEMBL6320, H5).

the CSD derived testing set (DT3), than the ChEMBL derived testing set (DT5b) which is likely a marker of poor generalisation in the model. The difference between the testing datasets could be due to true chemical diversity or simply differences intrinsic to the source of structures (CSD vs ChEMBL). In either case this suggests the need for a model with better generalisation. The difference in error distributions (Figure 3.3) highlights the differences between predictions, with little noticeable difference in quality of prediction on either dataset with chemical shift value, i.e. higher chemical shifts are predicted as well as lower values. There are significant outliers in the DT5b predictions, with several values badly underpredicted (by 5-10 ppm) in the 5-10 ppm chemical shift range.

3.2.2.2 Uncertainty Estimation

The pre-prediction variance correlates with the prediction error (Figure 3.4), and many large errors are associated with a higher variance. Tables 3.1 and 3.2 show the effect of variance cutoffs on datasets DT3 and DT5b. In both cases a very high variance cutoff, relative to the range of variance values, still produces a significant improvement in terms of the maximum error, whilst removing only 1 and 10 environments for a 0.7 ppm and 2.5 ppm reduction in maximum error

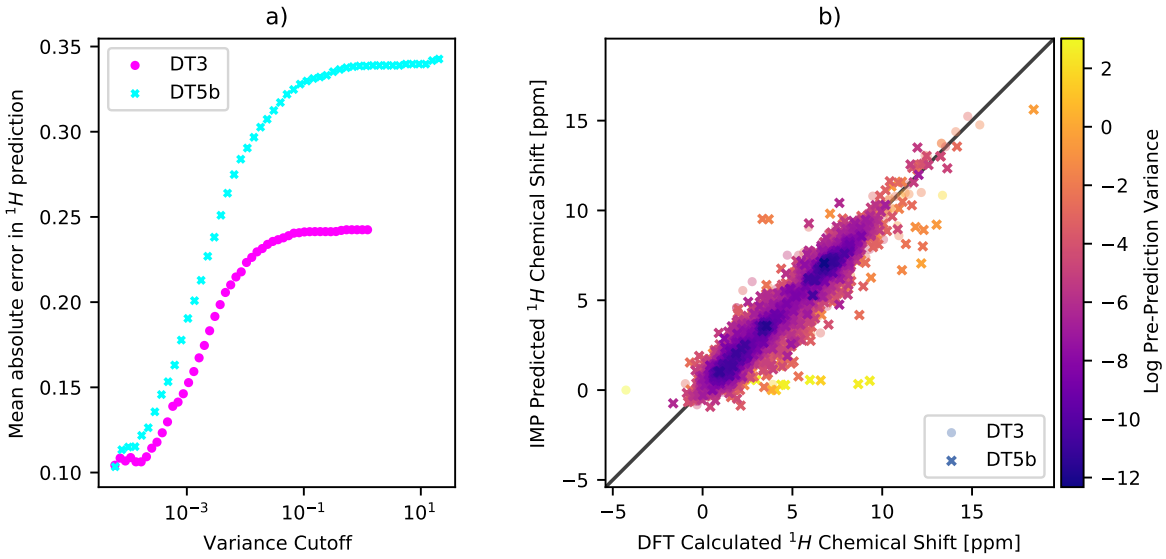

Figure 3.4: a) Error in predicted δ^1H for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated δ^1H for DT3 and DT5b, with variance values highlighted. Results for models trained using dataset 4 (DT4)

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.0001	5468	0.109	0.154	0.877	0.263
0.0005	3830	0.132	0.195	1.700	0.587
0.001	2879	0.150	0.222	1.794	0.749
0.005	737	0.208	0.318	3.278	1.240
0.01	323	0.222	0.338	3.304	1.329
0.05	36	0.239	0.375	3.384	1.598
0.1	12	0.241	0.379	3.384	1.630
0.5	1	0.242	0.384	3.579	1.668
1	1	0.242	0.384	3.579	1.668
5	0	0.243	0.388	4.268	1.699

Table 3.1: Effect on prediction error of removing environments with pre-prediction variance above a cutoff value for IMPRESSION δ^1H predictions against DFT calculations for dataset 3 (DT3). (Total δ^1H environments in DT3: 5,905)

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.0001	11441	0.115	0.160	0.862	0.279
0.0005	9562	0.154	0.219	1.663	0.656
0.001	7958	0.188	0.268	2.363	0.892
0.005	2965	0.264	0.380	3.364	1.536
0.01	1542	0.289	0.416	3.364	1.774
0.05	283	0.322	0.474	3.364	2.222
0.1	136	0.329	0.491	4.573	2.367
0.5	29	0.337	0.514	6.196	2.610
1	13	0.339	0.519	6.196	2.678
5	11	0.339	0.523	6.196	2.737
10	10	0.340	0.525	6.196	2.758
50	0	0.343	0.542	8.785	2.972

Table 3.2: Effect on prediction error of removing environments with pre-prediction variance above a cutoff value for IMPRESSION δ^1H predictions against DFT calculations for dataset 5b (DT5b). (Total δ^1H environments in DT5b: 11,885)

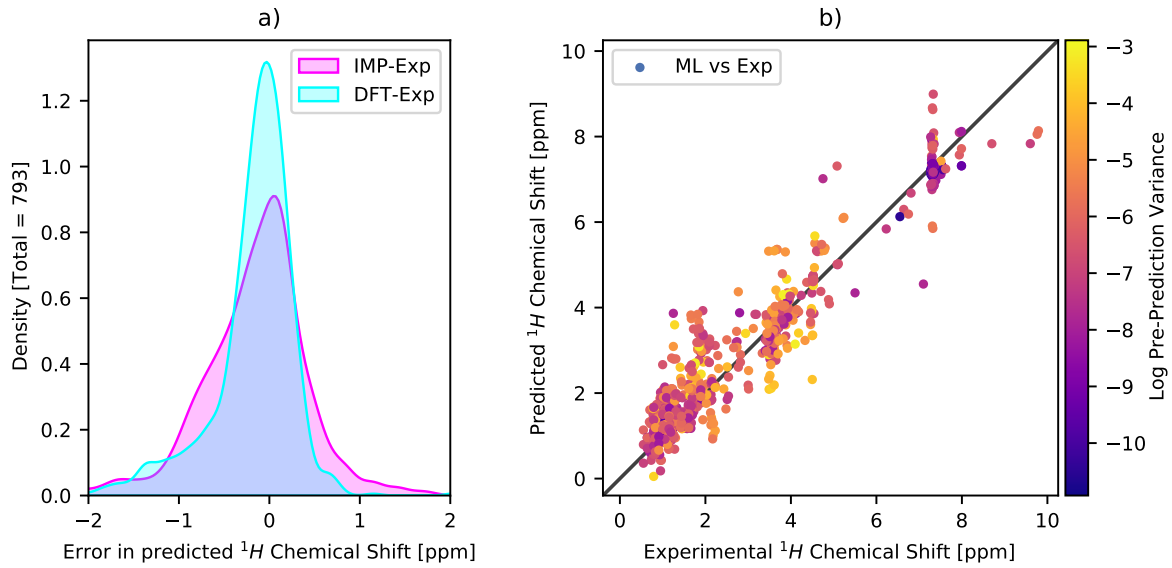

Figure 3.5: a) Error distributions for both IMPRESSION predicted and DFT calculated δ^1H relative to the experimentally measured values for experimental dataset 1b (DTe1b). DTe1b fit statistics: 0.44 ppm MAE, 0.61 ppm RMSD, 2.61 ppm MaxE. b) IMPRESSION predicted and experimentally measured δ^1H for DTe1b with variance highlighted. Results for models trained using dataset 4 (DT4)

respectively. It is difficult to suggest a general cutoff for the variance that might be imposed on a prediction model such as this, nor is that the intention of demonstrating its utility. Instead the pre-prediction variance can be used as a measure of relative uncertainty in any given prediction, and in a practical application special attention would be paid to high variance values relative to the whole set of predictions.

3.2.2.3 Performance relative to experiment

Predictions were also made for the molecules for the 46 molecules from experimental dataset 1b (DTe1b, Section 2.4.6.1) containing 906 δ^1H. The error between the IMPRESSION predicted and the experimentally measured values is 0.44 ppm MAE, 0.61 ppm RMSD, with a maximum error of 2.61 ppm. This must be considered in the context of the error between DFT and experiment (0.33 MAE, 0.50 RMSD, 2.22 MaxE).

The accuracy of IMPRESSION predicted δ^1H to DFT for DTe1b is 0.25 ppm MAE, 0.36 ppm RMSD, 2.57 MaxE. This accuracy is in line with the expectations set by validation on DT3 and DT5b, which demonstrates a degree of generalisation for this prediction accuracy. The prediction

error relative to experiment is encouraging, as the accuracy is similar to that of the underlying DFT method, suggesting the IMPRESSION predictions could be used in place of the DFT method in some circumstances. From the error distributions in Figure 3.5a there is a clear increase in the prediction error relative to the DFT error distribution, and the scatter plot (Figure 3.5b) highlights the presence of a significant number predictions with large error along the whole range of experimental values. Disappointingly in this case the pre-prediction variance is not indicative of prediction error, with the outlier values being associated with a wide range of variance values, indicated by the darker colored points in in Figure 3.5b.

3.2.3 $\delta^{13}C$ Prediction

3.2.3.1 Performance relative to DFT

The IMPRESSION generation 1 model for $\delta^{13}C$ prediction, trained on DT4 (772 molecules, 14,984 ^{13}C environments, Section 2.4.3), achieved a mean absolute error (MAE) of 3.50 ppm and a root mean squared deviation (RMSD) of 7.05 ppm against the CSD derived test set DT3 (306 molecules, 5262 ^{13}C environments). The maximum error in this set of predictions was 106.5 ppm. On the ChEMBL derived test set DT5b (400 molecules, 9912 ^{13}C environments) the model achieved an accuracy of 6.34 ppm MAE, 17.1 ppm RMSD with a maximum error of 271.7 ppm.

The error distributions in Figure 3.6b indicate that for DT3 and the vast majority of DT5b the prediction accuracy is very good, however Figure 3.6a shows a significant number of environments from DT5b which have a large prediction error. These environments have DFT calculated $\delta^{13}C$ in the 100-200 ppm range, yet the IMPRESSION predictions range from 150 ppm to almost 400 ppm. Upon further inspection of the molecules containing these poorly predicted environments, nearly all were found to contain nuclei not present in the DT4 training set used for this model. As shown in Table 2.1 in Chapter 2, DT5b contains P, S, Cl, and Br atoms, and clearly the model could not generalise to environments in molecules containing these nuclei. Reducing the dataset DT5b so that is just contains H/C/N/O/F atoms (233 remaining molecules), the prediction accuracy improves significantly to 4.62 ppm MAE, 8.97 ppm RMSD, 151.3 ppm MaxE (from 6.34 ppm MAE, 17.07 ppm RMSD, 271 ppm MaxE) and is comparable to the prediction accuracy on DT3. The same restriction improves the $\delta^{1}H$ prediction accuracy for DT5b from 0.34 ppm to 0.31 ppm MAE, a much more modest improvement. The accuracy for model predictions on molecules from DT5b containing different subsets of nuclei is shown in Table 3.3. The molecules with the

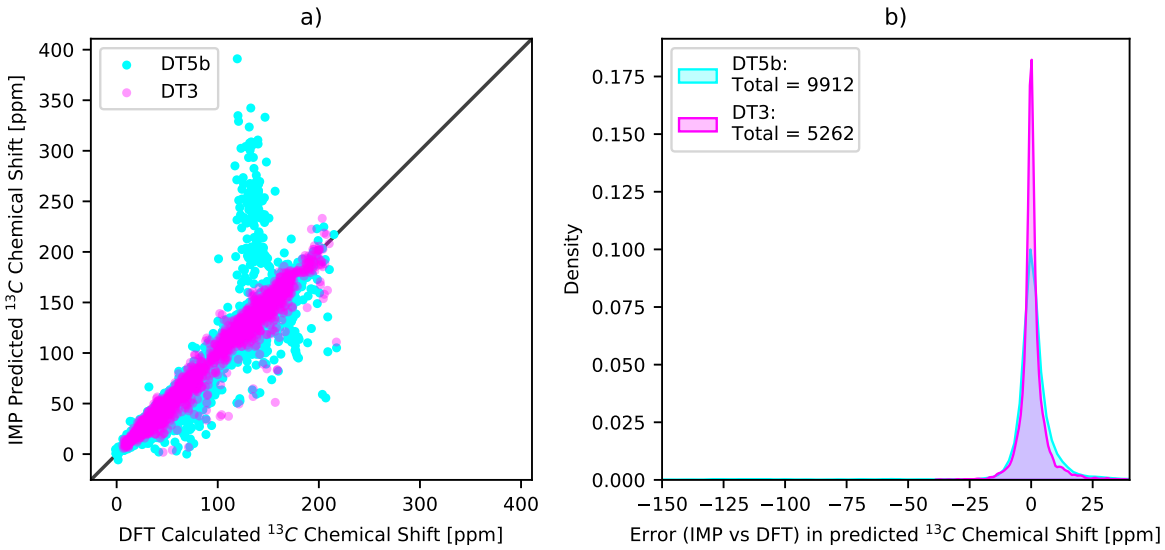

Figure 3.6: a) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{13}C$ for the DT3 and DT5b testing sets, 1 (DT3) and 97 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4)

worst prediction accuracy contain multiple nuclei not seen in DT4, the four worst groups contain bromine and either sulphur or chlorine. The inclusion of phosphorous nuclei on the other hand appears to be handled well by the model, with molecules with H,C,N,O,P having a prediction accuracy of 4.39 ppm MAE. This demonstrates that some degree of extrapolation to nuclei outside of those in the training set is possible.

The model predictions for the reduced dataset 5b (Figure 3.7a) are of similar accuracy across the whole range of chemical shift values. The most noticeable feature of the predictions is that very few environments are significantly over-predicted, with almost all the large errors coming from a significant under-prediction. Furthermore the error distributions in Figure 3.7b highlight how well the model generalises to the ChEMBL dataset (DT5b) once unknown nuclei are removed, with error distributions for both the reduced version of DT5b and DT3 having similar width and shape.

Nuclei in Mol.	No. Envs	MAE [ppm]	RMSD [ppm]	MaxE [ppm]
H C N	240	3.0652	4.8742	26.9215
H C O F	15	3.1983	4.0920	8.1488
H C N F	47	3.8384	6.2041	21.1505
H C N O F	567	4.0845	7.6795	69.4293
H C N O	3,840	4.1468	7.7657	107.6203
H C N O P	133	4.3905	6.8692	38.7524
H C N F S	22	5.2208	14.7653	67.1240
H C N O F S	437	5.8393	12.9635	78.5049
H C O	1,505	6.3066	12.2997	151.3202
H C N O S	1,183	7.2477	17.7931	144.4553
H C N O F P	63	7.6190	16.1977	99.5085
H C O Cl	67	7.7959	24.0347	186.4014
H C N O Br	65	8.4812	30.6364	173.9277
H C N O F S Cl	64	9.0489	23.3926	121.0222
H C N O F Cl	221	9.2859	26.2200	137.6577
H C N O Cl	554	9.9827	27.2952	168.0625
H C N Br	27	10.6124	38.1049	192.3585
H C O F S	17	11.0331	23.7265	78.3847
H C O S	99	11.6066	24.6860	140.7824
H C N S Cl	61	11.8357	29.4008	135.6957
H C N S	80	12.3328	29.3434	146.7029
H C N O F Br	71	12.5806	39.4500	214.7623
H C N O S Cl	278	14.3237	33.7088	152.4494
H C N Cl	36	14.4097	36.6224	130.7861
H C N O S Br	150	14.7205	41.9168	209.6141
H C N S Br	16	18.1613	45.6814	175.6439
H C N O Cl Br	34	22.2818	55.9204	208.5456
H C N O F Cl Br	20	29.0670	72.0087	271.6684

Table 3.3: $\delta^{13}C$ prediction accuracy for sets of molecules in testing dataset 5b containing different sets of nuclei

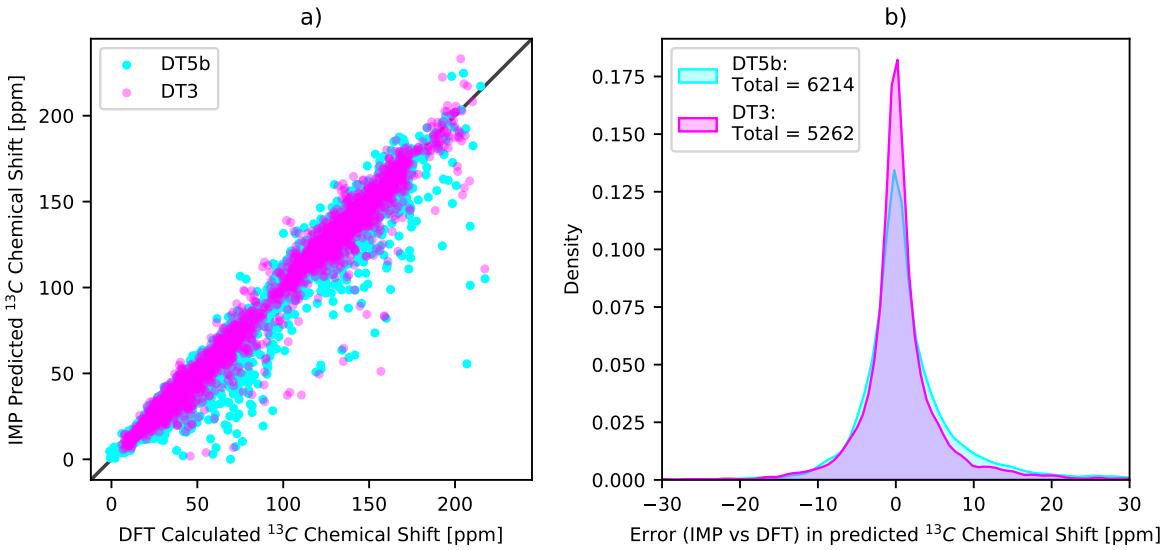

Figure 3.7: a) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for molecules containing only H/C/N/O/F atoms in dataset 3 (DT3) and dataset 5b (DT5b). DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 4.62 ppm MAE, 8.97 ppm RMSD, 151.3 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{13}C$ for molecules containing only H/C/N/O/F atoms in the DT3 and DT5b testing sets, 44 (DT3) and 109 (DT5b) values excluded from graph for clarity. Results for models trained using dataset 4 (DT4)

3.2.3.2 Uncertainty Estimation

The pre-prediction variance values for the $\delta^{13}C$ predictions correlate well with the prediction error for both DT3 and DT5b (Figure 3.8a). The pre-prediction variance also allows clear identification of the structures in DT5b with nuclei not present in the training set, discussed above, as these are highlighted with significantly higher variance values (brighter yellow points, Figure 3.8b). This example demonstrates the key utility of the pre-prediction variance metric, as the poor predictions in this dataset could have been readily identified without the true values for comparison. In a blind use-case where the target value is unknown, and so the prediction error is unknown, these poor predictions would still have been disregarded on the basis of unusually high variance. Applying even a relatively high filter on the pre-prediction variance of 100 removes 219 environments from DT5b and improves the MAE to 4.51 ppm (from 6.34 ppm), the RMSD to 8.43 ppm (from 17.1 ppm) and nearly halves the maximum error from 272 ppm to 151 ppm. The same variance cutoff for DT3 removes only 3 environments but does reduce the MAE from 3.50 ppm to 3.46 ppm, which is significant for such a small percentage of the dataset (< 0.1%). A summary of

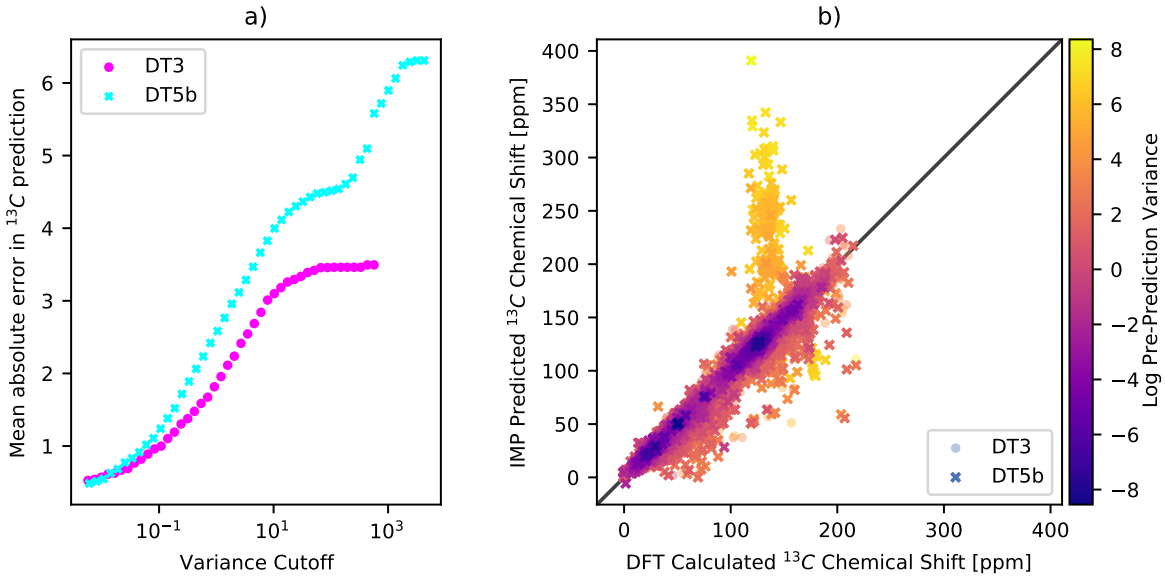

Figure 3.8: a) Error in predicted $\delta^{13}C$ for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated $\delta^{13}C$ for the DT3 and DT5b testing sets, with variance values highlighted. DT3 fit statistics: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b fit statistics: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. Models trained using dataset 4 (DT4)

these results, and the accuracy for further variance cutoff values is shown in Tables 3.4 and 3.5.

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.001	5,220	0.243	0.437	1.553	0.243
0.005	5,057	0.464	0.653	1.979	0.830
0.01	4,846	0.575	0.867	7.123	1.447
0.05	4,046	0.831	1.271	12.146	3.195
0.1	3,560	0.992	1.535	12.146	4.414
0.5	2,192	1.557	2.527	44.059	8.664
1	1,573	1.854	2.984	44.059	10.913
5	415	2.730	4.603	55.562	19.121
10	159	3.096	5.599	73.216	25.433
50	10	3.419	6.590	78.009	32.327
100	3	3.462	6.826	105.720	33.698
500	1	3.494	7.050	106.515	35.075
1000	0	3.496	7.052	106.515	35.075

Table 3.4: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 3. (Total $\delta^{13}C$ environments in DT3: 5,262)

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.005	9,746	0.482	0.830	6.055	0.736
0.01	9,561	0.536	0.858	6.055	1.267
0.05	8,650	0.961	1.523	12.878	4.064
0.1	7,944	1.198	1.866	12.878	5.701
0.5	5,293	2.125	3.219	21.047	11.698
1	3,878	2.537	3.898	65.649	15.156
5	1,121	3.562	5.854	73.104	30.247
10	599	3.977	6.845	107.620	38.399
50	255	4.445	8.264	151.320	49.814
100	219	4.510	8.438	151.320	51.333
500	98	5.294	12.130	152.450	89.558
1000	36	5.874	14.652	168.503	112.797
5000	0	6.336	17.069	271.668	132.253

Table 3.5: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 5b. (Total $\delta^{13}C$ environments in DT5b: 9,912)

3.2.3.3 Performance relative to experiment

Predictions were also made for the 46 molecules in experimental dataset e1b (DTe1b, Section 2.4.6.1), for which 654 experimentally measured $\delta^{13}C$ chemical shifts were available. The error between the IMPRESSION predicted and the experimentally measured values was 4.76 ppm MAE, 6.82 ppm RMSD, with a maximum error of 35.0 ppm. This must be considered in the context of the error between DFT and experiment (2.18 ppm MAE, 2.80 ppm RMSD, 15.9 ppm MaxE). The accuracy of the IMPRESSION predictions relative to the DFT calculated values for this dataset is 3.57 ppm MAE, 3.63 ppm RMSD, 25.5 ppm MaxE.

The IMPPRESSION to DFT prediction accuracy is similar the accuracy achieved on DT3 and the reduced set from DT5b. The accuracy relative to the experimental values is relatively good, as some reduction in accuracy is expected but the error is less than the combined IMPRESSION to DFT and DFT to experiment errors and holds well over the range of chemical shift values (Figure 3.9). A noticeable feature of these predictions is the appearance of an offset in both the error distributions (a) and scatter plot (b) in Figure 3.9, especially in the 75-130 ppm region. This could indicate that the scaling factors are not optimal, or that environments in this region are

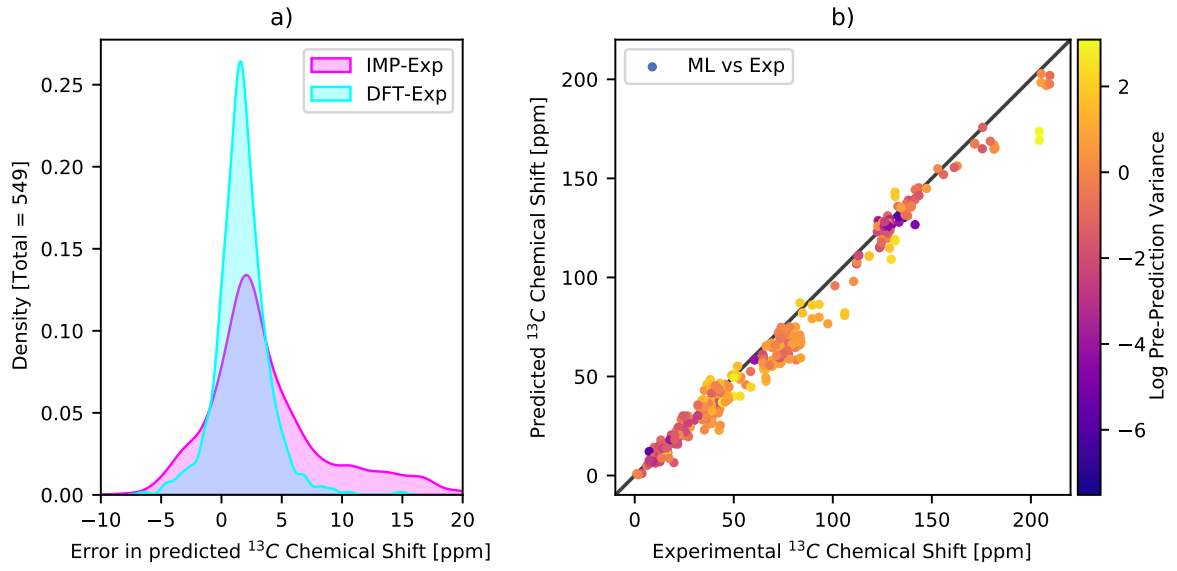

Figure 3.9: (a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{13}C$ relative to the experimentally measured values in experimental dataset 1b (DTe1b). (b) IMPRESSION predicted and experimentally measured $\delta^{13}C$ for DTe1b with variance highlighted. Fit statistics for DTe1b: 4.76 ppm MAE, 6.82 RMSD, 35.0 ppm MaxE. Models trained using dataset 4 (DT4)

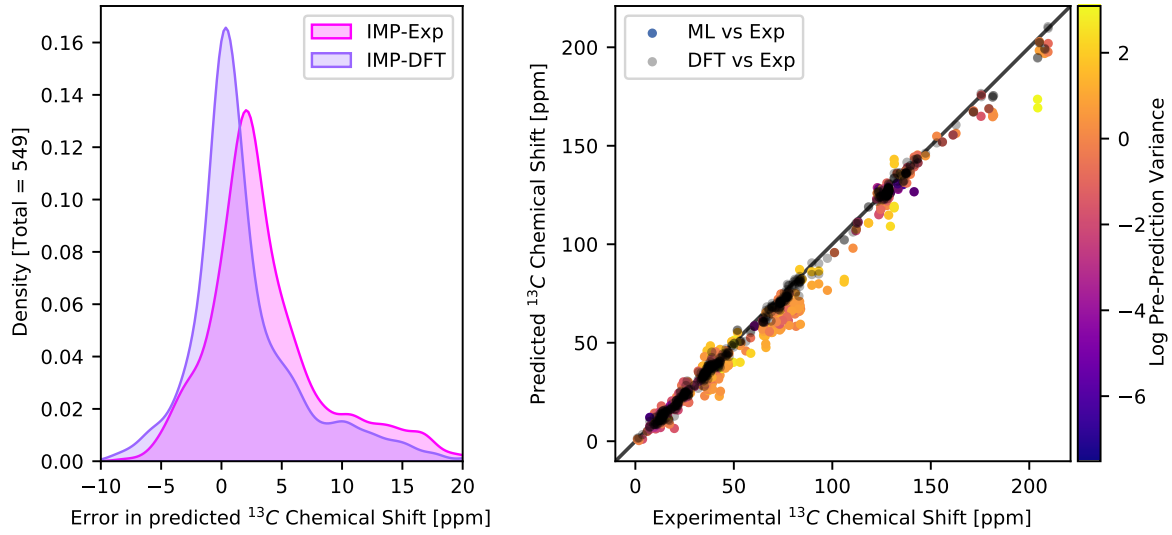

Figure 3.10: (a) error distributions for IMPRESSION predicted $\delta^{13}C$ relative to DFT calculated and experimentally measured values in experimental dataset 1b (DTe1b). (b) IMPRESSION predicted and DFT calculated $\delta^{13}C$ against experimentally measured $\delta^{13}C$ for DTe1b with variance highlighted. Fit statistics for ML to DTe1b: 4.76 ppm MAE, 6.82 ppm RMSD, 35.0 ppm MaxE. Fit statistics for DFT to DTe1b: 2.18 ppm MAE, 2.80 ppm RMSD, 15.9 ppm MaxE. Models trained using dataset 4 (DT4)

under-predicted.

Overlaying the IMPRESSION to DFT error distribution on to the IMPRESSION to experiment distribution (Figure 3.10a) suggests that the issue is at least in part due to the scaling, as the IMP-DFT distribution is nearly centered on zero. Furthermore, overlaying the DFT calculated values for DT1eb on to Figure 3.10b indicates that there is, at worst, only a minor scaling issue as the DFT-EXP points lie close to the $y = x$ line. It is therefore clear that the offset in IMPRESSION predictions is caused by both an underprediction in some chemical shift values, and a small scaling issue. Both are small issues, but combine to a significant offset in the predicted values in Figure 3.9.

The pre-prediction variance provides some benefit in this case by identifying the two worst predicted values in DTe1b, which can be seen at 200 ppm in Figure 3.9b. Applying a variance filter of 10 ppm to this data removes 10 values (out of 654) and reduces the maximum error from 35 ppm to 25 ppm (MAE reduces from 4.76 ppm to 4.58 ppm, RMSD reduces from 6.82 ppm to 6.43 ppm).

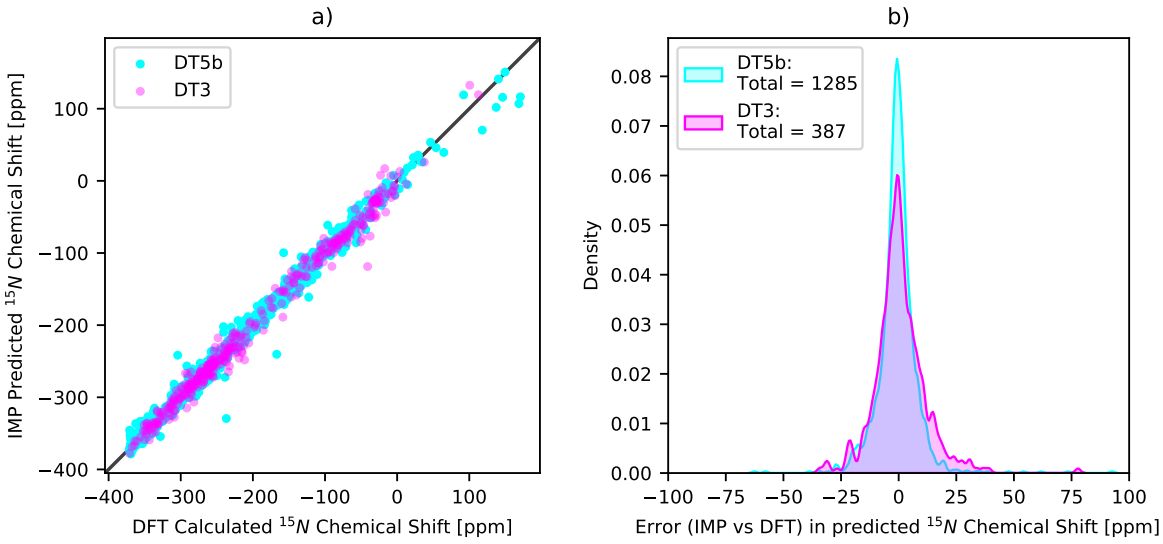

Figure 3.11: a) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for dataset 3 (DT3) and dataset 5b (DT5b). Fit statistics for DT3: 11.4 ppm MAE, 15.9 ppm RMSD, 67.9 ppm MaxE, Fit statistics for DT5b: 12.1 ppm MAE, 18.5 ppm RMSD, 216.5 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, 2 DT5b values excluded from graph for clarity. Models trained using dataset 4 (DT4)

3.2.4 $\delta^{15}N$ Prediction

3.2.4.1 Performance relative to DFT

The IMPRESSION generation 1 model for $\delta^{15}N$ prediction trained on DT4 (772 molecules, 1284 ^{15}N environments, Section 2.4.3) achieved a mean absolute error (MAE) of 11.4 ppm and a root mean squared deviation (RMSD) of 15.9 ppm against the CSD derived test DT3 (306 molecules, 387 ^{15}N environments, Section 2.4.2). The maximum error in this set of predictions was 67.9 ppm. On the ChEMBL derived test DT5b (400 molecules, 1285 ^{15}N environments, Section 2.4.4) the model achieved an accuracy of 12.1 ppm MAE, 18.5 ppm RMSD with a maximum error of 216.54 ppm. As shown in Figure 3.2, this accuracy compares reasonably well with the accuracy for $\delta^{13}C$ when the relative ranges of the parameters are taken into account, with a percentage mean absolute error of 2.36% (DT3) and 2.23% (DT5b) which is higher than the corresponding value for $\delta^{13}C$ on DT3 (1.66%), but lower than that for DT5b (2.90%). The range of $\delta^{13}C$ values is from roughly 0 ppm to 250 ppm, whereas the range of $\delta^{15}N$ is from -400 ppm to 200 ppm, a 450 ppm larger range. In a practical application this means the $\delta^{15}N$ prediction model discriminates equally well between environments across the chemical shift range as the the $\delta^{13}C$ model.

The model performs similarly on DT3 and DT5b with similar error distributions (Figure 3.11b) and little difference in the distribution and magnitude of individual large errors in the scatter plot (Figure 3.11a), indicating that the extra nuclei present in DT5b which cause issues with the $\delta^{13}C$, and to a lesser extent $\delta^{1}H$ prediction, do not affect the accuracy in $\delta^{15}N$ in the same way. There is variation in the accuracy across different subsets of nuclei for DT5b, ranging from the best predicted subset (the 4 environments in molecules containing H,C,N,O,Cl and Br) MAE of 2.99 ppm to the worst predicted subset (the 5 environments in molecules containing H,C,N,S, and Br) MAE of 22.93 ppm, but there is no significant difference between subsets containing only nuclei present in DT4 and those with extra nuclei.

3.2.4.2 Uncertainty Estimation

The pre-prediction variance correlates with the prediction error for DT5b, and to some degree for DT3, although this correlation is only seen for the lowest variance values (Figure 3.12a). The absence of a stronger correlation between prediction error and pre-prediction variance for DT3 could be due to the limited dataset size (387 environments) relative to the other datasets discussed in this chapter, which means the MAE for any one variance subset can be dominated by small number of values.

For DT5b the pre-prediction variance enables the identification of a significant number of the worst errors in the predictions (brighter yellow points in Figure 3.12b). Applying a variance filter of 0.01 ppm removes 195 environments and reduces the MAE from 12 ppm to 11 ppm, and reduces the maximum error from 217 ppm to 98 ppm (3.7). Applying a similar filter to the results from DT3 removes 65 environments and reduces the MAE from 11.43 ppm to 11.07 ppm, and the RMSD from 15.92 ppm to 15.78 ppm, and has no effect on the maximum error (3.6). The pre-prediction variance metric therefore provides some ability to identify poorly predicted environments as intended.

3.2.4.3 Additional Training Data

When compared to the number of available $\delta^{1}H$ and $\delta^{13}C$ values in the training dataset (DT4), there are significantly fewer $\delta^{15}N$ environments. To address this, dataset 5a (DT5a) is used in addition to DT4 for the training dataset for an additional IMPRESSION generation 1 $\delta^{15}N$ prediction model. The combined dataset (4+5a, labeled DT45) of 1967 molecules contains 6313

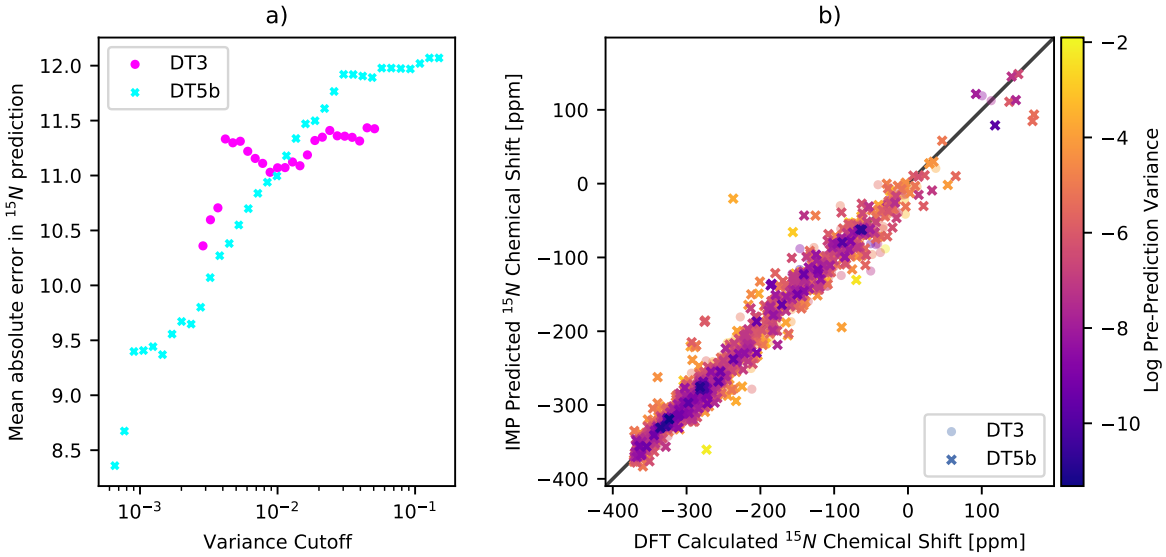

Figure 3.12: a) Error in predicted $\delta^{15}N$ for populations with different maximum variance for dataset 3 (DT3) and dataset 5b (DT5b). b) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 11.4 ppm MAE, 15.9 ppm RMSD, 67.9 ppm MaxE, Fit statistics for DT5b: 12.1 ppm MAE, 18.5 ppm RMSD, 216.5 ppm MaxE. Models trained using dataset 4 (DT4)

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]
5e-05	384	0.563	0.652	0.978
0.0001	380	5.451	11.624	30.507
0.0005	326	9.282	14.848	58.361
0.001	297	10.323	15.849	67.908
0.005	124	11.420	16.499	67.908
0.01	65	11.070	15.782	67.908
0.05	1	11.434	15.937	67.908
0.1	0	11.426	15.922	67.908

Table 3.6: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{15}N$ predictions against DFT calculations for dataset 3. (Total $\delta^{15}N$ environments in DT3: 387)

^{15}N environments. Including additional training data for the prediction models for carbon and proton chemical shift, or 1 bond proton-carbon coupling constants was investigated but due to the size of the FCHL representation used, the resulting training set size was too large and exceeded the available RAM during the training process. This is a key limitation of the KRR model architecture, as all of the training representations must be held in memory at the same time, this is not true of most neural network architectures.

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.0005	1,110	8.370	12.133	48.790	13.142
0.001	937	9.385	13.545	97.499	20.670
0.005	370	10.545	15.368	97.499	34.319
0.01	195	11.004	15.890	97.499	37.656
0.05	13	11.883	18.124	216.542	45.375
0.1	4	12.008	18.369	216.542	46.405
0.5	0	12.071	18.521	216.542	46.995

Table 3.7: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{15}N$ predictions against DFT calculations for dataset 5b. (Total $\delta^{15}N$ environments in DT5b: 1,285)

The $\delta^{15}N$ model trained using both datasets 4 and 5a (DT45) achieved an accuracy of 7.72 ppm MAE, 11.20 ppm RMSD, 77.8 MaxE for DT3 and 5.64 ppm MAE, 9.07 RMSD, 92.6 ppm MaxE for DT5b. The model trained on DT45 far outperforms the original model trained on DT4 (11.4 ppm and 12.1 ppm MAE for DT3 and DT5b respectively), and this improvement is reflected in the decreased number of outlying predictions for DT3 and DT5b (Figure 3.13a) as well as decreased error distribution width (Figure 3.13b). The model now performs better on DT3 than on DT5b, likely due to the fact that the majority of the training data now comes from the ChEMBL database (DT5a contains 5,029 ^{15}N environments to the 1,284 in DT4), and so is more similar to DT5b than DT3, where the structures were obtained from the CSD. The significant improvement in accuracy across both datasets is expected due to the vastly increased training dataset size, increasing from 1,284 $\delta^{15}N$ values in DT4 to 6,313 values in the combined dataset DT45. A summary of the relative prediction accuracy for both training and both testing datasets is shown in Table 3.8.

Unfortunately the pre-prediction variance for the DT45 trained model shows a much weaker correlation with the prediction error (Figure 3.14a). Apart from at very small variance values, there is little to no difference between predictions with different pre-prediction variances, predictions with high pre-prediction variance are not associated with higher prediction error and vice versa. The number of outlying values has reduced compared to the model trained on DT4 alone (Figure 3.15), however the values that remain are not associated with high pre-prediction variance (outlying values not highlighted brighter yellow in Figure 3.14b).

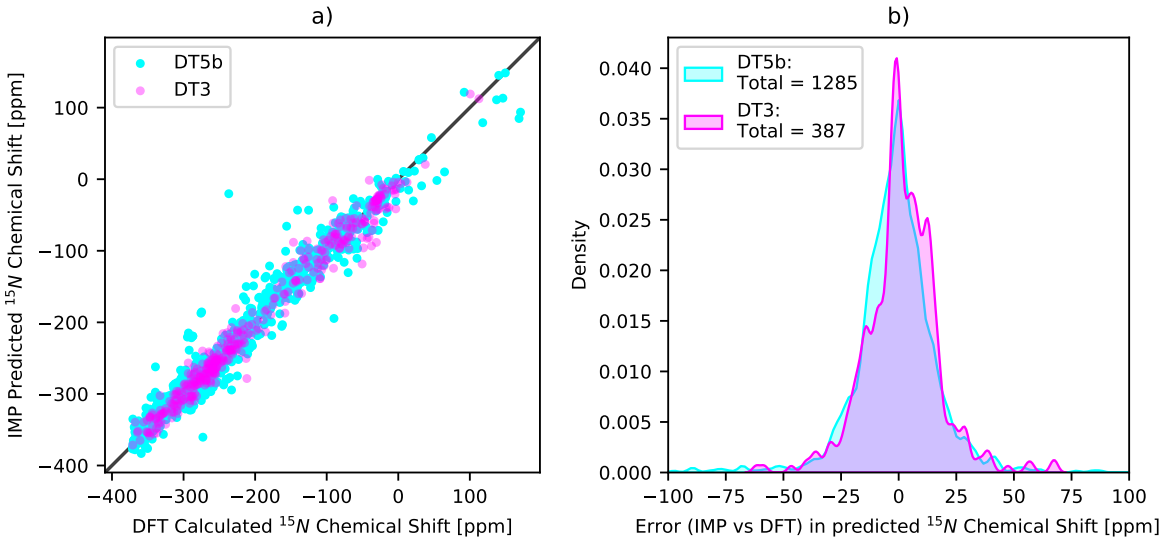

Figure 3.13: Model trained on both dataset 4 and 5a. a) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 7.72 ppm MAE, 11.20 ppm RMSD, 77.8 MaxE, fit statistics for DT5b: 5.64 ppm MAE, 9.07 RMSD, 92.6 ppm MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets.

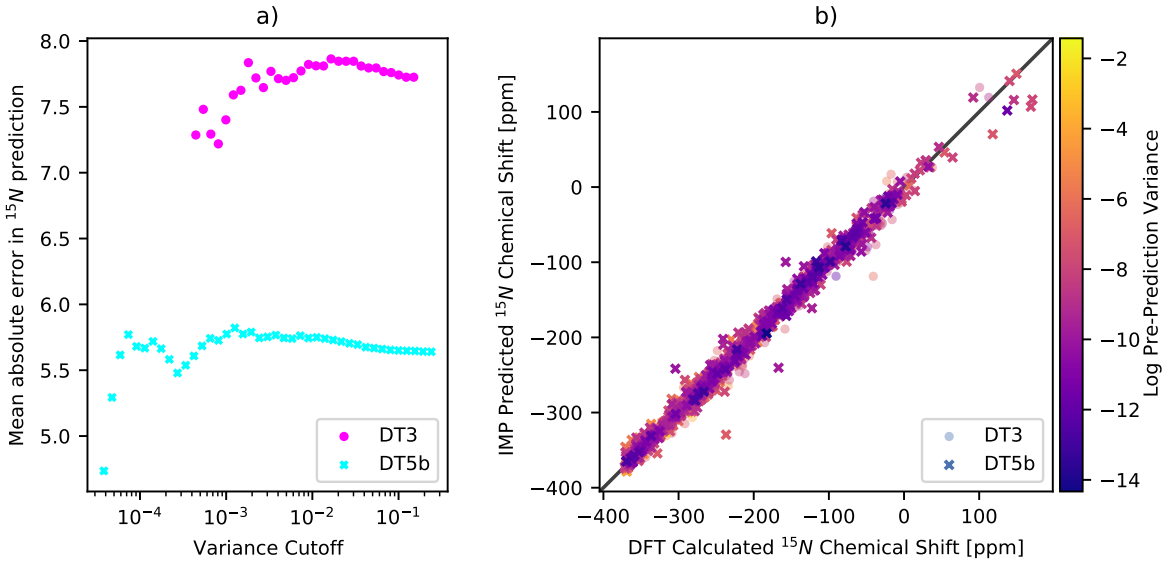

Figure 3.14: Model trained on both dataset 4 and 5a. a) Error in predicted $\delta^{15}N$ for populations with different maximum variance for the DT3 and DT5b testing sets. b) IMPRESSION predicted and DFT calculated $\delta^{15}N$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 7.72 ppm MAE, 11.20 ppm RMSD, 77.8 MaxE, fit statistics for DT5b: 5.64 ppm MAE, 9.07 RMSD, 92.6 ppm MaxE.

Training Dataset	Testing Dataset	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	% of Range [ppm]
DT4	DT3	11.426	15.922	67.908	2.362
DT4	DT5B	12.071	18.521	216.542	2.226
DT45	DT3	7.729	11.203	77.801	1.598
DT45	DT5B	5.640	9.074	92.622	1.040

Table 3.8: Summary of model accuracy in $\delta^{15}N$ prediction for dataset 3 (DT3) and dataset 5b (DT5b) for the IMPRESSION generation 1 models trained using dataset 4 (DT4) and a combination of dataset 4 a 5a (DT45)

3.2.4.4 Performance relative to experiment

Predictions were also made for the molecules in the experimental dataset 2 (DTe2). For the model trained on DT4, the error between the IMPRESSION predicted and the experimentally measured values is 33.04 ppm MAE, 46.43 ppm RMSD, with a maximum error of 141.3 ppm. For the model trained using DT45, the accuracy is 27.02 ppm MAE, 37.3 ppm RMSD, 110.7 ppm MaxE. The error between DFT and experiment for DTe2 is 8.73 ppm MAE, 11.17 ppm RMSD, 23.09 ppm MaxE. Unsurprisingly the model trained using the larger training dataset performs better again, however the accuracy to DFT for this model is 23.5 ppm MAE (31.9 ppm RMSD, 90.6 ppm MaxE) which is significantly worse than the accuracy for DT3 and DT5b. These predictions are relatively poor compared to the experimental predictions for $\delta^{1}H$ and $\delta^{13}C$, both in terms of the MAE relative to the range of experimental values and in terms of the accuracy relative to the accuracy of the DFT method. The MAE in $\delta^{15}N$ prediction as a percentage of the total range of values is 11.7% for the DT45 trained model, the corresponding percentages for $\delta^{1}H$ and $\delta^{13}C$ are 4.7% and 2.28%. This is however affected by the fact that the experimental $\delta^{15}N$ dataset (DTe2) is smaller (35 values) than those for $\delta^{1}H$ and $\delta^{13}C$ (906 and 654 values respectively) and so the range of values for $\delta^{15}N$ will be relatively smaller. A summary of these results is shown in Table 3.9.

Interestingly, the correlation between prediction error and pre-prediction variance is in this case not observed for the model trained on DT4 (higher error predictions are not colored brighter yellow in Figure 3.15b), but is observed for the model trained on DT45 (Several outlying points are coloured brighter yellow in Figure 3.16b). Applying a variance filter of 0.001 removes just 3 environments but improves the MAE from 27 ppm to 21 ppm, the RMSD from 37 ppm to 28 ppm, and the MaxE from 111 ppm to 87 ppm. The ability to identify poorly predicted environments does not improve the accuracy of the model, but their identification allows them to be removed

Target	Training Dataset	MAE	RMSD	MaxE	% of Range
^{1}H	DT4	0.437	0.609	2.615	4.728
^{1}H	DFT	0.326	0.496	2.218	3.525
^{13}C	DT4	4.758	6.817	35.046	2.281
^{13}C	DFT	2.183	2.798	14.866	1.047
^{15}N	DT4	33.040	46.434	141.313	14.298
^{15}N	DT45	27.024	37.315	110.709	11.694
^{15}N	DFT	11.745	14.459	37.794	5.083

Table 3.9: Accuracy in NMR prediction for chemical shift, for the experimental datasets DTe1b ($\delta^{1}H$ and $\delta^{13}C$) and DTe2 ($\delta^{15}N$) for DFT, IMPRESSION trained on DT4 and IMPRESSION trained on DT45 for $\delta^{15}N$

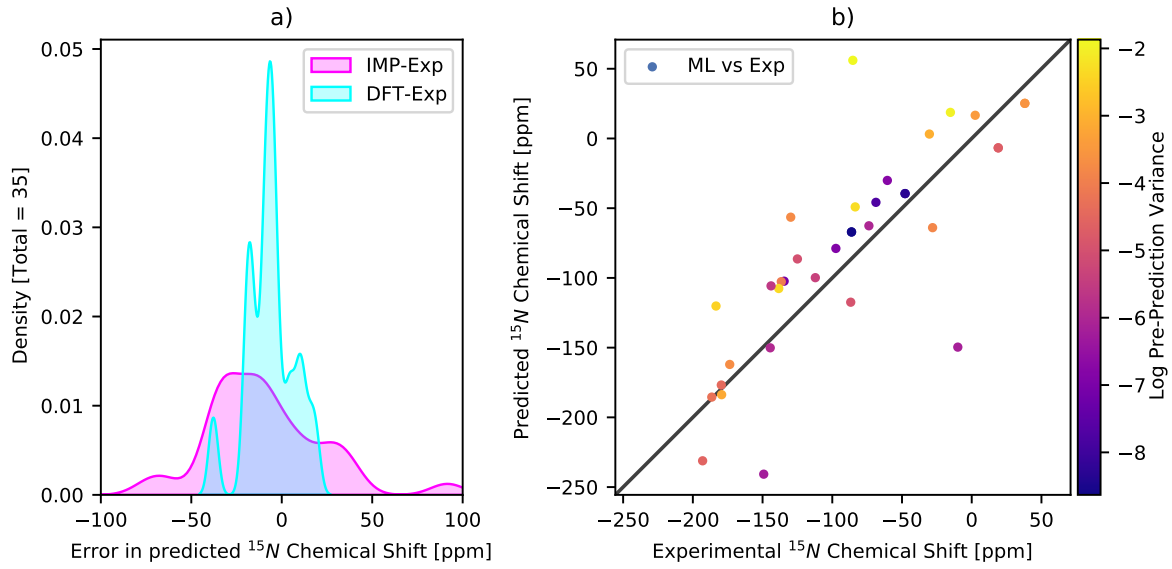

Figure 3.15: For the model trained using dataset 4 only. a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{15}N$ relative to the experimentally measured values in experimental dataset 2 (DTe2). b) IMPRESSION predicted and experimentally measured $\delta^{15}N$ with variance highlighted. Fit statistics for DTe2: 33.04 ppm MAE, 46.43 ppm RMSD, 141.3 ppm MaxE.

from an analysis where the accuracy of the predictions is important, and some loss of data is acceptable. An example of an analysis such as this is given in Chapter 5, and these issues are discussed further.

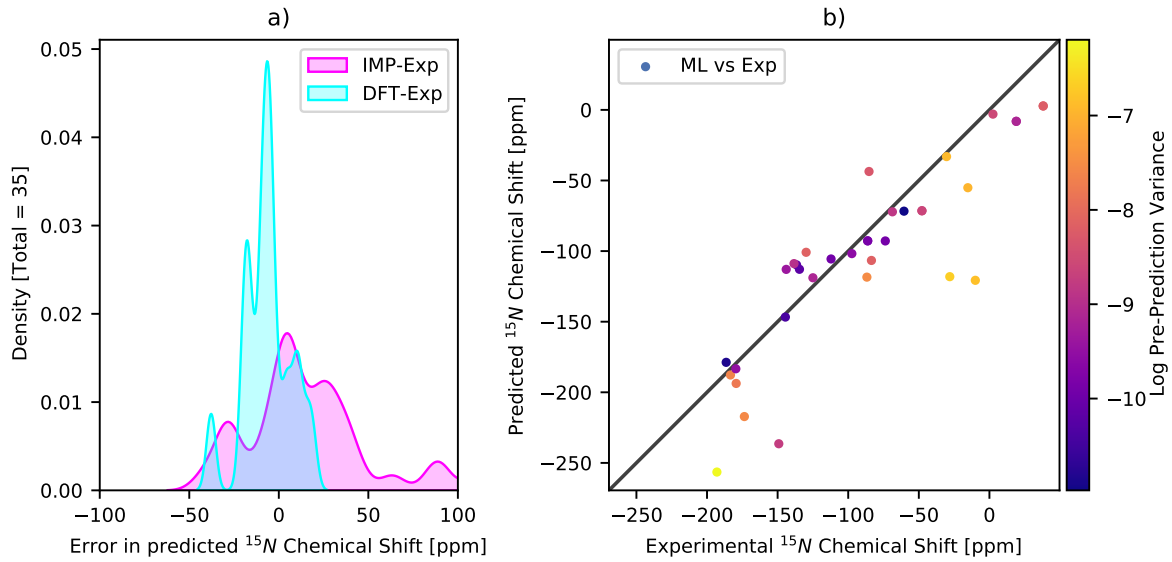

Figure 3.16: For the model trained on both dataset 4 and 5a. a) error distributions for both IMPRESSION predicted and DFT calculated $\delta^{15}N$ relative to the experimentally measured values in experimental dataset 2 (DTe2). b) IMPRESSION predicted and experimentally measured $\delta^{15}N$ with variance highlighted. Fit statistics for DTe2: 27.02 ppm MAE, 37.3 ppm RMSD, 110.7 ppm MaxE

3.2.5 $^{1}J_{CH}$ Prediction

3.2.5.1 Performance relative to DFT

The IMPRESSION generation 1 model trained on DT4 (772 molecules, 30,324 $^{1}J_{CH}$ environments, Section 2.4.3) achieved a mean absolute error (MAE) of 1.12 Hz and a root mean squared deviation (RMSD) of 1.71 Hz against the CSD derived test set DT3 (306 molecules, 5,262 $^{1}J_{CH}$ environments, Section 2.4.2). The maximum error in this set of predictions was 60.9 Hz. On the ChEMBL derived test set dataset 5b (400 molecules, 10,641 $^{1}J_{CH}$ environments, 2.4.4) the model achieved an accuracy of 1.83 Hz MAE, 3.20 Hz RMSD with a maximum error of 19.3 Hz.

The accuracy of the $^{1}J_{CH}$ predictions are also affected by the inclusion of nuclei not seen in the training dataset, as can be seen from the reduction in outlying values for DT5b in Figure 3.18a and the increase in height of the peak in Figure 3.18b.

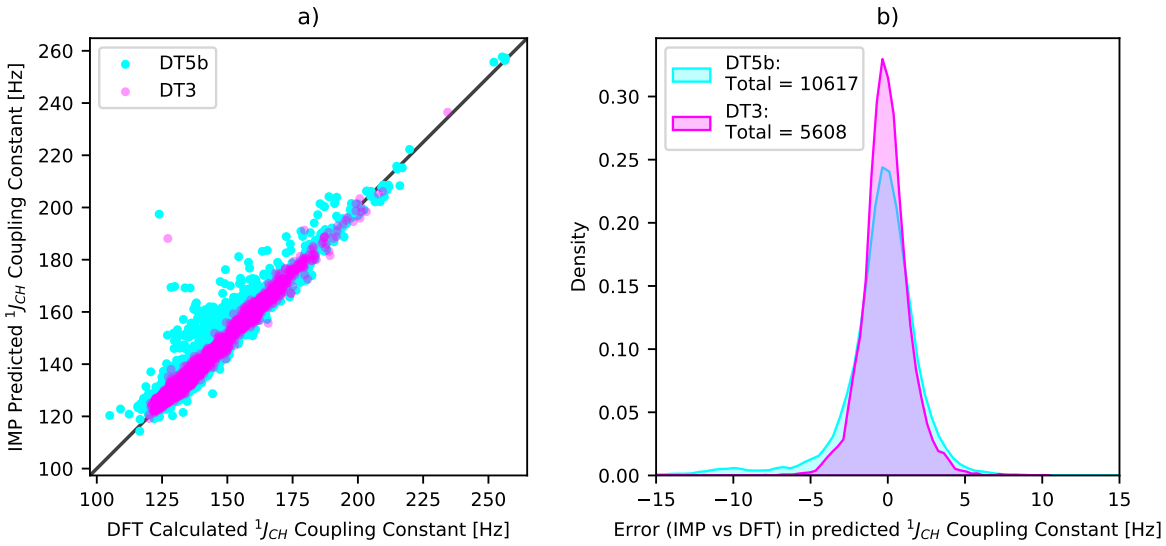

Figure 3.17: a) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, 1 (DT3) and 129 (DT5b) values excluded from graph for clarity. Models trained using dataset 4 (DT4)

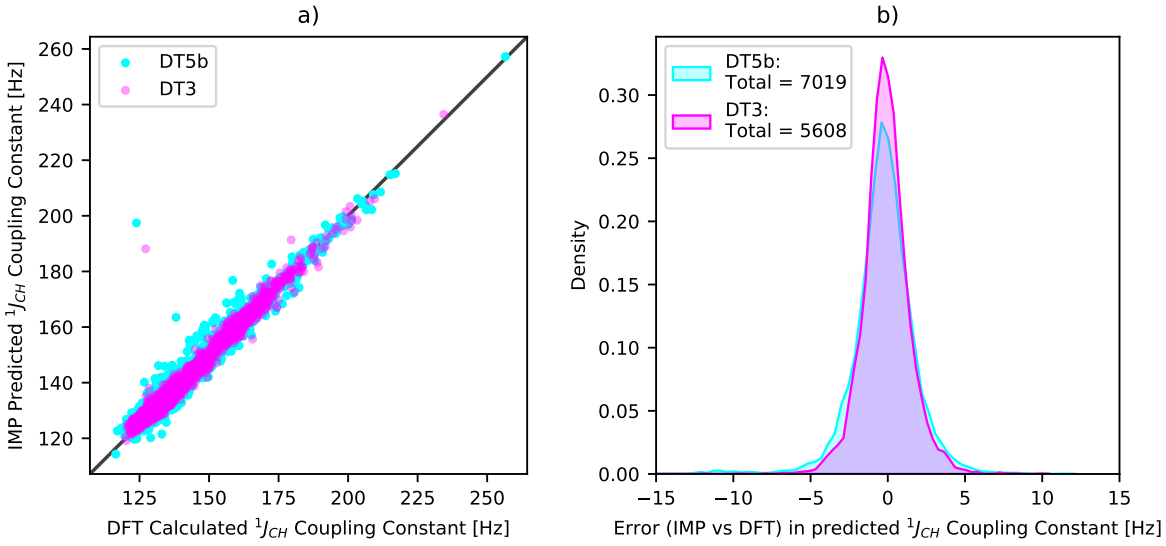

Figure 3.18: Molecules containing H,C,N,O,F atoms only. a) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.43 Hz MAE, 2.26 Hz RMSD, 19.3 Hz MaxE. b) Error distributions between IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, 2 (DT3) and 159 (DT5b) values excluded from graph for clarity. Models trained using dataset 4 (DT4)

3.2.5.2 Uncertainty Estimation

For the IMPRESSION model trained using DT4 to predict $^1J_{CH}$, the pre-prediction variance correlates with the prediction error for both DT3 and DT5b (with linear correction applied) (Figure 3.19a), and the two largest errors from DT3 and DT5b are associated with a higher variance (bright yellow points, Figure 3.19b). Many of the environments for the molecules in dataset 5b which contain atom types not present in the training set were also identified, as the region in Figure 3.19b highlighted in orange at 130-140 ppm (DFT calculated $^1J_{CH}$) corresponds with the region missing from the graph showing the reduced testing set in Figure 3.18. Once again the pre-prediction variance provides a method of identifying poorly predicted environments at the point of prediction.

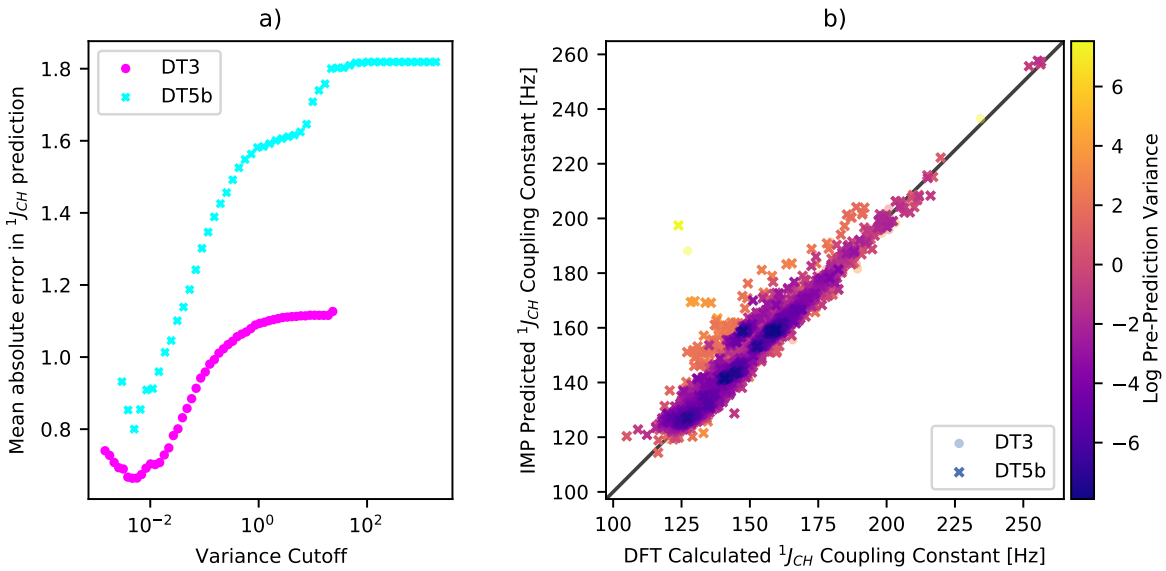

Figure 3.19: Linear correction applied to DT5b DFT values. a) Error in predicted $^1J_{CH}$ for populations with different maximum variance for the DT3 and DT5b testing sets. b) IMPRESSION predicted and DFT calculated $^1J_{CH}$ for the DT3 and DT5b testing sets, with variance values highlighted. Fit statistics for DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, fit statistics for DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. Models trained using dataset 4 (DT4)

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
0.001	5,475	0.746	0.838	1.673	0.921
0.005	4,936	0.659	0.812	2.977	1.490
0.01	4,448	0.705	0.897	3.416	1.992
0.05	2,688	0.865	1.124	4.696	3.120
0.1	1,694	0.953	1.247	5.797	3.693
0.5	331	1.064	1.406	7.840	4.352
1	95	1.092	1.457	9.987	4.726
5	11	1.113	1.499	11.814	5.004
10	4	1.116	1.507	11.814	5.071
50	0	1.127	1.713	60.920	5.641

Table 3.10: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 3. Total number of $^1J_{CH}$ environments in DT3: 5,608

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
0.005	10,098	0.802	1.607	11.626	2.268
0.01	9,614	0.909	1.683	11.626	3.616
0.05	7,047	1.182	1.936	18.824	8.329
0.1	5,014	1.322	2.094	18.824	10.375
0.5	1,300	1.540	2.328	18.824	11.474
1	602	1.583	2.377	18.824	11.579
5	185	1.618	2.436	18.824	11.950
10	85	1.708	2.716	26.928	14.441
50	4	1.809	3.058	35.431	18.047
100	1	1.819	3.117	40.978	18.710
500	1	1.819	3.117	40.978	18.710
1000	1	1.819	3.117	40.978	18.710
5000	0	1.825	3.198	73.556	19.308

Table 3.11: Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 5b. Total number of $^1J_{CH}$ environments in DT5b: 10,641

3.2.5.3 Performance relative to experiment

The error between the IMPRESSION predicted and the experimentally measured values in experimental dataset 2 (DTe2) is 6.01 Hz MAE, 11.18 Hz RMSD, with a maximum error of 54.3 Hz. The error between the DFT values and experiment is 2.16 Hz MAE, 3.23 Hz RMSD, 20.05 Hz MaxE for this same dataset. The error between IMPRESSION and DFT is 5.85 Hz MAE, 10.77 Hz RMSD, 54.8 Hz MaxE. This accuracy is disappointing relative to the accuracy achieved on DT3 and DT5b (post-correction), however this lack of accuracy is almost exclusively due to the poor prediction of the DFT values in this case, as the prediction error relative to experiment is only slightly higher than that relative to the DFT values. The error distribution (Figure 3.20a) between the IMPRESSION model and the experimental values shows that for the bulk of the predictions the accuracy is good, but a considerable number of environments the value is overpredicted by 10 Hz to 20 Hz.

Most of these large errors are associated with high pre-prediction variance (brighter yellow points in Figure 3.20b), and applying a modest variance filter of 10, removing 100 out of 721 environments, reduces the MAE to 3.12 Hz, the RMSD to 5.70 Hz and the maximum error to 29 Hz. Comparing the accuracy for DTe2, DT3 and DT5b for environments with pre-prediction variance less than 10, the results are much more similar. Though this does not improve the accuracy of the model, it at least achieves similar accuracy for environments that the model is similarly confident in predicting, if the pre-prediction variance is equated to a confidence in this situation.

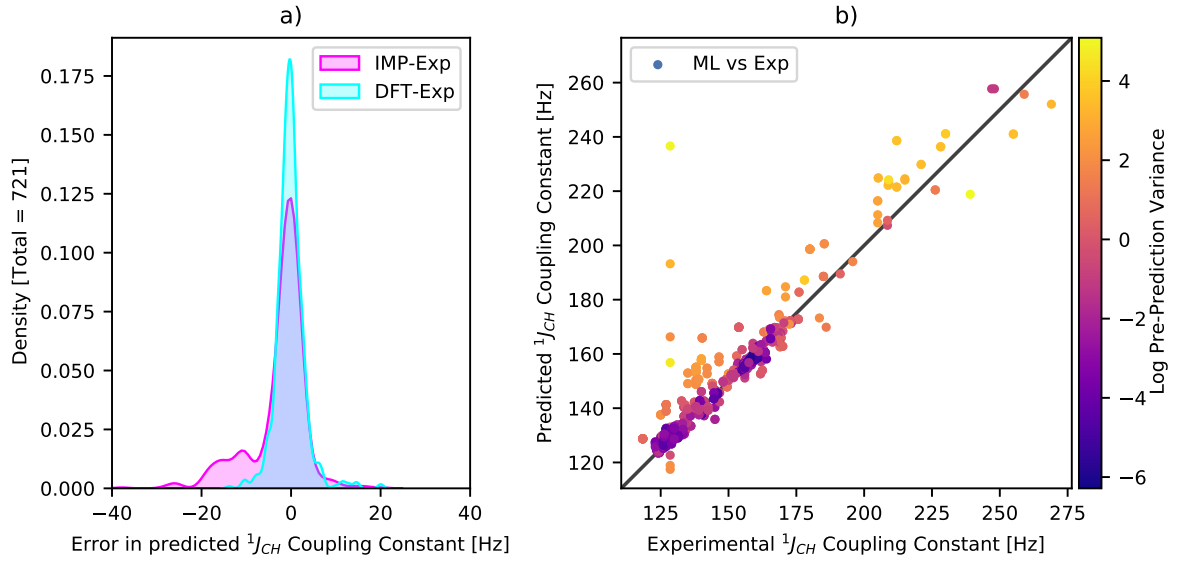

Figure 3.20: (a) error distributions for both IMPRESSION predicted and DFT calculated $^1J_{CH}$ relative to the experimentally measured $^1J_{CH}$ values for experimental dataset 2 (DTe2). (b) IMPRESSION predicted and experimentally measured $^1J_{CH}$ for DTe2 with variance highlighted. Fit statistics for DTe2: 6.01 Hz MAE, 11.18 Hz RMSD, 54.3 Hz MaxE. Models trained using dataset 4 (DT4)

3.3 Conclusion

The first generation IMPRESSION models predict δ^1H, $\delta^{13}C$, $\delta^{15}N$, and $^1J_{CH}$ DFT NMR parameters to an accuracy of between 0.5% and 3% of the range of the respective parameter, when tested against the DT3 and DT5b testing datasets.

The predictions are sufficiently accurate relative to experimentally measured values to be used in place of the underlying DFT method in certain circumstances. Furthermore the use of pre-prediction variance to highlight potentially innacurate predictions increases the utility of the predictions in practical applications.

The $^1J_{CH}$ prediction model in particular presents a unique improvement on existing models in the literature, being one of the first models to predict scalar coupling constants for 3-Dimensional molecules. The $\delta^{15}N$ model is also the first of its kind, and the δ^1H and $\delta^{13}C$ prediction models are competitive with the best and most recently published machine learning models in the literature.

Attempts to design a new generation of models, detailed in the next chapter, comprised the majority of the remainder of the work for this thesis.

CHAPTER 4

IMPRESSION GENERATION 2

4.1 Model Architecture and Training

The second generation NMR prediction model is based on recent advances in neural network architecture, and solutions generated by competition participants as part of a Kaggle competition to predict scalar coupling constants, and published in recent work by Bratholm et al [80]. The model is best described as a Graph Transformer Network (GTN) where molecules are represented as fully connected graphs. The transformer architecture is based on a mechanism called attention, originally developed for natural language processing [126]. Attention allows learnable weighting of different parts of the input data, effectively allowing a model to create its own input representation. In this case attention offers a clear advantage, as the best method of representing small molecules in machine learning problems is an unsolved problem, and so a machine-constructed representation may outperform existing methods.

4.1.1 Kaggle Competition

Kaggle is a popular website through which machine learning competitions are organised. The competitions cover a diverse range of subjects, training models to do everything from distinguishing ships and icebergs to predicting annual sales figures for restaurants. Lars Bratholm organised a kaggle competition in 2019 where the task was to accurately predict scalar coupling constants in molecules in the QM9 dataset, given training data from a subset of the same dataset.

The competition yielded many excellent solutions, with some central themes. Most of the top 10 solutions used some form of graph representation for the molecules, and many of them used some form of an attention mechanism to adaptively learn the representation of the molecules. Several of the highest scoring models are described in a recent publication [80].

The top solutions provided an initial starting point for the models produced as the second generation of IMPRESSION models. The crucial limitation of the solutions presented in the kaggle competition is that they were designed to train on and predict very small molecules (from QM9), and so required significant adaptation to work on larger molecules.

4.1.2 Molecules as Graphs

In a GTN, molecules can be represented as computational graphs, with each atom represented by a node. The graphs used in the models in this chapter are fully connected; there is an edge between every pair of nodes in the graph. The graph also contains node and edge features which are vectors associated with each node or edge in the graph. The only node feature used in this case was the atom type. The edge features used were the distance between and the number of bonds connecting the two atoms, as well as a numeric label representing the type of coupling the edge represents. The numeric label, referred to as the coupling type, is constructed by creating a list of all possible coupling constant labels ($^1J_{CH}$, $^2J_{HH}$, $^3J_{HH}$, etc) then assigning integer values to each item in the list, for example a coupling type label '0' refers to a $^1J_{CH}$ coupling. Due to the fact that edges are directional within the deep graph python library [127] used, each coupling is represented as a pair of edges. The feature vectors are combined into a single vector for each node and edge.

4.1.3 Graph Transformer Network (GTN)

The model architecture used in the second generation of IMPRESSION models is most similar to the Graph attention network [128] with gated residual connections [129], presented as solution 4 in the work by Bratholm et al [80]. A significant portion of the code from that solution was used to create the model, along with significant help from the original authors of the model.

Despite the models excellent performance in the prediction of NMR parameters for QM9 molecules, several important structural features in the model needed to be adapted to improve the performance on larger molecules. This came at the expense of some accuracy on the QM9

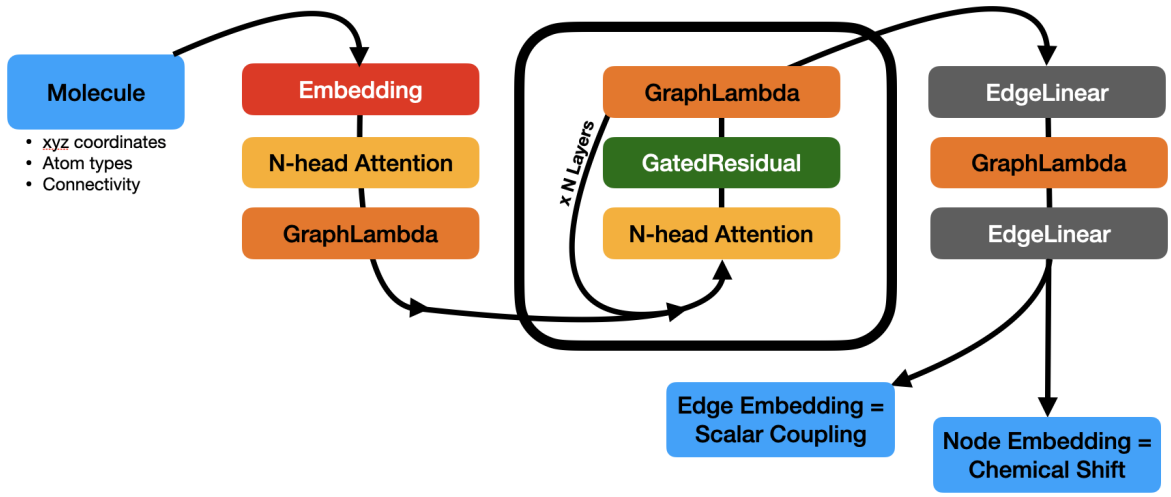

Figure 4.1: Simplified graph transformer network diagram.

molecules. Firstly it was found that fully connected graphs performed better than the graphs connected by only 1, 2 and 3 bond edges in the original model. Secondly fewer edge and node features were used in the IMPRESSION model (as described above), more similar to the model in solution 2[80], the original model from solution 4 utilised electronegativity, first ionization energy, and electron affinity for each atom type, as well as atom Mulliken charge. For the edges bond length, bond angle, and the dihedral angle were all used. Several more minor architectural features and parameters were altered in order to improve performance, as part of hyper-parameter tuning common in machine learning model production.

The network itself consists of the following sequential layers:

1. Embedding layer
2. Attention layer
3. Gated (Parametric) Residual Connection (PReLU) layer (GraphLambda)
4. Gated Residual Attention Layer
5. PReLU Layer (GraphLambda)
6. Gated Residual Attention Layer
7. PReLU Layer (GraphLambda)

8. Gated Residual Attention Layer
9. PReLU Layer (GraphLambda)
10. Linear Layer operating on edges only
11. PReLU Layer (GraphLambda)
12. Linear Layer operating on edges only

Where the embedding layer (1) creates the feature vector for the nodes from the atom type embedding and for the edges as a linear concatenation of the NMR coupling type embedding, distance vector, and path length vector. The attention layers (2,4,6,8) apply multiple independent attention mechanisms (multi-head attention, described in reference [128]) to the feature vectors in each node and vector. The results of each separate attention mechanism k are concatenated to give the new node feature vector using the feature vectors of all other nodes, and edges from the current node to every other node. e and η are the input feature vectors, n and e are the output feature vectors for edges and nodes respectively. The new node vector n is given:

$$n_i = \|_{k=1}^{K} \sigma\left(\sum_j \left(\alpha_{ijk}\eta_{ik}\right)e_{ijk}\right) \tag{4.1}$$

where

$$\alpha_{ij} = softmax_j\left(\sigma\left(A[\eta_i \| e_{ij} \| \eta_j]\right)\right) \tag{4.2}$$

σ is the Leaky Rectified Linear Unit (LReLU), a linear activation function which allows for very small negative values. Above 0 the LReLU returns the input value, below 0 a slope is applied to reduce the magnitude of the negative value. In this model a slope value of 0.2 was used. η_i is the vector for the i-th node in the graph, e_{ij} is the vector for the edge connecting nodes i and j. A is a learnable weight vector. The edge vectors are updated based on the concatenation of the source node, edge node, and destination node feature vectors.

$$e_{ij} = W_i j\left[n_i \| e_{ij} \| n_j\right] \tag{4.3}$$

where W is a learnable weight vector. The Gated residual connection applied to each of the attention layers (4,6,8) is applied as described in [129], and provides a shortcut layers of the network for information to pass through. The Gated (Parametric) Residual Connection PReLU

is applied to the outputs of the residual connection [130], similar to the LReLU, the PReLU randomly varies the slope of the function below zero in each iteration. The two final linear layers are fully connected across all edge feature vectors. The tunable weights in each layer of the network affect the values stored in the embedding tensor in each graph, the predicted chemical shift and scalar coupling values are extracted from these tensors (chemical shift from node embedding, scalar coupling from edge embedding). The embedding in each model in each pass contains values for all nodes and edges, therefore effectively containing information for each chemical shift and scalar coupling value in the molecule. Far more accurate predictions were however obtained by training separate models to predict separate properties, and discarding the other properties in each model. The flow of information through the various layers is shown schematically in Figure 4.1

Relative to the reported model by Bratholm [80], fewer layers (3 from 6), layer dimension (24 from 48), and attention heads (12 from 24) were used. This was primarily to account for the increase in size of the molecules used for this work, with molecules of up to 150 atoms as opposed to a maximum atom count of 29 in the QM9 dataset used in the original work. With the dimensions used in the original model, the number of molecule graphs used in each iteration of the network training (batch size) needed to be reduced too much in order to fit the model into memory on the GPU units available, which had a maximum available memory of 12GB. A batch size of 16 with the above model dimensions was used.

4.1.4 Model Training

The tunable parameters in all layers were optimised using the modified version of the LAMB optimiser [131] reported in [80] where the weight decay term is decoupled from the trust region calculations similarly to the AdamW modification to the Adam optimiser [132]. The target values for all parameters are scaled and normalised prior to training, and the conversion factors stored for prediction output and to report a scaled training loss. The loss function uses the mean absolute error across all NMR parameter targets: $\delta^{1}H$, $\delta^{13}C$, $\delta^{15}N$, $\delta^{17}O$, $\delta^{19}F$, $^{1}J_{CH}$, $^{1}J_{CC}$, $^{2}J_{CH}$, $^{2}J_{CC}$, $^{2}J_{HH}$, $^{3}J_{CH}$, $^{3}J_{CC}$, $^{3}J_{HH}$. Models were trained concurrently on all NMR parameter targets. The training utilised a cyclical learning rate scheduler for the optimiser for the first 35 epochs of training, cycling between values of 0.001 and 0.01. At epoch 35 the learning rate is fixed at 0.001 for a further 65 epochs, for a total of 100 training epochs. Models were trained using training

dataset 4 (described in Section 2.4.3, dataset 4 and 5a (described in Section 2.4.4) combined (hereon referred to as DT45), and the QM960k dataset.

4.2 Results

4.2.1 Model Training

The models trained on DT4, DT45 and QM960k optimised well within the 100 epoch optimisation, with a reasonably stable loss for the final 50 epochs for both in-sample and out-of-sample loss (Figure 4.2). The in-sample loss is the mean of the mean absolute errors across all NMR parameters, all values having been scaled and normalised prior to training, across the molecules used in training. The out of sample loss is the same loss calculated for the testing dataset DT3 (Section 2.4.2). The models optimise equally well for all target NMR parameters (with the exception of $\delta^{15}N$ in the QM960k model training), as is clear from the out of sample loss, plotted by target parameter (Figure 4.3). Different NMR parameters achieve a different loss value, this is expected as some parameters will be more difficult to predict.

For the models trained on DT4 and DT45, the out of sample loss is lower than the in sample loss for the entire training process (Figure 4.2a and 4.2b). This is generally unusual in machine learning models, however here this can be readily explained by the complexity of environments in dataset 4 (a result of the adaptive sampling process) and the therefore relatively simple set of environments in the dataset 3 testing dataset. This result is not seen in the model trained on QM960k (Figure 4.2c), as the model trained on the smaller molecules clearly struggles to predict parameters for the relatively larger molecules in dataset 3.

The learning curve for $\delta^{15}N$ in the QM960k trained model (Figure 4.3c) shows a clear difference to the curves for all other parameters and in all other models. In repeated model training curves this was also seen, and occasionally also seen for the $\delta^{17}O$ curve as well, though not in the final training runs. This would likely be solved by specific fine-tuning on this parameter alone, and this is recommended as part of future work.

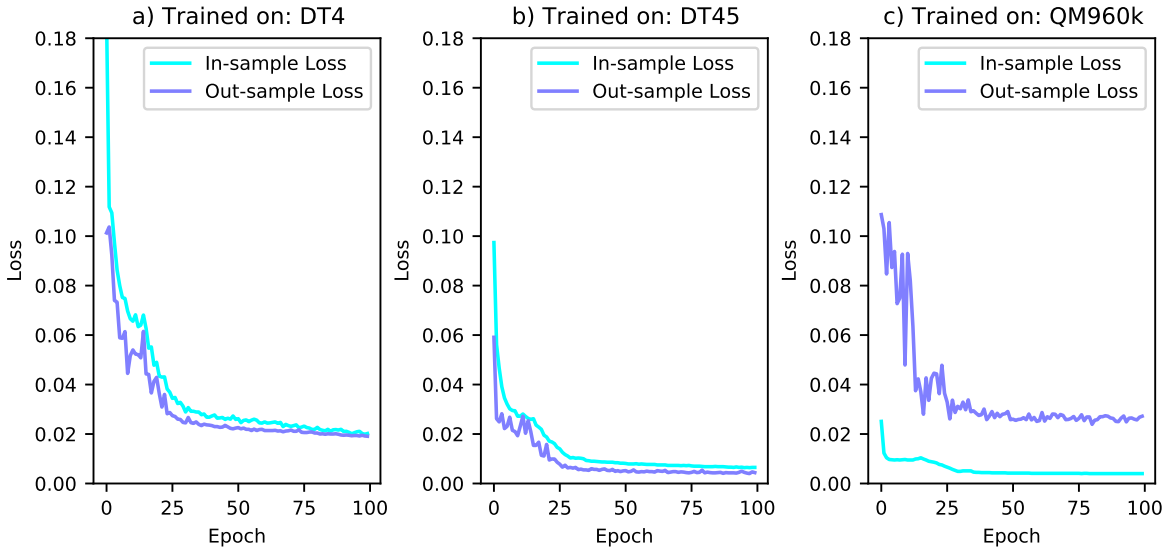

Figure 4.2: Out of sample for dataset 3 (DT3) and in sample loss during training for models trained on datasets 4 (DT4: a), 4 and 5a combined (DT45: b), and QM960k (c)

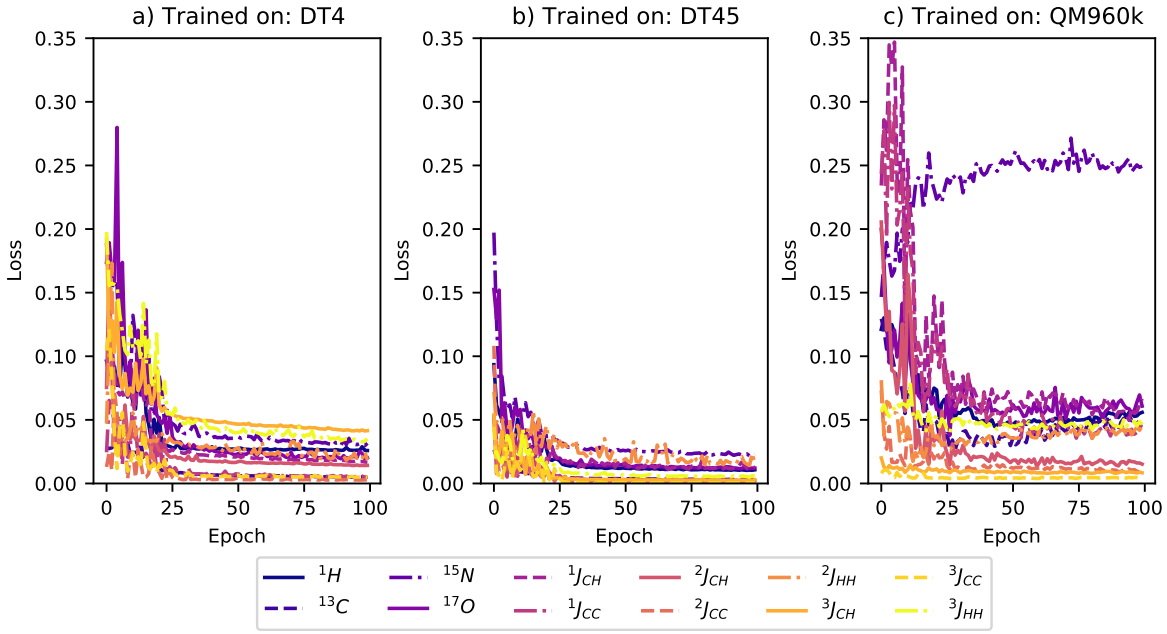

Figure 4.3: Out of sample for dataset 3 (DT3) loss split by target NMR parameter during training for models trained on datasets 4 (DT4: a), 4 and 5a combined (DT45: b), and QM960k (c)

4.2.2 Model Accuracy Summary

Figure 4.4 shows the accuracy in machine learning prediction across the three machine learning models trained (DT4, DT45, QM960k), against the three testing datasets (DT3: 4.4a, DT5b: 4.4b, QM91k: 4.4c) for two chemical shift parameters (δ^1H and $\delta^{13}C$) and two scalar coupling parameters ($^1J_{CH}$ and $^3J_{HH}$). Firstly, a consistent pattern is observed in terms of the accuracy across the different NMR parameters, with δ^1H being the most or nearly the most accurately predicted parameter relative to the range of values for each model across all testing datasets. The $^3J_{HH}$ couplings were the least accurately predicted for each model.

The QM9 trained model performed significantly worse in predicting the DT3 and DT5b datasets, but significantly better at predicting the QM91k dataset. Perhaps more surprisingly, the DT4 and DT45 models show the reverse pattern, performing worse on the QM91k dataset. The DT4/DT45 predictions on the QM91k dataset (Figure 4.4c) are more accurate than the QM960k trained predictions are on the DT5b datasets (Figure 4.4b), suggesting that the models trained on larger, drug-like, molecules appear to generalise better to the smaller molecule dataset than the models trained on smaller molecules generalise to the larger drug-like molecule testing dataset. The expected result would have been the DT4 and DT45 models performing similarly or better on the QM91k dataset than on the DT3 or DT5b testing datasets. The assumption is that QM91k contains smaller molecules and so covers a smaller region of chemical space, the type of structures in QM91k being also regularly found as part of larger molecules in the DT4 and DT5a datasets, and so the larger molecule datasets should allow a model to generalise to smaller molecules. What is clear from these results is that the absence of nearby atoms in a representation is an important structural feature, and so the inclusion of smaller molecules in a training set would be beneficial if prediction accuracy for smaller molecules is desired.

Whilst the QM960k trained model presents an excellent predictive accuracy on the QM91k test set (Figure 4.4c), surpassing results from recent publications in both $\delta^{13}C$ prediction [65] and $^3J_{HH}$ [67] (despite the issue with the 'mixed' option highlighted in Chapter 2), this accuracy is not replicated on the larger molecule datasets. Even though the performance of the QM960k trained model on DT3 (Figure 4.4a) is comparable to the DT4 and DT45 trained models performance on QM91k (Figure 4.4c), the accuracy of the QM960k trained model is 10 to 50 times worse on the larger datasets. This highlights the issue with using QM9 as a benchmark dataset, as molecules with fewer than 30 atoms (including H) are not regularly the subject of NMR studies, and the

accuracy achieved for QM9 trained models on QM9 datasets does not generalise to larger, more relevant molecules.

Due to the fact that the DT45 trained model attained a significantly better or similar accuracy than the DT4 trained model for every parameter in every testing dataset, analysis will be restricted to the DT45 and QM960k trained models for the remainder of this chapter. The QM960k trained models are not the best performing models, but represent the benchmark model considering the significant number of QM9 trained models presented in recent literature.

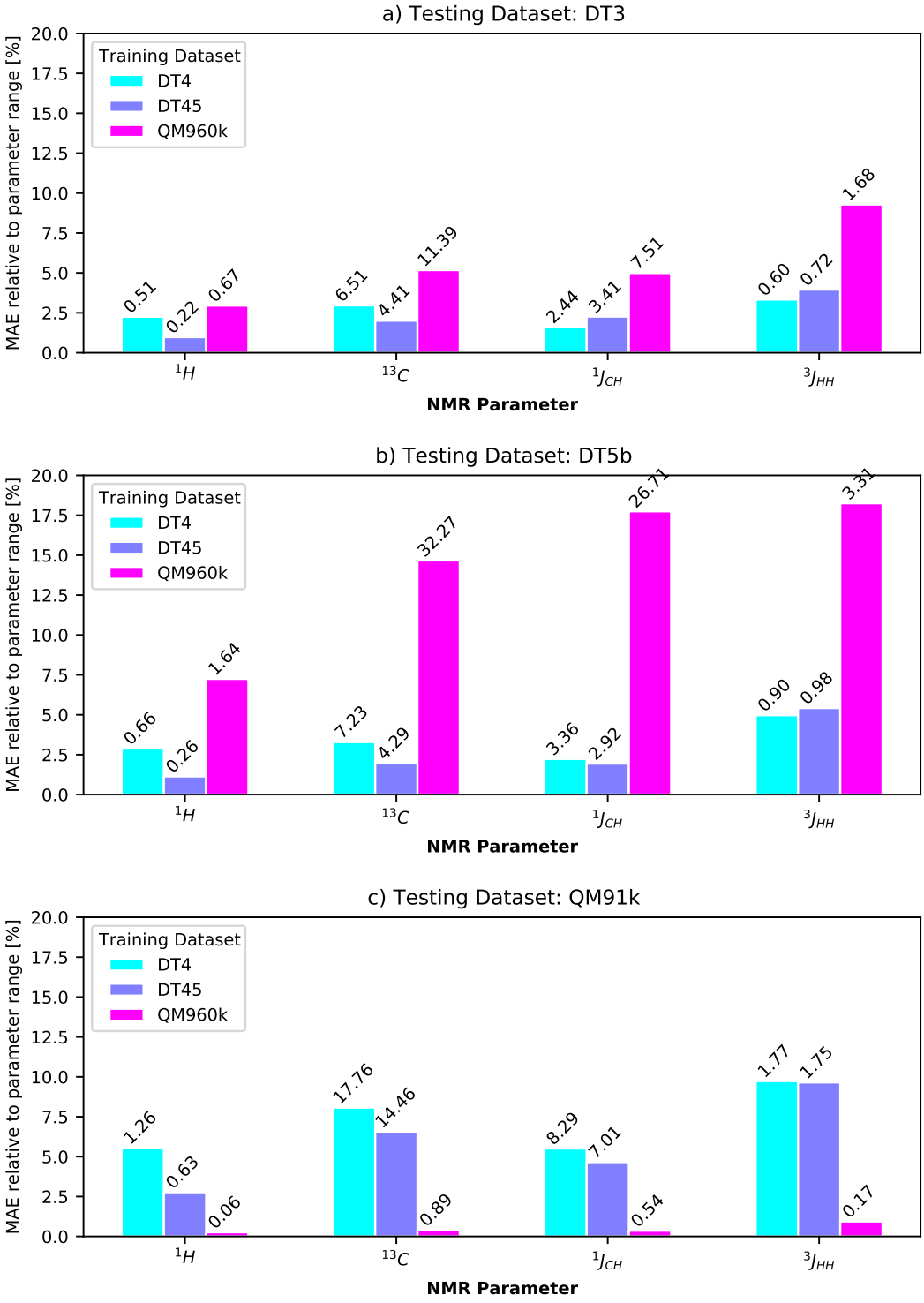

Figure 4.4: Comparison in model accuracy for testing datasets 3(DT3, a), 5b(DT5b, b), and QM91k(c). Bar height represents the mean absolute error as a percentage of the full range of values for that NMR parameter, each bar is annotated with the raw MAE values. The $^1J_{CH}$ and $^2J_{CC}$ bars for the model trained using QM9 in (b) are cut off for clarity, the relative MAE values are 42% and 28% respectively.

4.2.2.1 Uncertainty estimation

Following the same technique used for the generation 1 models (Chapter 3), 5 drop-out models were trained for each model and the variance across predictions in these 5 models appears to correlate well with the prediction error, specifically in terms of the reduction in large errors (Figure 4.5). The variance across the QM960k trained drop-out models tested against the QM91k dataset shows the weakest correlation, which is unsurprising due to both the high accuracy in these predictions, and the lack of chemical diversity within the QM9 dataset.

There is little variation between the correlation between pre-prediction variance and the largest prediction errors. The utility of the pre-prediction variance for each parameter will be discussed below, as the important factor not shown in Figure 4.5 is what percentage of the dataset needs to be removed in order to achieve the improvement in error shown. The important point here is that there is a correlation between poorly predicted NMR parameters and their pre-prediction variance, however whether this correlation can be utilised effectively depends on several factors including the number of large errors associated with a small pre-prediction variance relative to those with a large pre-prediction variance (discussed further in this chapter), and the nature of the specific task the models are being used for (discussed further in Chapter 5).

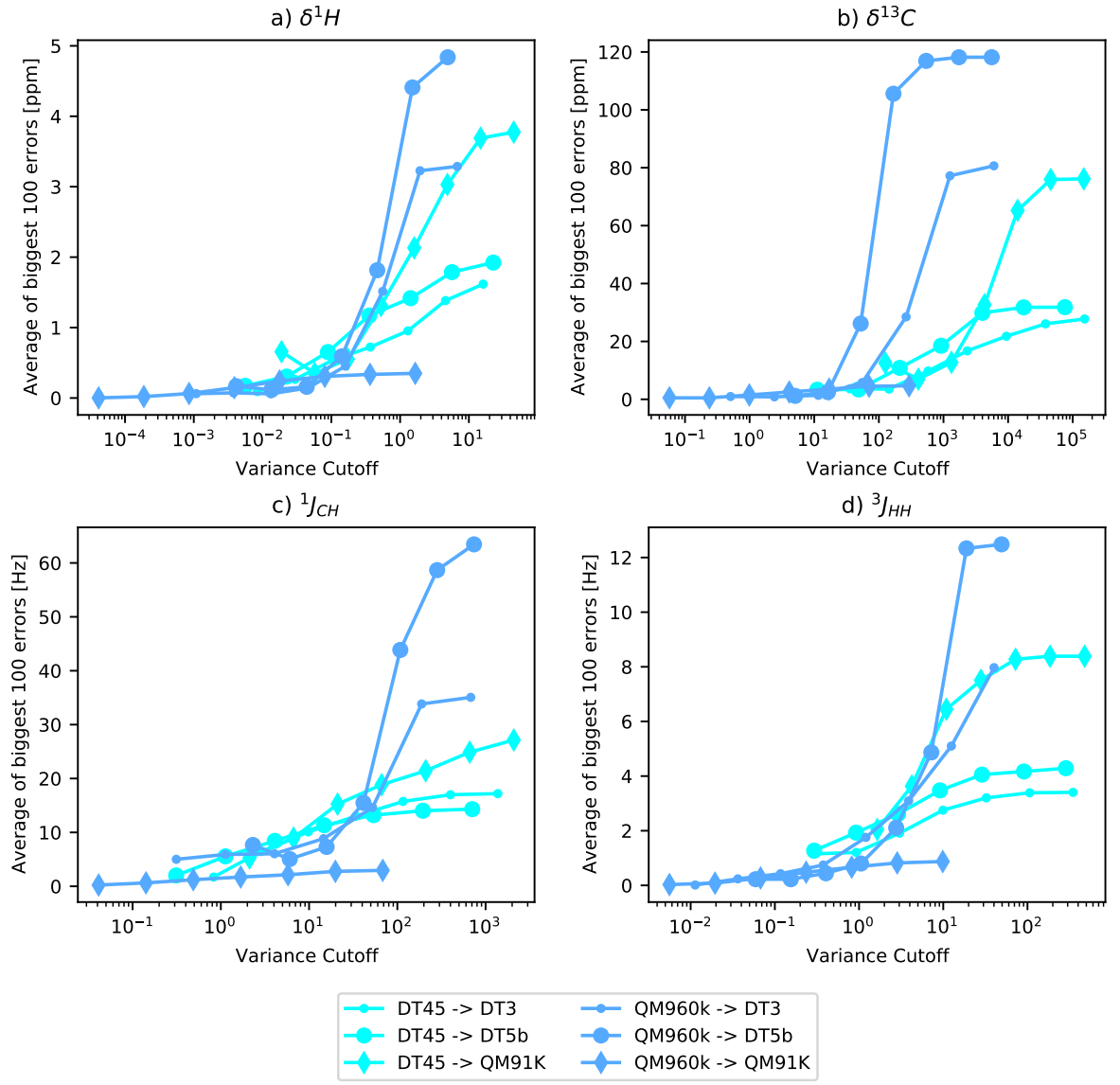

Figure 4.5: Mean absolute error for the worst predicted 100 environments for selected NMR parameters: δ^1H, $\delta^{13}C$, $^1J_{CH}$, $^3J_{HH}$. Errors presented for models trained on datasets 4 and 5 combined (DT45) and QM960k, tested on datasets 3 (DT3), 5b (DT5b), and QM91k.

4.2.3 δ^1H prediction

4.2.3.1 Performance relative to DFT for δ^1H prediction

The generation 2 model trained using DT45 (combination of dataset 4: Section 2.4.3 and dataset 5a: Section 2.4.4, 2,372 molecules, 66,805 δ^1H values) model achieves an accuracy of 0.22/0.25 ppm MAE, 0.36/0.36 ppm RMSD, 8.0/6.0 ppm MaxE when tested on the DT3 (306 molecules,

5,905 δ^1H values, Section 2.4.2) and DT5b (400 molecules, 11,885 δ^1H values, Section 2.4.4) testing datasets respectively (Figures 4.6a and 4.8a). When tested against the QM91k testing dataset (752 molecules, 6,949 δ^1H values, Section 2.4.5), the accuracy is 0.64 ppm MAE, 1.00 ppm RMSD, 6.30 ppm MaxE (Figures 4.6b and 4.8b. These results are summarised in Table 4.1. The predictions for this model slightly improve on those of the KRR based IMPRESSION generation 1 prediction model (Trained on DT4 only) where the mean absolute error was 0.24 ppm and 0.34 ppm for DT3 and DT5b respectively. Whilst this accuracy improvement can be explained primarily by the increase in training dataset size, rather than by any improvement in architecture, the ability to use larger datasets is one of the key advantages of using neural network type architectures over kernel ridge regression.

Training Dataset	Testing Dataset	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE as % of Range
DT4 (Gen 1)	DT3	0.243	0.388	4.268	1.044
DT4 (Gen 1)	DT5b	0.343	0.542	8.785	1.480
DT45	DT3	0.221	0.355	8.007	1.124
DT45	DT5B	0.246	0.364	5.957	1.224
DT45	QM91K	0.628	0.982	6.290	5.295
QM960k	DT3	0.670	1.001	9.053	3.399
QM960k	DT5B	1.645	1.973	9.683	8.196
QM960k	QM91K	0.059	0.085	1.561	0.494

Table 4.1: Accuracy in δ^1H prediction across the three testing datasets, for the DT45 and QM960k trained models. as well as the generation 1, KRR model

The QM960k trained model achieves an accuracy of 0.06 ppm MAE, 0.09 ppm RMSD, 1.6 ppm MaxE when tested against the QM91k test set (Figures 4.7b and 4.9b), and 0.67/1.65 ppm MAE, 1.00/1.97 ppm RMSD, 9.05/9.68 ppm MaxE when tested against the DT3 and DT5b testing datasets respectively (Figures 4.7a and 4.9a). The accuracy of the QM960k trained model is, as expected, far better than the DT45 trained model on the QM91k dataset, this can be seen most clearly in the difference between figures 4.6b and 4.7b.

The performance of the QM960k trained model on the DT3 testing dataset is significantly worse than the DT45 trained model, with a mean absolute error 3-4 times larger (0.67 ppm vs 0.22 ppm MAE). The performance on DT5b is even worse, where the mean absolute error is 7-8 times larger. The difference in performance against DT3 and DT5b is especially clear in the error distribution in Figure 4.8a, when contrasted with the relatively similar error distributions across

all three datasets for the model trained on DT45 (Figure 4.8).

This demonstrates that, for δ^1H, using larger molecules, even in a smaller training dataset presents a clear advantage over using a larger dataset of smaller molecules. The QM960k training dataset contains 565,420 δ^1H values, compared to only 66,805 values in DT45, yet the prediction accuracy for the the DT45 trained model generalises better to QM91k than the QM960k model generalised to DT3 or DT5b. It is also important to note that the molecules in ChEMBL represent drug-like molecules, for which the accuracy in NMR prediction is of far more use than for the QM9 molecules, and so the accuracy of the models on DT5b is of more importance in real-world applications.

4.2.3.2 Uncertainty estimation for δ^1H prediction

Both the QM960k and DT45 trained models were able to identify poorly predicted environments on the basis of the pre-prediction variance, as shown by the outlying values highlighted in brighter yellow in Figures 4.6a, 4.6b, and 4.7a. This correlation is not seen for the QM960k trained model predictions on QM91k, though there are almost no outlying values to identify, primarily due to the lack of diversity in the chemistry of QM9 molecules. The DT45 trained model displays a correlation between pre-prediction variance and the δ^1H value (Figure 4.6), with higher values being predicted just as accurately, but with higher associate variance. Furthermore the largest error (DFT calculated δ^1H approximately -4.5 ppm) is associated with a relatively low variance. These two factors reduce the effectiveness of the pre-prediction variance in this case. It is clear from the distribution of points with very low variance (dark blue points in Figure 4.6) that the correlation between prediction error and pre-prediction variance is much stronger for the lowest pre-prediction variance values, as these are highly accurate. Unfortunately the practical application of the pre-prediction variance requires a strong correlation for high pre-prediction variance values, and so it is unlikely to be of use for the DT45 trained model in δ^1H prediction. For example, a variance cutoff of 5 ppm reduces the mean absolute error in δ^1H prediction from 0.221 ppm to 0.215 ppm for predictions on dataset 3, however this requires removing 852 environments, 14% of the total.

The QM960k trained model shows a stronger correlation between pre-prediction variance and prediction error for DT3 and DT5b, with no variation across the chemical shift range. The largest error in the dataset 3 predictions is removed as one of 3 environments removed with a 5

ppm variance filter, lowering the maximum error from 3.5 ppm to 3.3 ppm. The distribution in Figure 4.7 shows that pre-prediction variance can identify highly accurate predictions, however these make up a relatively small subset of the total. For environments in DT3 and DT5b with pre-prediction variance less than 0.1 (391 environments in DT3 and 194 environments in DT5b, 3.2% of the combined dataset), the QM960k model prediction error (0.24 ppm MAE, 0.47 ppm RMSD, 4.19 ppm MaxE) is comparable to that for the predictions for the DT45 model on the entirety of DT5b (0.25 ppm MAE, 0.36 ppm RMSD, 6.0 ppm MaxE). The QM960k model is therefore capable of defining environments at the point of prediction for which its accuracy will match those of the DT45 trained model. The effect of variance cutoffs on the accuracy of both models against both testing datasets are shown in Tables 4.2, 4.3, 4.4, and 4.5.

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors [ppm]
0.05	5571	0.183	0.240	0.884	0.380
0.1	5163	0.191	0.260	1.503	0.532
0.5	3580	0.190	0.263	1.646	0.766
1	2873	0.198	0.309	8.007	0.935
5	852	0.215	0.338	8.007	1.396
10	55	0.220	0.350	8.007	1.567
50	0	0.221	0.355	8.007	1.617

Table 4.2: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION δ^1H predictions against DFT calculations for dataset 3. Total δ^1H environments in DT3: 5905

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors [ppm]
0.05	11122	0.177	0.232	1.050	0.467
0.1	10232	0.198	0.274	2.191	0.711
0.5	5424	0.225	0.310	2.191	1.215
1	3150	0.234	0.322	2.389	1.311
5	144	0.252	0.361	5.211	1.721
10	26	0.256	0.372	5.211	1.853
50	0	0.258	0.377	5.211	1.923

Table 4.3: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION δ^1H predictions against DFT calculations for dataset 5b. Total δ^1H environments in DT5b: 11,885

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors [ppm]
0.05	5677	0.088	0.156	1.687	0.155
0.1	5514	0.113	0.191	1.687	0.277
0.5	4397	0.248	0.426	5.277	1.244
1	3081	0.435	0.735	5.277	2.604
5	3	0.668	0.994	6.819	3.288
10	0	0.670	1.001	9.053	3.349

Table 4.4: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION δ^1H predictions against DFT calculations for dataset 3. Total δ^1H environments in DT3: 5905

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors [ppm]
0.05	11794	0.161	0.230	0.700	0.161
0.1	11691	0.242	0.467	4.192	0.413
0.5	10859	0.530	0.810	4.192	2.029
1	8762	1.145	1.513	9.683	3.549
5	0	1.645	1.973	9.683	4.838

Table 4.5: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION δ^1H predictions against DFT calculations for dataset 5b. Total δ^1H environments in DT5b: 11,885

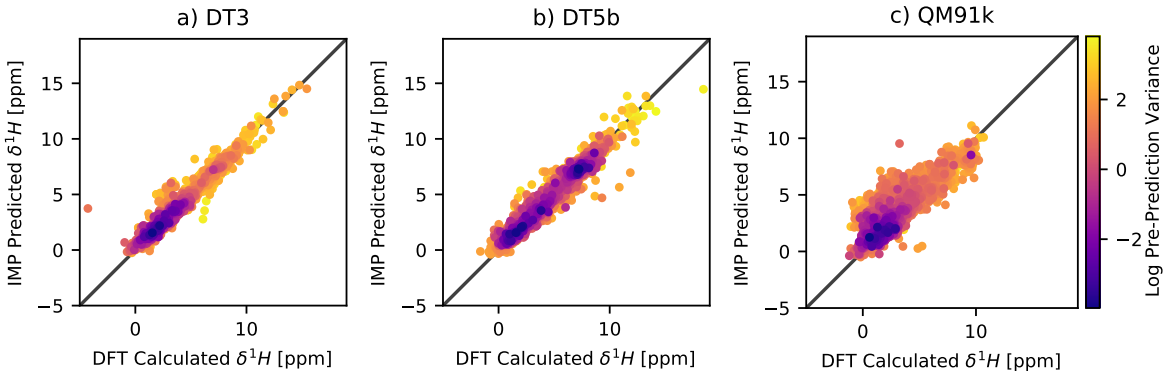

Figure 4.6: For the model trained using DT45: IMPRESSION predicted and DFT calculated δ^1H, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. DT3 fit statistics: 0.22 ppm MAE, 0.36 ppm RMSD, 8.0 ppm MaxE, DT5b fit statistics: 0.25 ppm MAE, 0.36 ppm RMSD, 6.0 ppm MaxE, QM91k fit statistics: 0.64 ppm MAE, 1.00 ppm RMSD, 6.30 ppm MaxE.

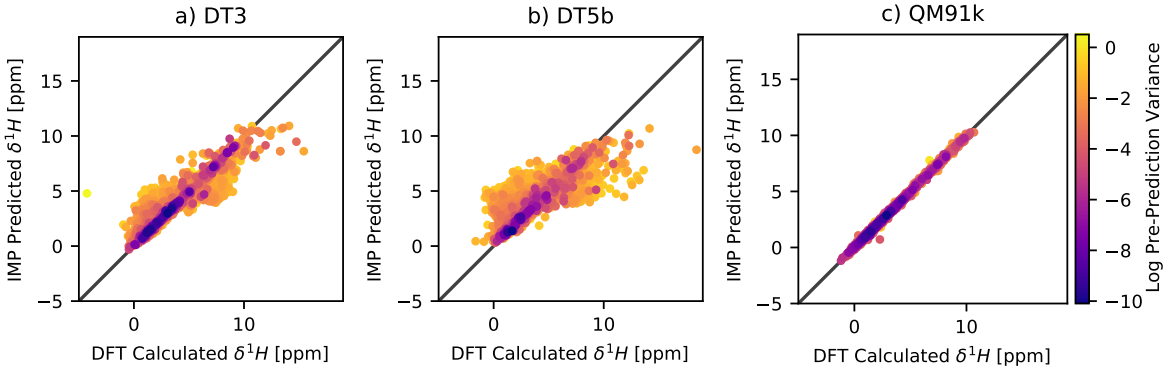

Figure 4.7: For the model trained using QM960k: IMPRESSION predicted and DFT calculated δ^1H, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. DT3 fit statistics: 0.67 ppm MAE, 1.00 ppm RMSD, 9.05 ppm MaxE, DT5b fit statistics: 1.65 ppm MAE, 1.97 ppm RMSD, 9.68 ppm MaxE, QM91k fit statistics: 0.06 ppm MAE, 0.09 ppm RMSD, 1.56 ppm MaxE.

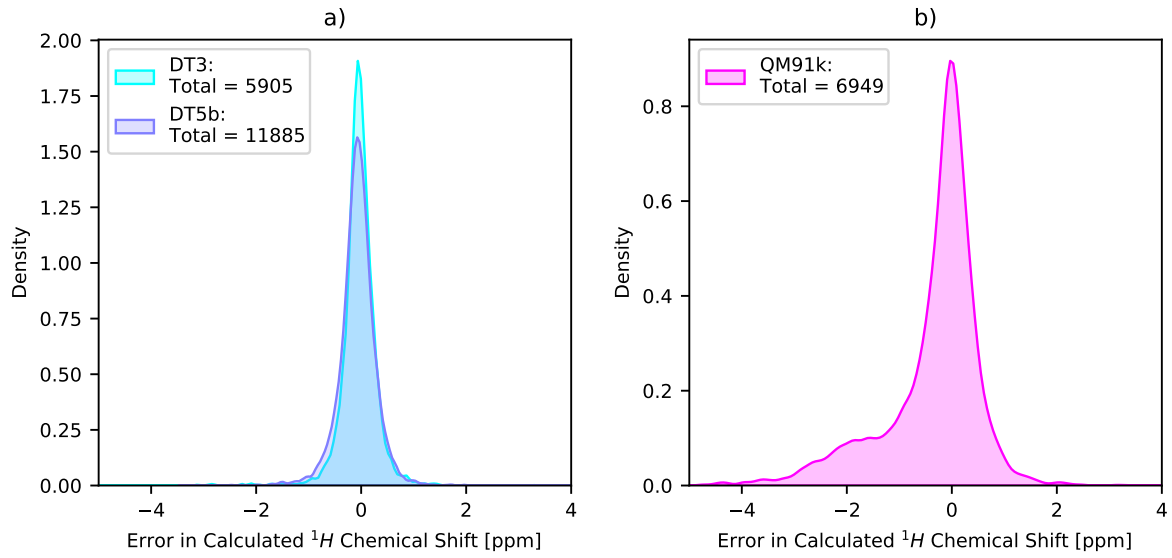

Figure 4.8: For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated δ^1H, for the DT3, DT5b (a) and the QM91k (b) testing datasets. DT3 fit statistics: 0.22 ppm MAE, 0.36 ppm RMSD, 8.0 ppm MaxE, DT5b fit statistics: 0.25 ppm MAE, 0.36 ppm RMSD, 6.0 ppm MaxE, QM91k fit statistics: 0.64 ppm MAE, 1.00 ppm RMSD, 6.30 ppm MaxE.

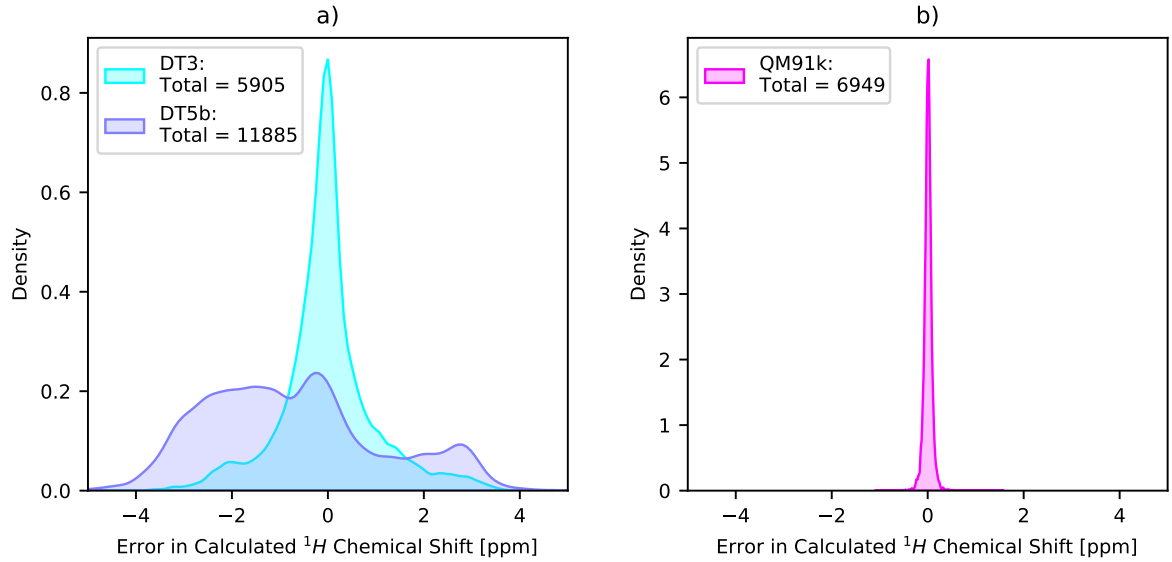

Figure 4.9: For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated δ^1H, for the DT3, DT5b (a) and the QM91k (b) testing datasets. DT3 fit statistics: 0.67 ppm MAE, 1.00 ppm RMSD, 9.05 ppm MaxE, DT5b fit statistics: 1.65 ppm MAE, 1.97 ppm RMSD, 9.68 ppm MaxE, QM91k fit statistics: 0.06 ppm MAE, 0.09 ppm RMSD, 1.56 ppm MaxE.

4.2.3.3 Prediction accuracy and molecule size for $\delta^1 H$ prediction

The QM91k dataset contains considerably smaller molecules than the DT5b dataset, it is also the case that DT4 and DT5a contain considerably larger molecules than the QM960k dataset. Therefore a key factor affecting the performance of the DT45 trained model on QM91k and the performance of the QM960k trained model on DT5b is likely to be the size of the molecules. Dividing the testing sets into subsets of molecules with number of atoms within a certain range, and plotting the rolling average of mean absolute error against the mean molecule size for each subset highlights the effect of molecule size on the prediction accuracy (Figure 4.10). The accuracy in predictions made by the DT45 trained model decrease as the molecule size reduces below 25 atoms, however is relatively stable for molecules with more than 30 atoms. For the QM960k trained model the opposite pattern is observed, with the prediction accuracy decreasing as molecule size increases for testing datasets 3 and 5b, but remaining stable for the molecules in testing dataset QM91k, where the maximum number of atoms is around 25.

This demonstrates that the prediction accuracy between the two models is more similar for similarly size molecules, however there is still a difference between the prediction accuracy for models trained on DT45 or QM960k, in predicting molecules from DT3, DT5b, or QM91k. This is visible in Figure 4.10 as the mean absolute errors for the subset with a mean molecule size of 20 atoms are different for each testing and training dataset combination.

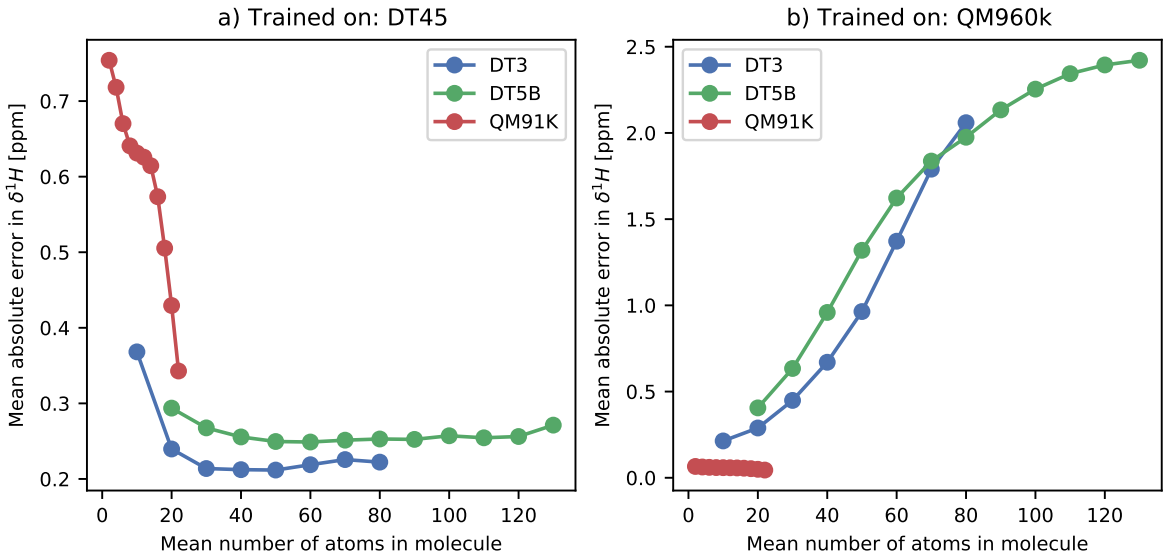

Figure 4.10: Accuracy in δ^1H prediction across the three testing datasets (DT3, DT5b, QM91k) for subsets of molecules with different size. For the model trained using DT45 (a) and the model trained using QM960k (b)

4.2.3.4 Performance relative to experiment for δ^1H prediction

The predictions on the δ^1H experimental dataset 1b (DTe1b, 46 molecules, 906 δ^1H values, 2.4.6.1) follow the pattern observed above for molecules with more than 20 atoms, namely that the DT45 trained model significantly outperforms the QM960k trained model. The accuracy for the DT45 trained model in predicting experimental δ^1H is 0.39 ppm MAE, 0.58 ppm RMSD, 2.86 ppm MaxE. This is more accurate than the generation 1 model trained on DT4 (0.44 ppm MAE, 0.61 ppm RMSD) however the maximum error is higher, it was 2.61 ppm in that case. The QM960k trained model accuracy is 0.78 ppm MAE, 1.10 ppm RMSD, 3.88 ppm MAE. The accuracy of the underlying DFT calculations in calculating the experimental δ^1H values is 0.33 ppm MAE, 0.50 ppm RMSD, 2.22 ppm MaxE. A summary of the prediction accuracy for DTe1b is shown in Table 4.6.

The DT45 trained model demonstrates a very good prediction accuracy relative to the experimental results in this case, with accuracy similar to the underlying DFT method. This suggests that this prediction model could replace the DFT NMR calculations and provide similarly accurate predictions in a fraction of the time. The similarity in prediction accuracy between the DT45 trained IMPRESSION model and the DFT calculations is highlighted by the similarity in error

distributions in Figure 4.11a.

Target	Training Dataset	MAE [ppm]	RMSD [ppm]	MaxE [ppm]
δ^1H	DT45	0.388	0.580	2.861
δ^1H	QM960k	0.776	1.098	3.879
δ^1H	DFT	0.326	0.496	2.218

Table 4.6: Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the δ^1H experimental test set (DTe1b).

For the DT45 trained model, the effectiveness of the pre-prediction variance in indicating prediction error is again hampered by the apparent correlation between variance and δ^1H value. Despite this it is clear from Figure 4.11b that a significant number of environments with high pre-prediction variance are associated with high error, however applying any variance filter removes too many accurate predictions to prove useful. The pre-prediction variance functions significantly better for the QM960k trained model, as indicated by Figure 4.11b. Removing environments with a variance greater than 1 (235 environments, approximately 30% of the dataset) improves the accuracy to 0.55 ppm MAE, 0.87 ppm RMSD, 1.76 ppm MaxE, which is a significant improvement, albeit at the cost of removing almost a third of the environments.

The poor performance of the QM960k trained model is highlighted by the increased width in the error distribution in Figure 4.12a, and the large number of outlying values in the scatter plot in Figure 4.11b. The relatively poor prediction accuracy can be partially explained by the size of the molecules in the test dataset (DTe1b). The prediction accuracy on those molecules from the test set with fewer than 40 atoms is 0.36 ppm MAE, 0.56 ppm RMSD, 2.73 ppm MaxE (307 environments). Conversely the prediction accuracy on those with greater than 40 atoms is 1.04 ppm MAE, 1.33 ppm RMSD, 3.88 ppm MaxE (486 environments). The accuracy on the smaller molecules is nearly identical to the accuracy of the DT45 trained model: 0.36 ppm MAE, 0.57 ppm RMSD, 2.86 ppm MaxE. The majority of the outlying values are also associated with a higher pre-prediction variance (brighter yellow points in Figure 4.11b), and the removal of prediction with pre-prediction variance higher than 0.5 improves the mean absolute error from 0.78 ppm to 0.40 ppm, similar to the prediction accuracy of the DT45 model. This variance cutoff does however remove 415 out of the 906 values in DTe1b.

Further to this, the two molecules shown in Figure 4.12 highlight the issues with training on small molecules. These two molecules are representative of an issue which causes the vertical line

of similar points at around 7.3ppm on the X-axis in Figure 4.12. The cause of the vast majority of these errors are aromatic protons in structures with complex 3D shape, which causes other parts of the molecule to be close in space to these protons. Such structures are highly unlikely to occur in the QM9 dataset due to the limited number of atoms in each structure. As a result, the representation which has been learned through training is incapable of dealing with these types of structures, and so incorrectly alters the prediction value based on some aspects of the structure which, as is clear from the experimental data, do not significantly affect the chemical shift of these protons.

The models therefore are similarly accurate in δ^1H prediction on molecules with size between 20 and 40 atoms, when these molecules are obtained from a different source to any of those used for training in either model. The models are also similarly accurate for environments with a similar pre-prediction variance. The advantage of the DT45 trained model is however clearly demonstrated here, as over the entire, unfiltered dataset it achieves a significantly better accuracy, nearly matching the accuracy of the DFT calculations used to train the model.

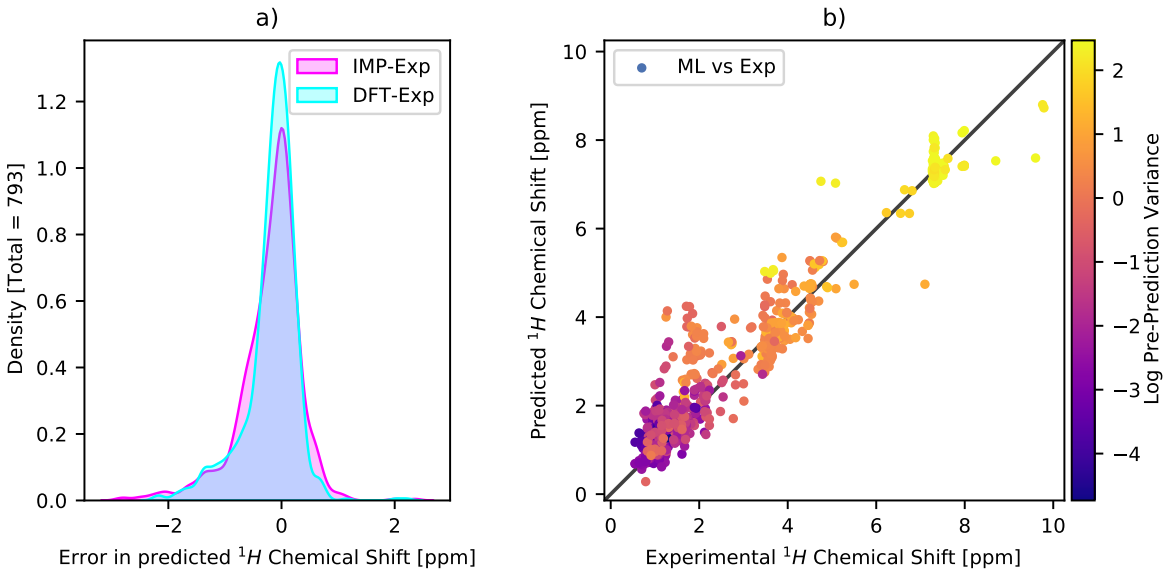

Figure 4.11: For the DT45 trained model predictions on the δ^1H experimental testing dataset DTe1b. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured δ^1H, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 0.39 ppm MAE, 0.58 ppm RMSD, 2.86 ppm MaxE.

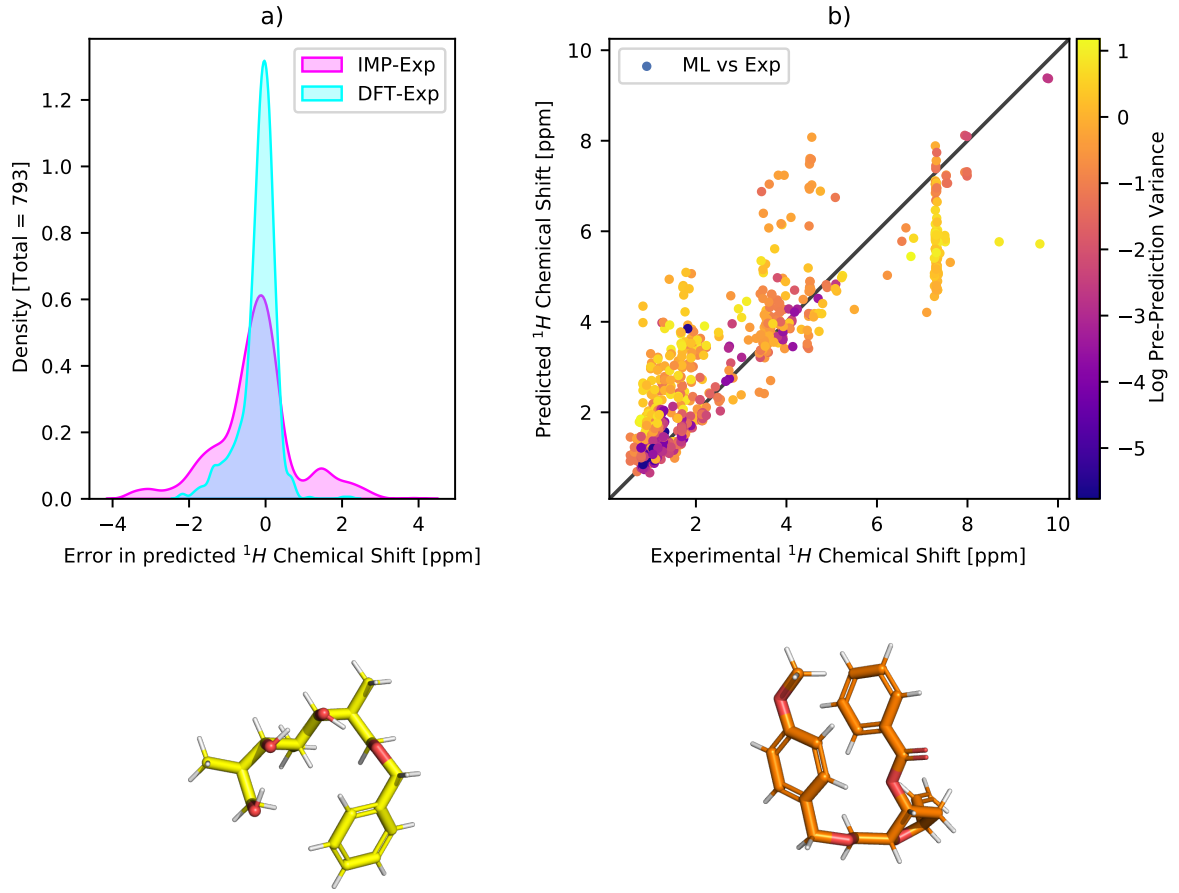

Figure 4.12: For the QM960k trained model predictions on the δ^1H experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured δ^1H, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 0.78 ppm MAE, 1.10 ppm RMSD, 3.88 ppm MaxE. The two structures are representative of the structures which cause the similar set of errors around 7.3ppm in (b).

4.2.4 $\delta^{13}C$ prediction

4.2.4.1 Performance relative to DFT for $\delta^{13}C$ prediction

The generation 2 model trained using DT45 (combination of dataset 4: Section 2.4.3 and dataset 5a: Section 2.4.4, 2,372 molecules, 56,349 $\delta^{13}C$ values) achieves an accuracy of 4.41/4.31 ppm MAE, 6.71/6.31 ppm RMSD, 90.8/64.1 ppm MaxE when tested on the DT3 (306 molecules, 5,262 $\delta^{13}C$ values, Section 2.4.2) and DT5b (400 molecules, 9,912 $\delta^{13}C$ values, Section 2.4.4) testing datasets respectively (Figure 4.13a, and Figure 4.15a).

The accuracy for the IMPRESSION generation 1 model (trained on DT4) is 3.5/6.34 ppm MAE, 7.05/17.1 ppm RMSD, 106.5/271 ppm MaxE for DT3 and DT5b respectively. The DT45 generation 2 model performs considerably better against the DT5b testing set than the generation 1 model, which is an expected result due to the inclusion of training data from DT5a, which contains molecules similar to those in DT5b. The generation 2 model trained using DT45 performs worse against DT3 than the generation 1 model trained using DT45 in terms of mean absolute error, but presents a small improvement in the root mean squared error and maximum error. This similarity suggests the change in model architecture between the two generations (KRR to GTN) provides no benefit in the prediction of $\delta^{13}C$ for DT3, however the fact that the generation 2 model retains similar accuracy on DT3 whilst now providing significantly improved predictions for DT5b demonstrates the advantage of the GTN framework. The GTN framework used in the generation 2 models allows for increased training dataset sizes, in this case via the inclusion of DT5a. This increase in training set size has produced a model capable of nearly replicating the predictions on DT3 for the generation 1 model, whilst expanding the same prediction accuracy to a far wider range of structures.

When tested against the QM91k testing dataset (752 molecules, 4,751 $\delta^{13}C$ values, Section 2.4.5), the accuracy for the DT45 trained generation 2 model is 14.8 ppm MAE, 22.3 ppm RMSD, 103.0 ppm MaxE (Figure 4.13b and Figure 4.15b). This follows the pattern seen across all predicted NMR parameters in this section, where the models trained using larger molecules struggle to provide accuracy predictions for the smaller molecules in QM91k. It is clear however that the DT45 trained model generalises better across the three testing datasets than the QM960k trained model, as is visible by the similarity in error distributions across Figure 4.15, and the relative dissimilarity between the three distributions in Figure 4.16.

The QM960k trained model achieves an accuracy of 0.88 ppm MAE, 1.26 ppm RMSD, 25.4 ppm MaxE when tested against the QM91k test set (Figure 4.14b and Figure 4.16b), potentially surpassing the accuracy (1.88 ppm MAE) reported in recent work on $\delta^{13}C$ prediction on QM9 molecules [65], though for a much larger QM9 testing dataset. The model performs poorly on testing DT3 and DT5b with accuracy of 11.39/32.27 ppm MAE, 19.56/45.90 ppm RMSD, 120.4/150.5 ppm MaxE (Figure 4.14a and Figure 4.16a). These results follow the pattern seen in the $\delta^{1}H$ predictions above, where the QM960k trained model achieves a very high accuracy on the QM91k testing dataset, but this model fails to generalise to the larger molecules. As

mentioned previously the larger molecules are more relevant to the practical application of these molecules, and so this suggests the QM960k trained model would be of little use in applications on larger molecules, though this will be discussed further in Chapter 5. The results for both models are summarised in Table 4.7.

Training Dataset	Testing Dataset	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE as % of range
DT4 (Gen 1)	DT3	3.496	7.052	106.5	1.523
DT4 (Gen 1)	DT5B	6.336	17.069	271.7	2.870
DT45	DT3	4.413	6.713	90.819	2.089
DT45	DT5B	4.305	6.312	64.130	1.970
DT45	QM91K	14.455	21.934	97.783	6.683
QM960k	DT3	11.388	19.559	120.392	5.391
QM960k	DT5B	32.269	45.902	150.470	14.762
QM960k	QM91K	0.891	1.286	25.411	0.412

Table 4.7: Accuracy in $\delta^{13}C$ prediction across the three testing datasets, for the DT45 and QM960k trained models. as well as the generation 1, KRR model

4.2.4.2 Uncertainty estimation for $\delta^{13}C$ prediction

The correlation between pre-prediction variance and prediction error shows a similar pattern as in $\delta^{1}H$ prediction, namely that there is a strong correlation for the most accurate predictions (darker blue points in Figures 4.13 and 4.14), but the relationship is less useful at identifying the largest errors (Outlying values are not consistently highlighted brighter yellow in Figures 4.13 and 4.14). The pre-prediction variance would likely be of little use in the application of the generation 2 $\delta^{13}C$ models, based on these results. The effect of variance cutoffs on the accuracy of both models against both testing datasets are shown in Tables 4.8, 4.9, 4.10, and 4.11.

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
10	5260	1.805	1.981	2.622	1.805
50	5246	3.109	3.941	9.135	3.109
100	5203	3.482	4.457	12.048	3.482
500	4815	3.773	5.253	22.400	9.321
1000	4448	4.159	5.745	28.044	12.368
5000	2995	4.483	6.618	78.920	19.873
10000	2207	4.495	6.603	78.920	21.838
50000	217	4.446	6.637	78.920	26.724
1e+08	0	4.413	6.713	90.819	27.755

Table 4.8: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 3. Total $\delta^{13}C$ environments in DT3: 5,262

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
5	9910	4.594	4.594	4.665	4.594
10	9901	2.756	3.223	4.902	2.756
50	9788	2.880	3.898	11.789	3.480
100	9504	3.102	4.396	30.748	7.219
500	6931	3.326	4.689	43.259	14.768
1000	4915	3.572	5.096	58.377	19.019
5000	1006	4.096	6.137	64.729	30.147
10000	333	4.217	6.335	66.176	31.407
50000	5	4.289	6.437	66.176	31.829
1e+08	0	4.290	6.437	66.176	31.829
5e+08	0	4.290	6.437	66.176	31.829

Table 4.9: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 5b. Total $\delta^{13}C$ environments in DT5b: 9,912

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
0.5	5259	0.966	1.134	1.779	0.966
1	5251	0.703	0.818	1.779	0.703
5	5212	0.890	1.204	3.307	0.890
10	5177	1.276	2.341	16.369	1.276
50	5017	2.653	6.518	78.169	5.449
100	4823	3.988	9.230	78.169	12.765
500	3562	7.707	14.930	100.454	51.313
1000	2459	10.040	18.882	120.392	73.536
5000	39	11.395	19.607	120.392	80.662
10000	0	11.388	19.559	120.392	80.662

Table 4.10: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 3. Total $\delta^{13}C$ environments in DT3: 5,262

Max Variance	No. Envs. Removed	MAE [ppm]	RMSD [ppm]	MaxE [ppm]	MAE of 100 largest errors[ppm]
5	9903	1.185	1.432	2.880	1.185
10	9888	2.789	4.641	18.153	2.789
50	9779	15.907	35.278	119.606	20.987
100	9569	29.632	49.789	143.785	89.171
500	6325	43.430	56.941	150.470	116.360
1000	3593	39.282	52.696	150.470	118.021
5000	6	32.283	45.915	150.470	118.178
10000	0	32.269	45.902	150.470	118.178

Table 4.11: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $\delta^{13}C$ predictions against DFT calculations for dataset 5b. Total $\delta^{13}C$ environments in DT5b: 9,912

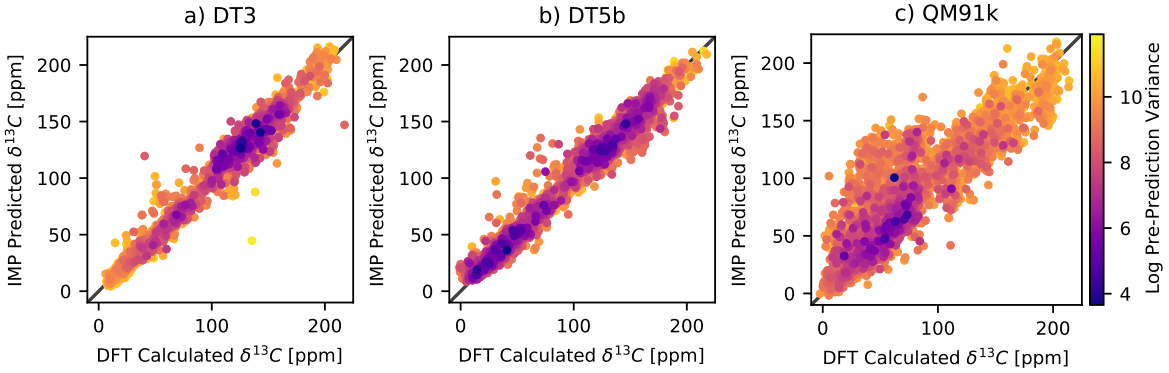

Figure 4.13: For the model trained using DT45: IMPRESSION predicted and DFT calculated $\delta^{13}C$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 4.41 ppm MAE, 6.71 RMSD, 90.82 MaxE, fit statistics for DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.13 ppm MaxE, fit statistics for QM91k: 14.5 ppm MAE, 21.9 ppm RMSD, 97.8 ppm MaxE.

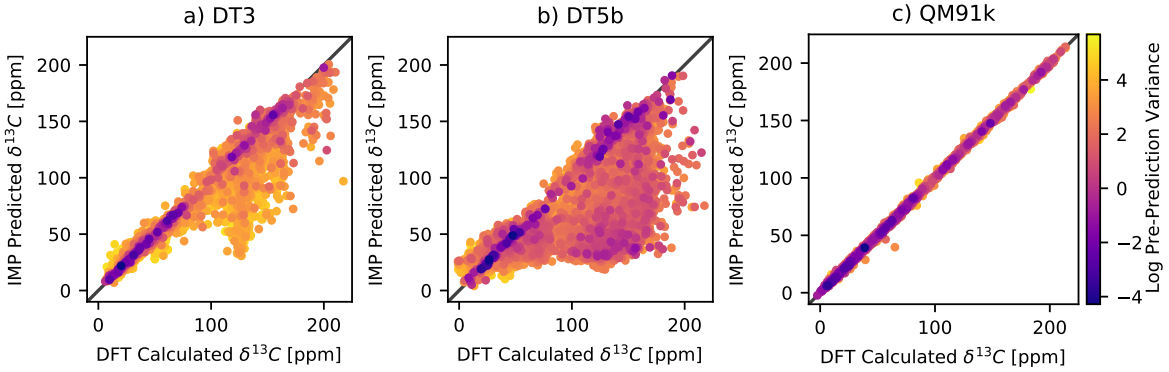

Figure 4.14: For the model trained using QM960k: IMPRESSION predicted and DFT calculated $\delta^{13}C$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 11.4 ppm MAE, 19.6 RMSD, 120.4 MaxE, fit statistics for DT5b: 32.3 ppm MAE, 45.9 ppm RMSD, 150.5 ppm MaxE, fit statistics for QM91k: 0.89 ppm MAE, 1.29 ppm RMSD, 25.4 ppm MaxE.

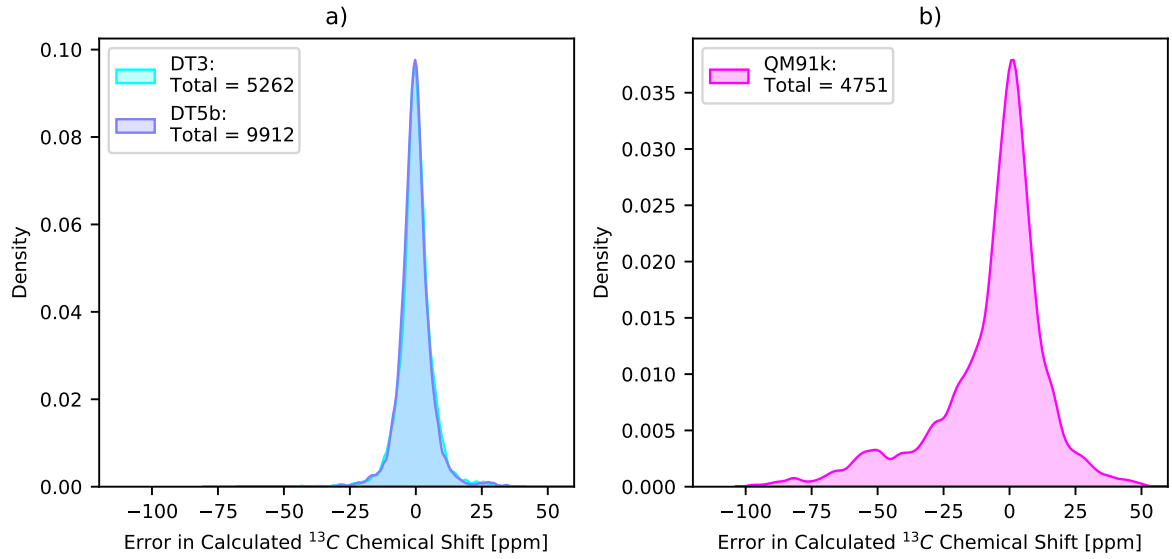

Figure 4.15: For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{13}C$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 4.41 ppm MAE, 6.71 RMSD, 90.82 MaxE, fit statistics for DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.13 ppm MaxE, fit statistics for QM91k: 14.5 ppm MAE, 21.9 ppm RMSD, 97.8 ppm MaxE.

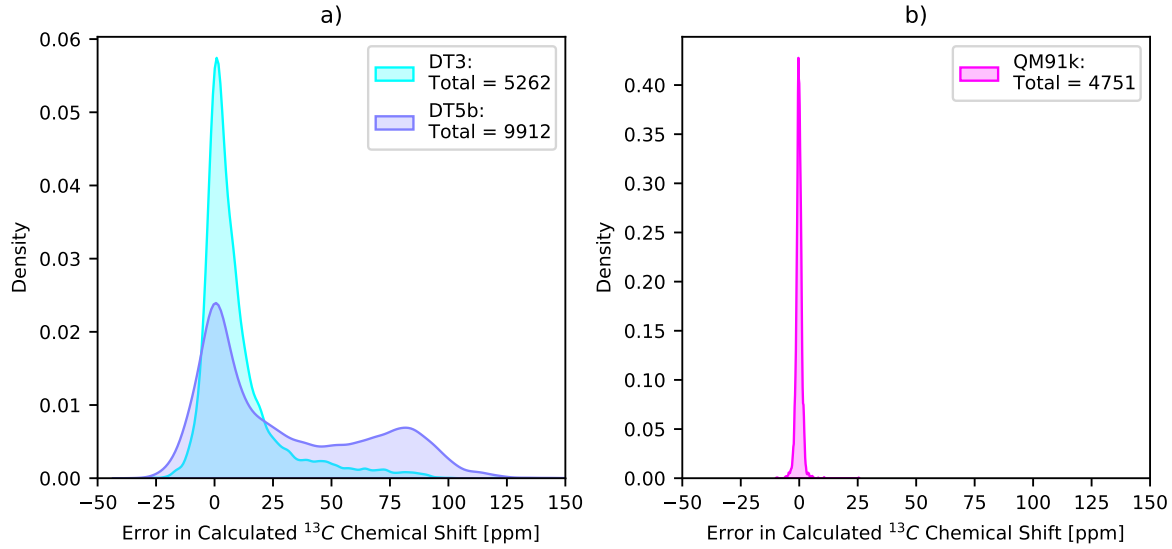

Figure 4.16: For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated $\delta^{13}C$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 11.4 ppm MAE, 19.6 RMSD, 120.4 MaxE, fit statistics for DT5b: 32.3 ppm MAE, 45.9 ppm RMSD, 150.5 ppm MaxE, fit statistics for QM91k: 0.89 ppm MAE, 1.29 ppm RMSD, 25.4 ppm MaxE.

4.2.4.3 Prediction accuracy and molecule size for $\delta^{13}C$ prediction

Similarly to the $\delta^{1}H$ predictions, the accuracy in predictions for $\delta^{13}C$ for each model depends on the size of the molecule (Figure 4.17). The DT45 trained model performs poorly on the smallest molecules (from QM91k) with an accuracy of 27.0 ppm MAE, 32.3 ppm RMSD, 60.1 ppm MaxE on the 30 environments from molecules with fewer than 10 atoms. Conversely the QM960k trained model performs poorly on the largest molecules (from DT5b) with an accuracy of 41.8 ppm MAE, 55.4 ppm RMSD, 150 ppm MaxE on the 4237 environments from molecules with greater than 80 atoms. As in the case of $\delta^{1}H$ prediction, both models fail to generalise successfully to molecules with significantly different size than those in their respective training datasets, however due to the nature of the DT45 training dataset, the DT45 trained model retains its prediction accuracy over a much larger range of molecule sizes.

It is also clear that the size of the molecules is not the only factor affecting model prediction accuracy, as there is still a difference in MAE for each of the two models across the three testing datasets, even when this difference is accounted for. This is visible in the mean absolute error of molecule subsets with mean number of atoms equal to 20 in Figure 4.17. There remains a clear bias in prediction accuracy towards molecules from the same source as the training data for a given model.

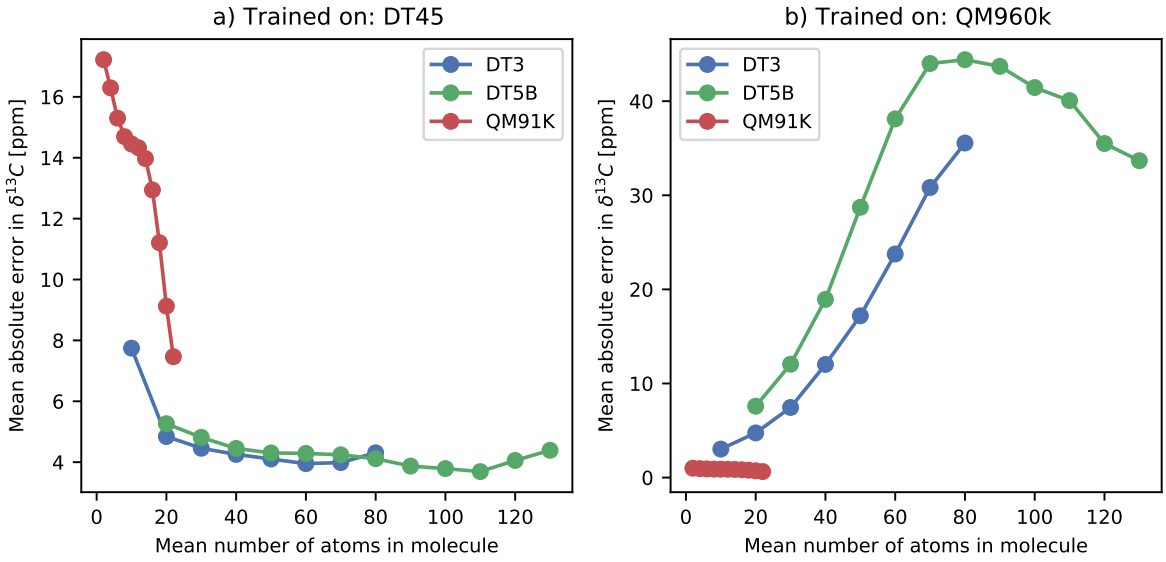

Figure 4.17: Accuracy in $\delta^{13}C$ prediction across the three testing datasets for subsets of molecules with different size, for the models trained using DT45 (a), and QM960k (b).

4.2.4.4 Performance relative to experiment for $\delta^{13}C$ prediction

The DT45 trained model obtains an accuracy of 3.76 ppm MAE, 5.25 ppm RMSD, 25.5 ppm MaxE on the experimentally measured $\delta^{13}C$ values from experimental testing dataset 1 (DTe1b, 46 molecules, 654 $\delta^{1}3C$ values, 2.4.6.1). This is significantly better than the generation 1 model trained using DT4: 4.76 ppm MAE, 6.82 ppm RMSD, MaxE 35.0 ppm. The QM960k trained model achieves a considerably worse accuracy: 7.15 ppm MAE, 11.92 ppm RMSD, 57.270 ppm MaxE. The accuracy of the underlying DFT method relative to the experimental values is 2.18 ppm MAE, 2.80 ppm RMSD, 14.9 ppm MaxE. A summary of these results is shown in Table 4.7. The DT45 trained model provides a much closer prediction accuracy to the underlying DFT method than the QM960k trained model, this is highlighted by the similarity in error distributions relative to experiment in Figure 4.18a and the relative dissimilarity between the error distributions in Figure 4.19a. The QM960k model predictions also contain several more large errors than the DT45 model predictions, visible in Figure 4.19b.

The pre-prediction variance provides a small but useful filter for the predictions for both models, in contrast to the results for the DFT testing datasets above. Removing environments with a variance of greater than 5000 for the QM960k trained model, and greater than 50,000 for the DT45 trained model removes just 5 (DT45) and 23 (QM960k) environments, but improves the mean absolute error from 3.76 ppm to 3.69 ppm for the DT45 trained model, and from 7.15 ppm to 6.78 ppm for the QM960k trained model. This is a small improvement, but demonstrates the potential utility of the pre-prediction variance in identifying unreliable predictions.

When splitting the test dataset into molecules with fewer or more than 40 atoms, the difference between the prediction accuracy of the two models displays the same pattern as in the $\delta^{1}H$ predictions. For the 208 environments from molecules with fewer than 40 atoms, the prediction accuracy of both models is similar (3.45/4.15 ppm MAE, 4.59/5.78 ppm RMSD, 17.8/19.2 ppm MaxE for models trained on DT45 and QM960k respectively). For the 341 environments from molecules with more than 40 atoms however, the accuracy of the two models diverge (9.41/3.52 ppm MAE, 14.7/4.89 ppm RMSD, 57.2/25.5 ppm MaxE for models trained on DT45 and QM960k respectively).

The QM960k trained model provides an advantage over the DT45 trained model in the prediction of environments from smaller molecules, however this difference is relatively small compared to the difference between prediction accuracy for larger molecules. This again suggests

that the DT45 trained model will perform better in the prediction of $\delta^{13}C$ in the practical application of these models, as the molecules of interest in such an application are likely to be larger than 30 atoms, and for smaller molecules the difference between the model prediction accuracy is minor.

Target	Training Dataset	MAE [ppm]	RMSD [ppm]	MaxE [ppm]
$\delta^{13}C$	DT45	3.759	5.245	25.448
$\delta^{13}C$	QM960k	7.151	11.922	57.270
$\delta^{13}C$	DFT	2.183	2.798	14.866

Table 4.12: Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the $\delta^{13}C$ experimental test set (DTe1b).

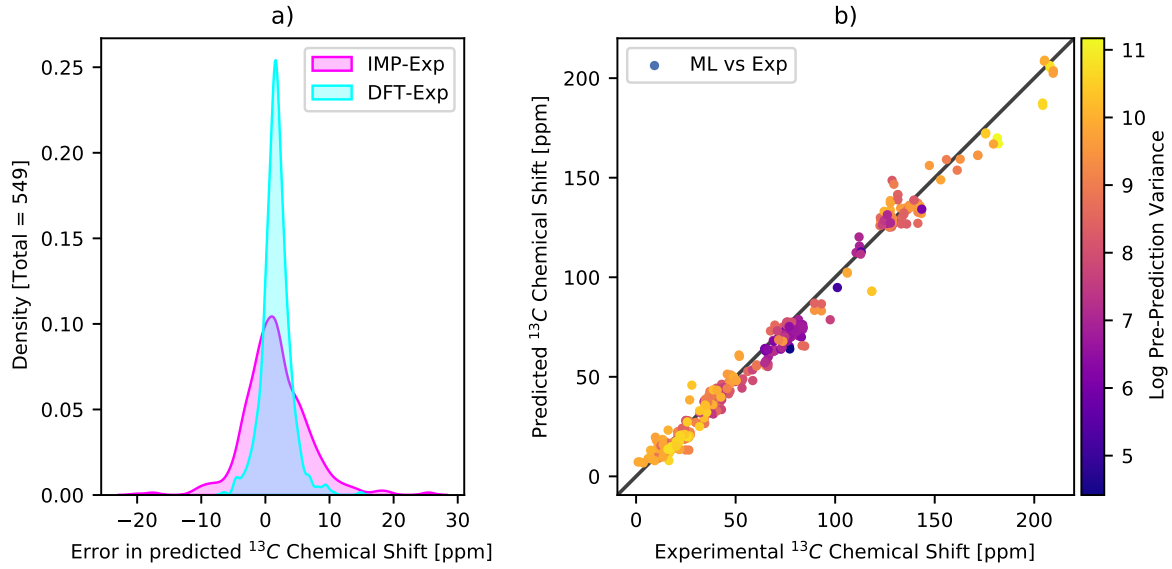

Figure 4.18: For the DT45 trained model predictions on the $\delta^{13}C$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $\delta^{13}C$, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 3.76 ppm MAE, 5.25 ppm RMSD, 25.45 ppm MaxE.

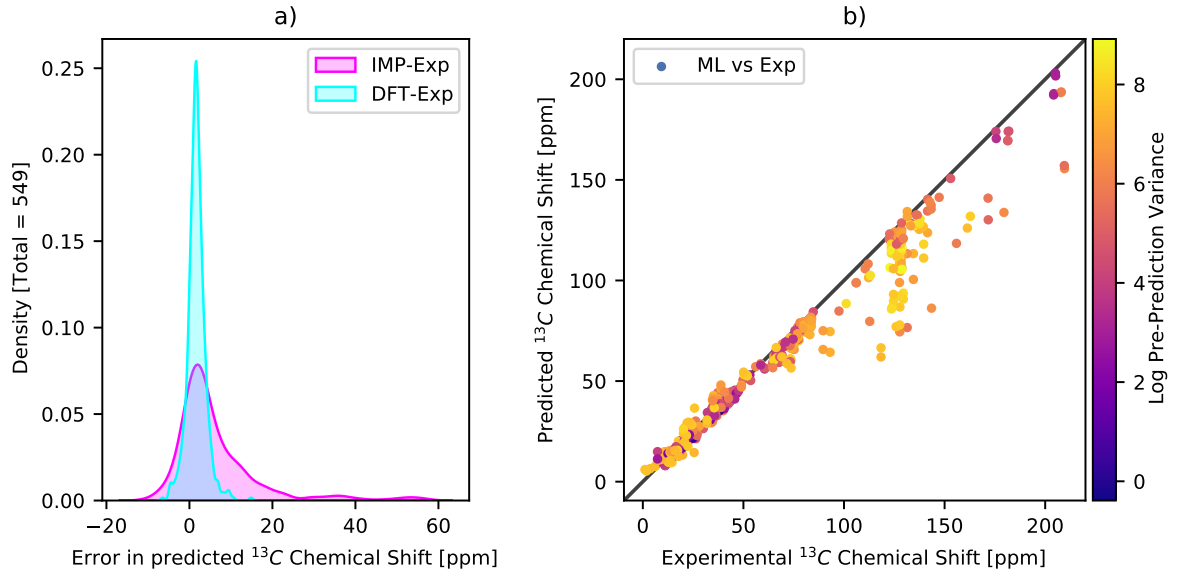

Figure 4.19: For the DT45 trained model predictions on the $\delta^{13}C$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $\delta^{13}C$, with pre-prediction variance highlighted (b). Fit statistics for DTe1b: 7.15 ppm MAE, 11.9 ppm RMSD, 57.3 ppm MaxE

4.2.5 $^{1}J_{CH}$ prediction

4.2.5.1 Performance relative to DFT for $^{1}J_{CH}$ prediction

The generation 2 model trained using DT45 (combination of dataset 4: Section 2.4.3 and dataset 5a: Section 2.4.4, 2,372 molecules, 121,436 $^{1}J_{CH}$ values) achieves an accuracy of 3.41/2.92 Hz MAE, 4.79/3.82 Hz RMSD, 51.6/23.4 Hz MaxE when tested on the DT3 (306 molecules, 5,608 $^{1}J_{CH}$ values, Section 2.4.2) and DT5b (400 molecules, 10,641 $^{1}J_{CH}$ values, Section 2.4.4) respectively (Figure 4.20a, and Figure 4.20a).

This accuracy is significantly worse for DT3 than the IMPRESSION generation 1 model trained using DT4 (1.29 Hz MAE, 2.16 Hz RMSD, 114.6 Hz MaxE), and significantly better for DT5b (7.29 Hz MAE, 8.02 Hz RMSD, 152.8 Hz MaxE). This could be a result of differences between molecules in DT4 and molecules in DT5a, which now dominate the training dataset, causing the model to prioritise prediction accuracy for molecules similar to those in DT5b, at the expense of accuracy on molecules in DT3.

When tested against the QM91k testing dataset (752 molecules, 6,284 $^{1}J_{CH}$ values, Section 2.4.5), the accuracy for the DT45 trained model is 7.01 Hz MAE, 8.69 Hz RMSD, 45.2 Hz MaxE

(Figure 4.20b and Figure 4.22b). The poor accuracy against the QM91k testing dataset follows the same pattern as observed for the chemical shift results above, the models trained using the larger structures from DT4 and DT5a lose a significant amount of accuracy in generalising to the smaller molecules in QM9.

The QM960k trained model achieves an accuracy of 0.54 Hz MAE, 0.77 Hz RMSD, 8.72 Hz MaxE when tested against the QM91k test set (Figure 4.21b and Figure 4.23b, and 7.51/26.7 Hz MAE, 10.5/31.6 Hz RMSD, 43.0/69.7 Hz MaxE when tested against the DT3 and DT5b testing datasets respectively 4.21a and Figure 4.23a. A summary of the accuracy for both models on the three testing sets is shown in Table 4.13.

Training Dataset	Testing Dataset	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE as % of Range
DT4 (Gen 1)	DT3	1.127	1.713	60.92	0.861
DT4 (Gen 1)	DT5B	1.825	3.198	73.556	1.098
DT45	DT3	3.408	4.789	51.622	2.981
DT45	DT5B	2.918	3.823	23.401	1.925
DT45	QM91K	7.011	8.693	45.183	5.070
QM960k	DT3	7.514	10.544	43.017	6.573
QM960k	DT5B	26.708	31.600	69.688	17.620
QM960k	QM91K	0.541	0.769	8.721	0.391

Table 4.13: Accuracy in $^1J_{CH}$ prediction across the three testing datasets 3 (DT3), 5b (DT5b) and QM91k, for the DT45 and QM960k trained models, as well as the generation 1, KRR model.

The accuracy of the QM960k trained model is also worse than the reported accuracy for the LightGBM based model [67] (1.82 Hz RMSD against a larger subset of QM9 molecules), however even if the improvement in accuracy for the LightGBM model also leads to an equivalent improvement in prediction accuracy on datasets such as DT3 and especially DT5b, the DT45 trained model is likely to provide significantly more accurate predictions for larger molecules.

4.2.5.2 Uncertainty Estimation for $^1J_{CH}$ prediction

The pre-prediction variance for the DT45 trained model follows a similar pattern as seen above for δ^1H prediction, as the pre-prediction variance appears to correlate more with the value of the NMR parameter than with the error in prediction (Figure 4.20). This is unsurprising as the environments for the δ^1H prediction are also a part of the $^1J_{CH}$ prediction, and so the relative dissimilarity of environments relative to both parameters will share some features.

The pre-prediction variance for the QM960k trained model is again unhelpful in the identification of large errors in the QM91k testing dataset, largely due to the lack of any large errors (Lack of outlying values in Figure 4.21b). In the prediction of DT3 and DT5b the largest outlying values above 225 Hz (DFT value) are identified, (brighter yellow points in Figure 4.21a). In this case a variance filter of 500 for the QM960k trained model improves the MAE relative to DT5b from 26.7 Hz to 26.3 Hz, removing 200 (less than 2% of the dataset). As can be seen by the majority of the points in 4.21a and Table 4.17 however, a very large proportion of the dataset needs to be discounted for the accuracy to begin to match that of the DT45 trained model. This indicates that the QM960k trained model will likely not be useful in the prediction of $^1J_{CH}$ values for larger molecules, even with a pre-prediction variance filter. The effect of variance cutoffs on the accuracy of both models against both testing datasets are shown in Tables 4.14, 4.15, 4.16, and 4.17.

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
5	4673	2.813	3.916	28.404	8.498
10	3810	2.903	3.935	28.404	10.112
50	1129	3.204	4.395	29.639	13.966
100	362	3.303	4.567	32.337	15.409
500	21	3.395	4.769	51.622	17.064
1000	4	3.404	4.782	51.622	17.152
5000	0	3.408	4.789	51.622	17.196

Table 4.14: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 3. Total $^1J_{CH}$ environments in DT3: 5,608

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
1	10385	7.753	8.549	15.433	11.174
5	8274	7.567	8.407	21.082	14.255
10	5531	6.964	7.790	24.509	15.531
50	540	6.781	7.602	29.440	18.295
100	303	6.814	7.646	29.440	18.664
500	33	6.873	7.729	30.131	19.486
1000	0	6.886	7.746	30.131	19.589

Table 4.15: For the model trained using DT45. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 5b. Total $^1J_{CH}$ environments in DT5b: 10,641

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
5	5480	5.900	6.380	13.622	6.650
10	5262	5.501	6.068	17.214	8.288
50	3977	4.900	6.002	30.732	14.000
100	2621	6.083	8.457	40.268	29.188
500	9	7.522	10.552	43.017	35.059
1000	0	7.514	10.544	43.017	35.059

Table 4.16: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 3. Total $^1J_{CH}$ environments in DT3: 5,608

Max Variance	No. Envs. Removed	MAE [Hz]	RMSD [Hz]	MaxE [Hz]	MAE of 100 largest errors[Hz]
50	9808	6.136	8.606	37.112	19.046
100	8378	13.299	18.328	53.549	42.502
500	200	26.320	31.178	69.639	62.754
1000	0	26.708	31.600	69.688	63.448

Table 4.17: For the model trained using QM960k. Effect of difference maximum variance cutoffs on accuracy metrics for IMPRESSION $^1J_{CH}$ predictions against DFT calculations for dataset 5b. Total $^1J_{CH}$ environments in DT5b: 10,641

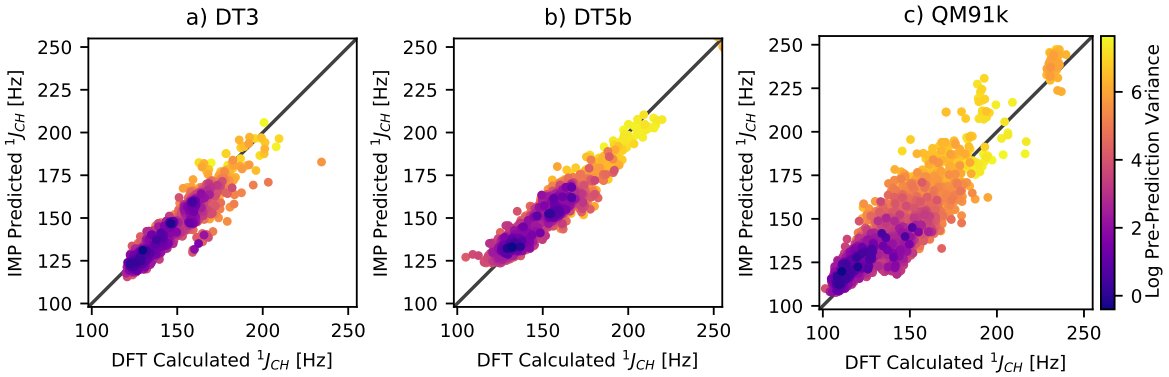

Figure 4.20: For the model trained using DT45: IMPRESSION predicted and DFT calculated $^1J_{CH}$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, fit statistics for DT5b: 2.92 Hz MAE, 3.82 Hz RMSD, 23.4 Hz MaxE, fit statistics for QM91k: 7.01 Hz MAE, 8.69 Hz RMSD, 45.2 Hz MaxE

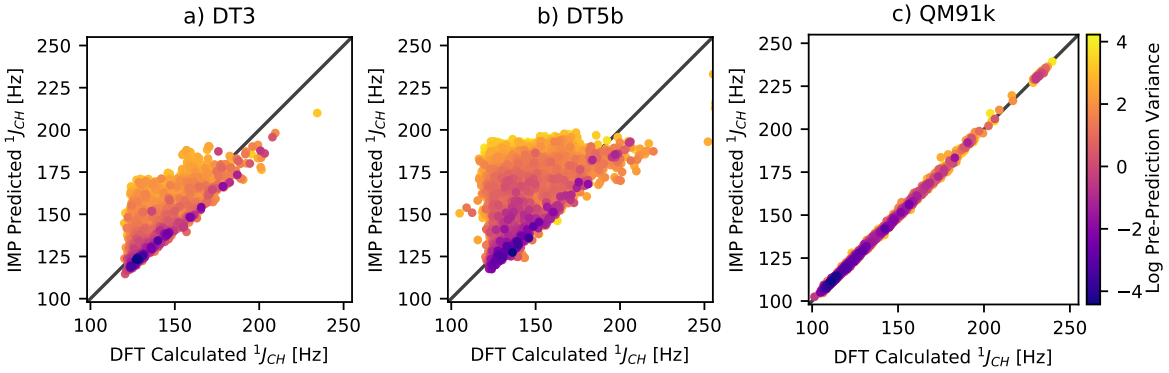

Figure 4.21: For the model trained using QM960k: IMPRESSION predicted and DFT calculated $^1J_{CH}$, with pre-prediction variance highlighted, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 7.51 Hz MAE, 10.5 Hz RMSD, 43.0 Hz MaxE, fit statistics for DT5b: 26.7 Hz MAE, 31.6 Hz RMSD, 69.7 Hz MaxE, fit statistics for QM91k: 0.54 Hz MAE, 0.77 Hz RMSD, 8.72 Hz MaxE

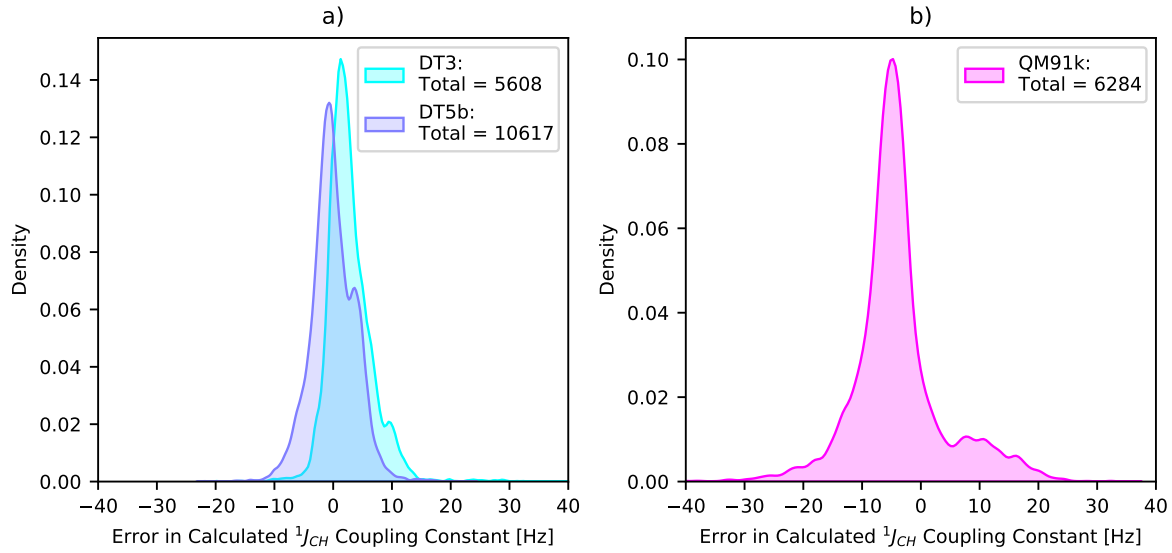

Figure 4.22: For the model trained using DT45: Error distribution between IMPRESSION predicted and DFT calculated $^1J_{CH}$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, fit statistics for DT5b: 2.92 Hz MAE, 3.82 Hz RMSD, 23.4 Hz MaxE, fit statistics for QM91k: 7.01 Hz MAE, 8.69 Hz RMSD, 45.2 Hz MaxE

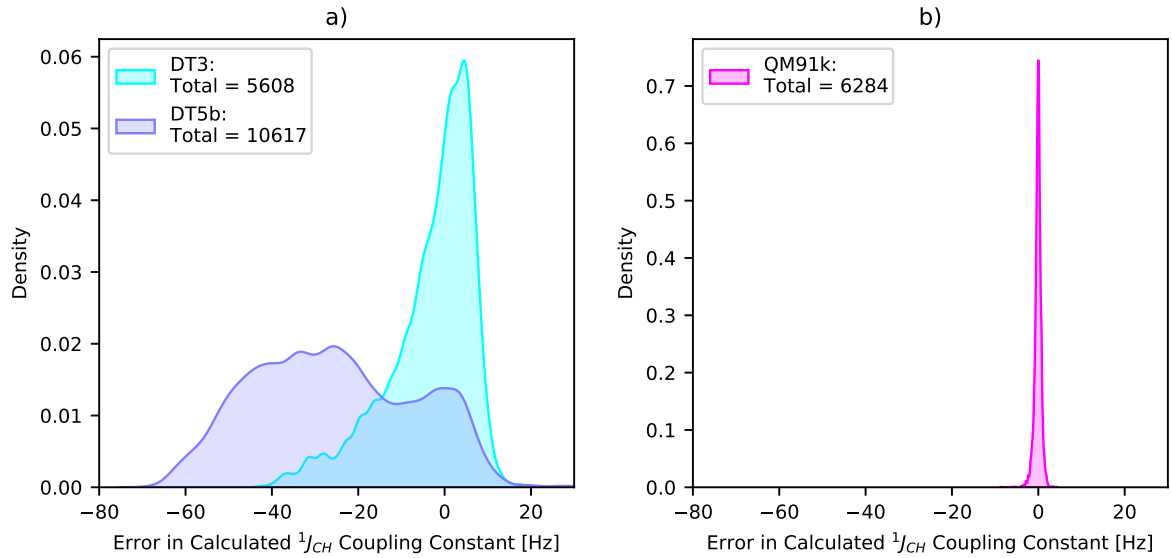

Figure 4.23: For the model trained using QM960k: Error distribution between IMPRESSION predicted and DFT calculated $^1J_{CH}$, for the DT3 (a), DT5b (b) and the QM91k (c) testing datasets. Fit statistics for DT3: 7.51 Hz MAE, 10.5 Hz RMSD, 43.0 Hz MaxE, fit statistics for DT5b: 26.7 Hz MAE, 31.6 Hz RMSD, 69.7 Hz MaxE, fit statistics for QM91k: 0.54 Hz MAE, 0.77 Hz RMSD, 8.72 Hz MaxE

4.2.5.3 Prediction accuracy and molecule size for $^1J_{CH}$ prediction

Similarly to the chemical shift predictions, the accuracy in $^1J_{CH}$ predictions for each model depends on the size of the molecule (Figure 4.24), although the effect is different for dataset 3 in this case.

The DT45 trained model performs poorly on the smallest molecules (from QM91k) with an accuracy of 13.0 Hz MAE, 18.5 Hz RMSD, 40.5 Hz MaxE on the 11 environments from molecules with fewer than 10 atoms. Conversely the QM960k trained model performs poorly on the largest molecules (from DT5b) with an accuracy of 45.1 Hz MAE, 46.5 Hz RMSD, 75.6 Hz MaxE on the 5075 environments from molecules with greater than 80 atoms.

For molecules with fewer than 25 atoms in DT3 and QM91k, the prediction accuracy from the QM960k trained model increases for smaller molecules. The QM960k model accuracy peaks for molecules with between 20 and 40 atoms in both DT3 and QM91k, with a mean absolute error of 4.73 Hz for DT3 and 0.42 Hz for QM91k. The accuracy for the QM960k trained model for datasets 3 and QM91k for each molecule size subset is shown in Tables 4.18 and 4.19.

The prediction accuracy for the DT45 trained model appears to decrease for the smallest molecules and largest molecules in dataset 3, with the peak accuracy occurring for molecules with between 10 and 20 atoms. It is uncertain what the cause of this is. The increase in prediction error for larger molecules is not seen for the DT45 trained models for chemical shift prediction (Figures 4.15 and 4.8).

Min Size	Max Size	No. Envs.	MAE [Hz]	RMSD [Hz]	MaxE [Hz]
5	10	11	1.084	1.507	3.484
10	20	5203	0.572	0.808	8.721
20	40	1959	0.422	0.576	2.866

Table 4.18: QM960k trained model predictions on testing dataset QM91k, split by molecule size.

Min Size	Max Size	No. Envs.	MAE [Hz]	RMSD [Hz]	MaxE [Hz]
5	10	10	9.734	10.961	17.334
10	20	99	7.848	8.411	14.338
20	40	2279	4.727	5.625	24.404
40	80	3306	8.255	11.403	43.017
80	160	175	25.280	26.426	40.268

Table 4.19: QM960k trained model predictions on testing dataset 3, split by molecule size.

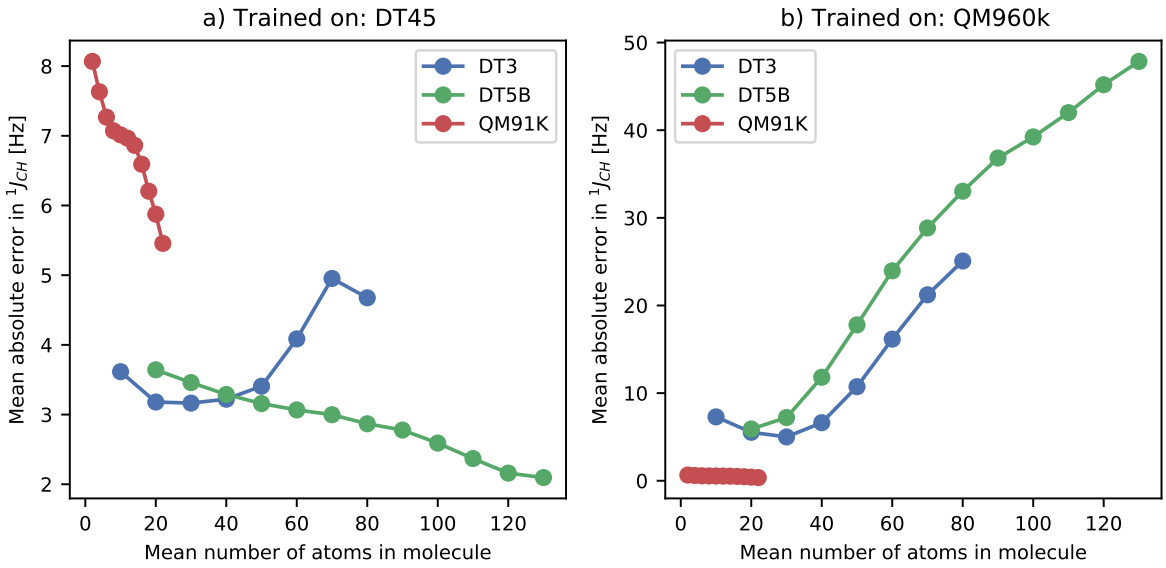

Figure 4.24: Accuracy in $^1J_{CH}$ prediction across the three testing datasets for subsets of molecules with different size, for the model trained using DT45 (a) and QM960k (b).

4.2.5.4 Experimental Validation for $^1J_{CH}$ prediction

The DT45 trained model achieves an accuracy of 6.69 Hz MAE, 10.59 Hz RMSD, 73.2 Hz MaxE against the 721 experimentally measured values in the $^1J_{CH}$ experimental dataset 3 (DTe3, 131 molecules, Section 2.4.6.3) (Figure 4.25). This compares well with the accuracy of the generation 1 model: 6.01 Hz MAE, 11.18 Hz RMSD, MaxE 54.3 Hz. The pre-prediction variance on the experimental predictions identifies several of the largest errors, and removing just 53 of the environments (around 7% of the dataset) reduces the MAE to 6.50 Hz, the RMSD to 10.22 and the MaxE to 55.8 Hz. The accuracy of the DFT calculations compared to the experimental values is 2.16 Hz MAE, 3.23 Hz RMSD, 20.1 Hz MaxE. The IMPRESSION predictions are relatively much poorer compared to the underlying DFT method in this case than in the chemical shift predictions.

The model trained on QM960k performs roughly as well as the DT45 trained model, with an accuracy of 6.45 Hz MAE, 10.13 Hz RMSD, and 60.8 Hz MaxE (Figure 4.26). Although neither of the models perform as well as hoped (similar accuracy to the underlying DFT model, as seen in chemical shift prediction), the fact that the QM960k trained model achieves a similar accuracy as the DT45 trained model is significant, though likely this is emphasised due to the small size of the molecules in the DTe3 dataset, with the majority being smaller than 20 atoms.

As with the DT45 trained model, a modest improvement to the QM960k prediction accuracy can be made by removing environments with high pre-prediction variance. Removing 232 environments (all with variance above 10) lowers the mean absolute error to 5.4 ppm. This is however a large proportion of the dataset (over 30%).

Target	Training Dataset	MAE [Hz]	RMSD [Hz]	MaxE [Hz]
$^1J_{CH}$	DT45	6.694	10.594	73.182
$^1J_{CH}$	QM960k	6.449	10.134	60.801
$^1J_{CH}$	DFT	2.158	3.226	20.050

Table 4.20: Accuracy of DFT calculations as well as predictions from the DT45 and QM960k trained models relative to the experimental values from the ^1JCH experimental test set DTe3.

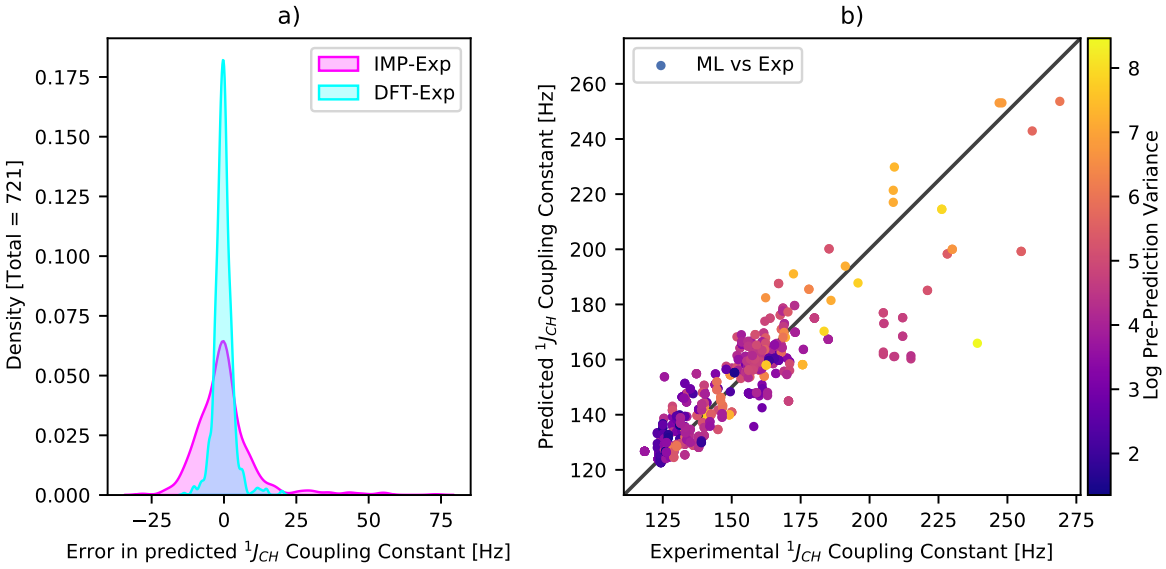

Figure 4.25: For the DT45 trained model predictions on the $^1J_{CH}$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $^1J_{CH}$, with pre-prediction variance highlighted (b). Fit statistics for DTe3: 6.69 Hz MAE, 10.6 Hz RMSD, 73.2 Hz MaxE

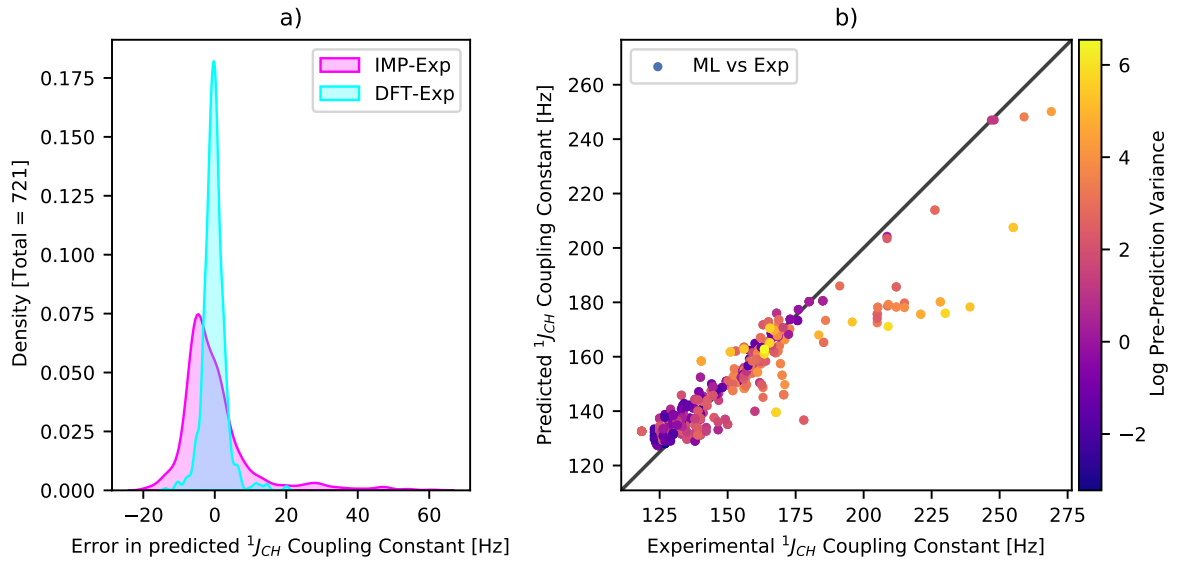

Figure 4.26: For the QM960k trained model predictions on the $^1J_{CH}$ experimental testing dataset. Error distributions between IMPRESSION and Experiment and between DFT and Experiment (a). IMPRESSION predicted against experimentally measured $^1J_{CH}$, with pre-prediction variance highlighted (b). Fit statistics for DTe3: 6.50 Hz MAE, 10.1 Hz RMSD, 60.8 Hz MaxE.

4.2.6 Further Scalar Coupling Prediction

The generation two models are trained to predict further scalar coupling constants: $^1J_{CC}$, $^2J_{CH}$, $^2J_{CC}$, $^2J_{HH}$, $^3J_{CH}$, $^3J_{CC}$, $^3J_{HH}$. These parameters were not investigated as a part of the generation 1 IMPRESSION prediction models, and the experimental validation of these parameters is beyond the scope of this thesis, however some analysis can be made of the quality of predictions for these NMR parameters relative to the underlying DFT method and relative to the performance on other parameters.

The accuracy of each parameter relative to the range of DFT calculated values across the three testing datasets (3, 5, QM91k) for models trained on DT45 and QM960k are shown in Figure 4.27. In general the parameters follow the same pattern as those discussed above (^1H and ^{13}C chemical shifts, and $^1J_{CH}$ coupling constants) , where the predictions for the DT45 model are reasonably accurate for the DT3 and DT5b testing datasets (MAE 3% or less of the range of DFT calculated values), but less accurate for the QM91k testing dataset (MAE greater than 3% in most cases, greater than 5% in 4 out of 7 parameters). Conversely the QM960k trained model predictions are less accurate for the DT3 and DT5b testing datasets (MAE greater than 5% of the range of DFT calculated values) and more accurate on the QM91k dataset (MAE less than 2% for

NMR Parameter	Reported RMSD [Hz] against QM9 [67]	QM960k trained model against QM91k RMSD [Hz]
$^2J_{CH}$	0.82	0.20
$^2J_{HH}$	0.48	0.18
$^3J_{CH}$	1.07	0.40
$^3J_{HH}$	0.67	0.17

Table 4.21: Accuracy comparison between recent published work and the QM960k trained model

all parameters).

The accuracy of the QM960k trained model against the QM91k testing dataset is better than the reported accuracy from the most recent work on prediction of several coupling constants: $^2J_{CH}$, $^2J_{HH}$, $^3J_{CH}$, $^3J_{HH}$ (Table 4.21) [67]. It is important to note that the reported accuracy in this case was for testing against a much larger subset of molecules from QM9, however both that dataset and the one used in this work were selected at random from molecules in the QM9 dataset. It is reasonable to suggest that the accuracy of the QM960k trained model on the DT3 and DT5b testing datasets is as good or better than the accuracy of the reported model from Shibata et al.

There are currently no reported machine learning models which predict the remaining $^{13}C-^{13}C$ coupling constants predicted by the generation 2 models. The percentage accuracy for these coupling constants ($^1J_{CC}$, $^2J_{CC}$, $^3J_{CC}$) is however similar to those of the other scalar coupling constants, and so the prediction accuracy on these parameters is also representative of the accuracy in recently published work if they were adapted for the prediction of these coupling constants.

The DT45 trained model therefore represents the most accurate predictions for the coupling constants: $^1J_{CC}$, $^2J_{CH}$, $^2J_{CC}$, $^2J_{HH}$, $^3J_{CH}$, $^3J_{CC}$, $^3J_{HH}$ on molecules similar to those in datasets 3 and 5b. Considering datasets 3 and 5b were chosen to be representative (in terms of size, structural diversity and constituent atoms) of organic molecules frequently targeted by NMR studies, the predictions from this model are likely to prove useful in practical applications, where predictions from QM9 based models are not.

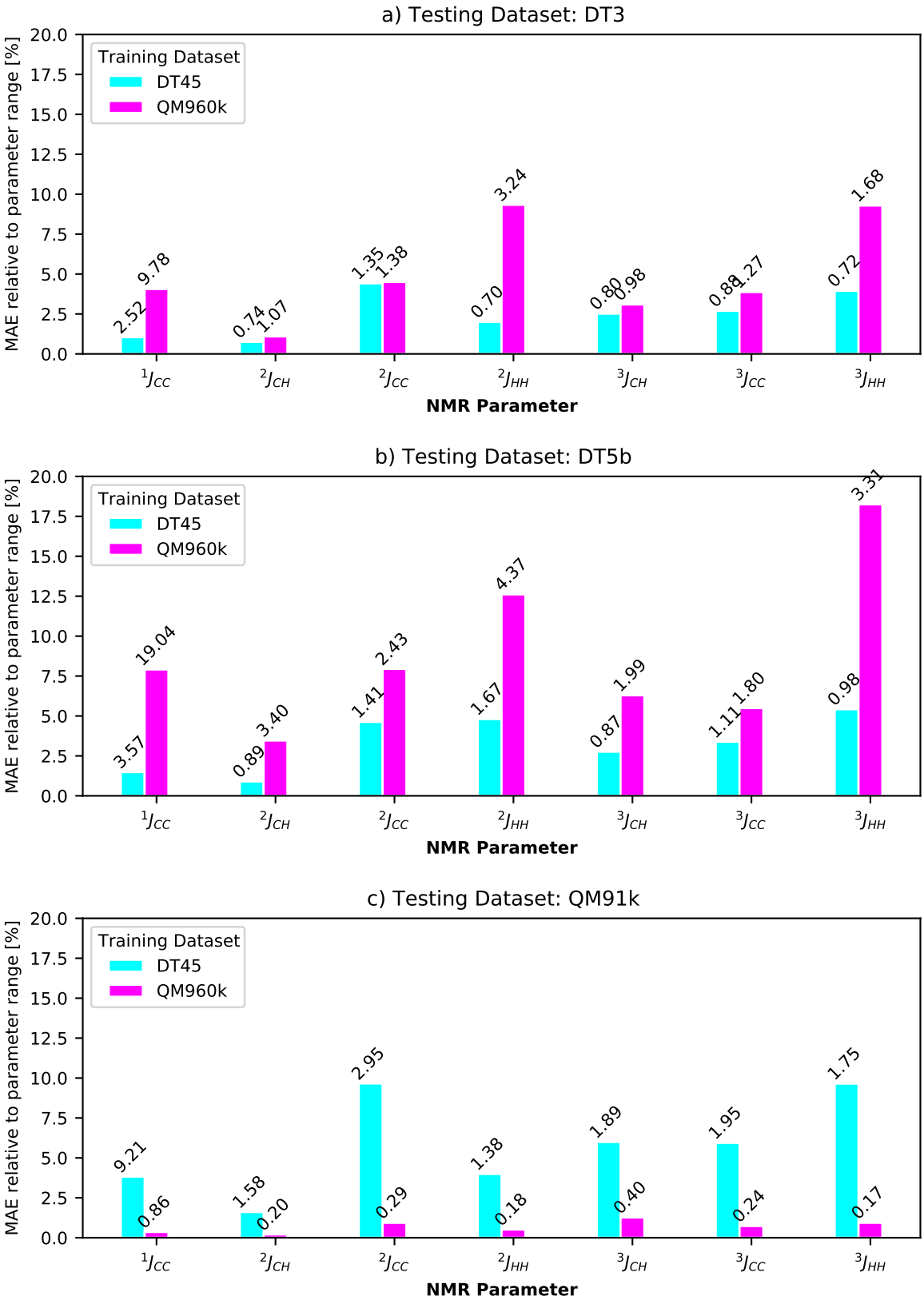

Figure 4.27: Comparison in model accuracy for testing datasets 3(a), 5b(b), QM91k(c). Bar height represents the mean absolute error as a percentage of the full range of values for that NMR parameter, each bar is annotated with the raw MAE values.

4.3 Comparison to NMRShiftDB

As discussed in Chapter 1, NMRShiftDB provides an open source NMR prediction tool [83]. The performance of this prediction tool is assessed relative to the two main models produced as a part of this project (IMPRESSION generation 1 and IMPRESSION generation 2) on a set of randomly selected compounds from DT5b (dataset 5b, discussed in section 2.4.4). The full results are shown below in tables 4.22 and 4.23.

The results for $\delta^{13}C$ prediction show no clear indication of whether the NMRShiftDB model provides better predictions than either of the two models produced, with the NMRShiftDB model outperforming the two IMPRESSION models in roughly half of the molecules for which predictions were available. The inability for the NMRShiftDB model to produce predictions for almost half of the submitted molecules however provides an indication that the IMPRESSION models drastically increase the range of molecules for which a similar accuracy can be achieved.

For $\delta^{1}H$ prediction there is a far clearer picture, with both IMPRESSION models providing significantly more accurate predictions for every molecule tested, in most cases with a mean absolute error of better than half that obtained from the NMRShiftDB predictions. As noted in 1, the NMRShiftDB model is designed to predict experimental chemical shifts and so there is some advantage provided to the IMPRESSION models here (all of these errors are relative to the DFT computed values), however even adjusting for a possible increase in the NMRShiftDB accuracy (the error between the chosen DFT method and experimental values can be estimated at around 0.2-0.3ppm) the IMPRESSION models are significantly more accurate, and provide accurate predictions for a much wider range of molecules than the NMRShfitDB model.

Molecule ID	Result	MAE NMR-ShiftDB [ppm]	MAE IMPGen1 [ppm]	MAE IMPGen2 [ppm]
CHEMBL1075841	Site Error	N/A	4.07	3.34
CHEMBL1084953	Success	2.85	2.17	3.93
CHEMBL1086530	Invalid atom(s)	N/A	13.00	4.08
CHEMBL1094672	Site Error	N/A	4.68	4.02
CHEMBL1096781	Success	4.04	7.08	3.52
CHEMBL1213982	Success	3.71	2.53	4.68
CHEMBL174668	Site Error	N/A	3.44	4.13
CHEMBL4116108	Success	1.88	4.54	6.23
CHEMBL4116148	Success	2.49	3.50	4.18
CHEMBL437851	Site Error	N/A	9.87	2.53
CHEMBL501943	Site Error	N/A	2.58	3.52
CHEMBL507540	Site Error	N/A	2.98	3.36
CHEMBL538928	Success	3.27	11.03	6.96
CHEMBL573427	Site Error	N/A	5.59	2.86
CHEMBL574221	Invalid Atom(s)	N/A	18.16	6.03
CHEMBL579584	Success	2.73	4.30	5.32
CHEMBL595793	Success	5.36	5.08	6.29
CHEMBL608847	Success	3.01	3.83	4.04
CHEMBL6225	Success	5.46	5.32	8.32
CHEMBL6889	Success	9.86	3.09	2.92

Table 4.22: Comparison between NMRShiftDB, IMPRESSION generation 1, and Impression generation 2 for $\delta^{13}C$ chemical shift. MAE = Mean Absolute Error.

Molecule ID	Result	MAE NMR-ShiftDB [ppm]	MAE IMPGen1 [ppm]	MAE IMPGen2 [ppm]
CHEMBL1075841	Site Error	N/A	0.23	0.23
CHEMBL1084953	Success	0.22	0.13	0.16
CHEMBL1086530	Invalid atom(s)	N/A	0.34	0.22
CHEMBL1094672	Site Error	N/A	0.39	0.29
CHEMBL1096781	Success	0.75	0.33	0.27
CHEMBL1213982	Success	0.53	0.21	0.31
CHEMBL174668	Site Error	N/A	0.33	0.25
CHEMBL4116108	Success	1.76	0.39	0.38
CHEMBL4116148	Success	0.79	0.20	0.28
CHEMBL437851	Site Error	N/A	0.32	0.20
CHEMBL501943	Site Error	N/A	0.26	0.31
CHEMBL507540	Site Error	N/A	0.35	0.23
CHEMBL538928	Success	0.33	0.26	0.23
CHEMBL573427	Site Error	N/A	0.41	0.24
CHEMBL574221	Invalid Atom(s)	N/A	0.59	0.58
CHEMBL579584	Success	0.87	0.31	0.24
CHEMBL595793	Success	1.01	0.20	0.51
CHEMBL608847	Success	1.05	0.22	0.23
CHEMBL6225	Success	0.71	0.50	0.33
CHEMBL6889	Success	0.77	0.20	0.22

Table 4.23: Comparison between NMRShiftDB, IMPRESSION generation 1, and Impression generation 2 for $\delta^1 H$ prediction. MAE = Mean Absolute Error.

4.4 IMPRESSION Generation 1 vs IMPRESSION Generation 2

The primary purpose of the further development of machine learning models is to make improvements upon the first generation of machine learning models. For the purposes of this section the IMPRESSION generation 1 model trained on DT4 (Dataset 4, derived from the CSD, discussed in section 2.4.3) and the IMPRESSION generation 2 model trained on DT45 (the combination of both datasets 4 and 5) are treated as the final output of the two model development processes.

In the development of the generation 2 models (based on graph transformer network architecture), the aims were to improve the prediction accuracy for the molecules in DT5b (dataset 5b, derived from ChEMBL structures, discussed in section 2.4.4) in particular. It was also hoped that an improved architecture and a larger training set may also yield improved predictions on DT3 (dataset 3, derived from CSD structures, discussed in section 2.4.2), however this is clearly not the case.

Firstly, looking at δ^1H prediction, the relative error distributions of the two models for DT3 and DT5b are shown in Figure 4.29, the four error distributions shown here are all relatively similar, indicating that the accuracy of both models is comparable across both datasets for this parameter. The mean absolute errors in predictions are similar, for DT3 the MAE is 0.24 ppm for generation 1 and 0.22 for generation 2, for DT5 the MAEs are 0.34 ppm and 0.36 ppm respectively. There is therefore little advantage in the generation 2 model in terms of accuracy for δ^1H prediction, however the other advantages of the generation 2 model, in terms of further

Target	Generation	Testing dataset	MAE	RMSD	MaxE
δ^1H	1	DT3	0.24 ppm	0.39 ppm	4.27 ppm
δ^1H	2	DT3	0.22 ppm	0.36 ppm	8.01 ppm
δ^1H	1	DT5b	0.34 ppm	0.54 ppm	8.78 ppm
δ^1H	2	DT5b	0.27 ppm	0.36 ppm	5.96 ppm
$\delta^{13}C$	1	DT3	3.50 ppm	7.05 ppm	106.5 ppm
$\delta^{13}C$	2	DT3	4.41 ppm	6.71 ppm	90.1 ppm
$\delta^{13}C$	1	DT5b	6.34 ppm	17.1 ppm	271.7 ppm
$\delta^{13}C$	2	DT5b	4.31 ppm	6.31 ppm	64.1 ppm
$^1J_{CH}$	1	DT3	1.12 Hz	1.71 Hz	60.9 Hz
$^1J_{CH}$	2	DT3	3.41 Hz	4.79 Hz	51.6 Hz
$^1J_{CH}$	1	DT5b	1.83 Hz	3.20 Hz	19.3 Hz
$^1J_{CH}$	2	DT5b	2.92 Hz	3.80 Hz	23.4 Hz

Table 4.24: Caption

development and speed of prediction, are still applicable.

For $\delta^{13}C$ prediction, the generation 1 model shows a significant difference in performance when making predictions on the two datasets DT3 and DT5b, with MAEs of 3.5 ppm and 6.3 ppm respectively. The generation 2 model however produces similar accuracy against both datasets, with MAEs of 4.41 ppm and 4.31 ppm for DT3 and DT5b respectively. These differences in error distributions are clear in Figure 4.30, and indicate that whilst the generation 2 model has worse accuracy on DT3 than the first generation, there is clearly an improvement in the generalisation of the model. This is arguably one of the most important qualities in a machine learning system, and so this represents a significant improvement from the first to the second generation.

The results in the comparison of predictions for $^{1}J_{CH}$ prediction are different again, as can be clearly seen from the error distributions in Figure 4.31. The generation 2 model performs worse against both datasets, with MAEs between 2 and 3 times worse than the generation 1 model. This is a disappointing result from a model development perspective, however it is interesting that the new architecture and increased training dataset has yielded different results across the three main parameters investigated. The potential solution for improving $^{1}J_{CH}$ prediction in generation 2 lies clearly in the expansion of the training datasets, which is one of the main advantages of moving from a kernel ridge regression architecture to a neural network style model. The practical limits of the training dataset size of generation 1 have been nearly realised in the model discussed here, however the training dataset for the generation 2 model can be vastly increased, and relative to recent work in neural network prediction models, the existing dataset here is very small.

In terms of the pre-prediction variance performance, both generations suffer from a similar issue in most cases, namely that in most circumstances, the correlation between pre-prediction variance and prediction error breaks down at low errors. In other words the pre-prediction variance is a good indicator of very poor predictions, however identifying predictions with even 2 or 3 times worse prediction accuracy does not seem possible with the pre-prediction variance as it is implemented here. This effect can be clearly seen in Figure 4.28, where the flatness of the curve for low variance percentiles shows the lack of correlation between this value and the mean absolute error. Furthermore these results show that in several cases the generation 2 model shows little to no correlation between the two values, and in fact for the prediction of $\delta^{13}C$ for DT3 it appears to show a small negative correlation (Figure 4.28c).

In the cases where there is at least a correlation for high variance values, there is still a significant utility in using the pre-prediction variance, as it does effectively exclude the largest errors, especially in the generation 1 model. A hypothetical situation could be imagined where, for the generation 1 model, values are labelled with the pre-prediction variance percentile (to make this comparable across different parameters), and this value used to weight comparisons such as those discussed in Chapter 5. It would be highly important in these cases to determine the percentiles based on an independent test dataset, and to assign new values into the existing percentiles, otherwise the labels would be tightly dependent on the quality of predictions for the given molecule being analysed.

The results here are promising in terms of the potential utility of the generation 2 model, as they already show a significant improvement in $\delta^{13}C$ prediction accuracy, and successfully improve the generalisation of $\delta^{1}H$ prediction. However the poor $^{1}J_{CH}$ prediction accuracy indicates this is not a simple improvement in prediction model across all metrics. It is likely that through expanding the training dataset for the generation 2 model that the accuracy can be significantly increased, beyond that of the generation 1 model.

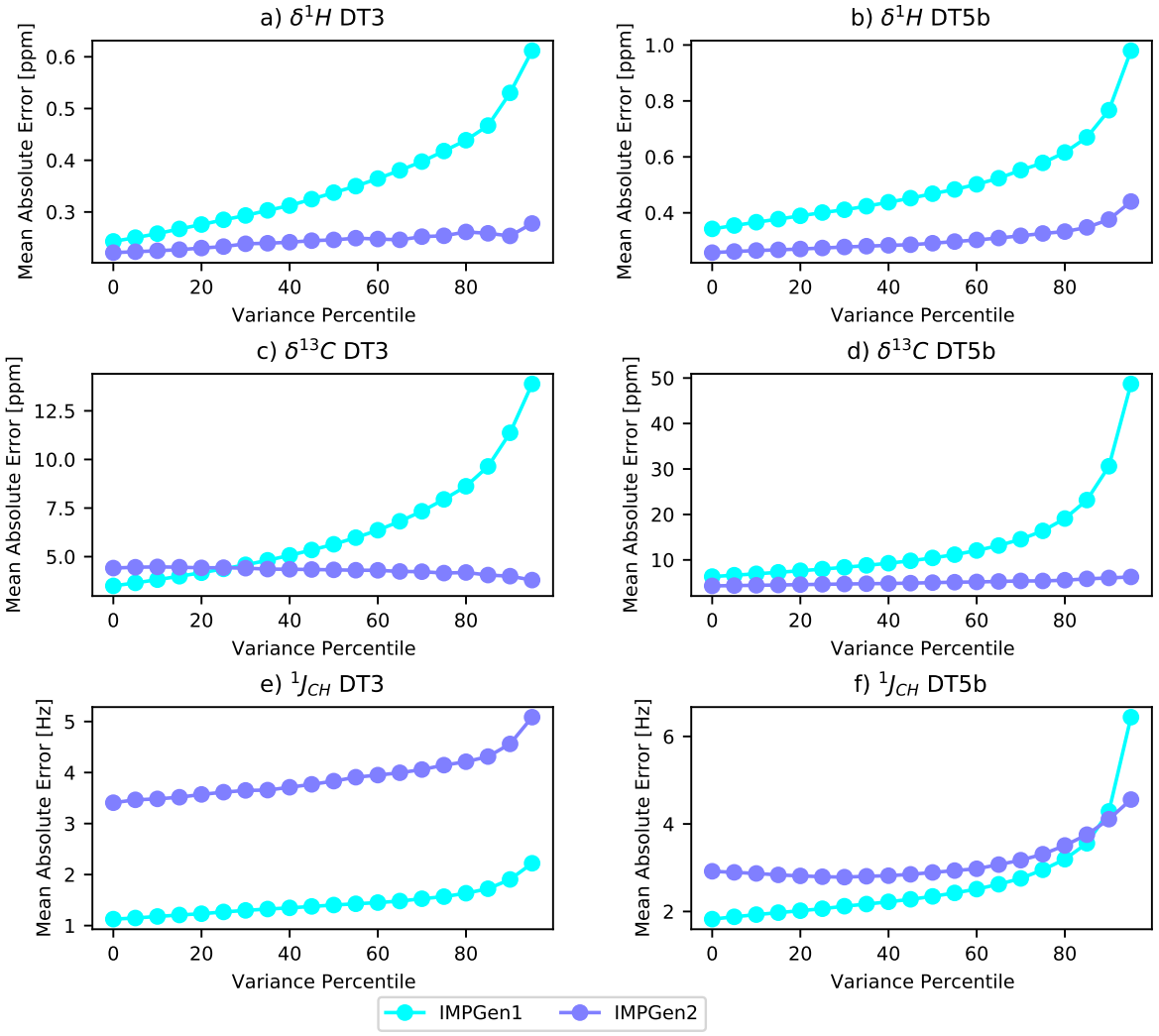

Figure 4.28: Comparison between IMPRESSION generation 1 trained using DT4 and IMPRESSION generation 2 trained using DT45, in terms of the correlation between the mean absolute error and the pre-prediction variance. For three NMR parameters against both testing datasets: δ^1H for DT3 (a), δ^1H for DT5b (b), $\delta^{13}C$ for DT3 (c), $\delta^{13}C$ for DT5b (d), $^1J_{CH}$ for DT3 (e), $^1J_{CH}$ for DT5b (f).

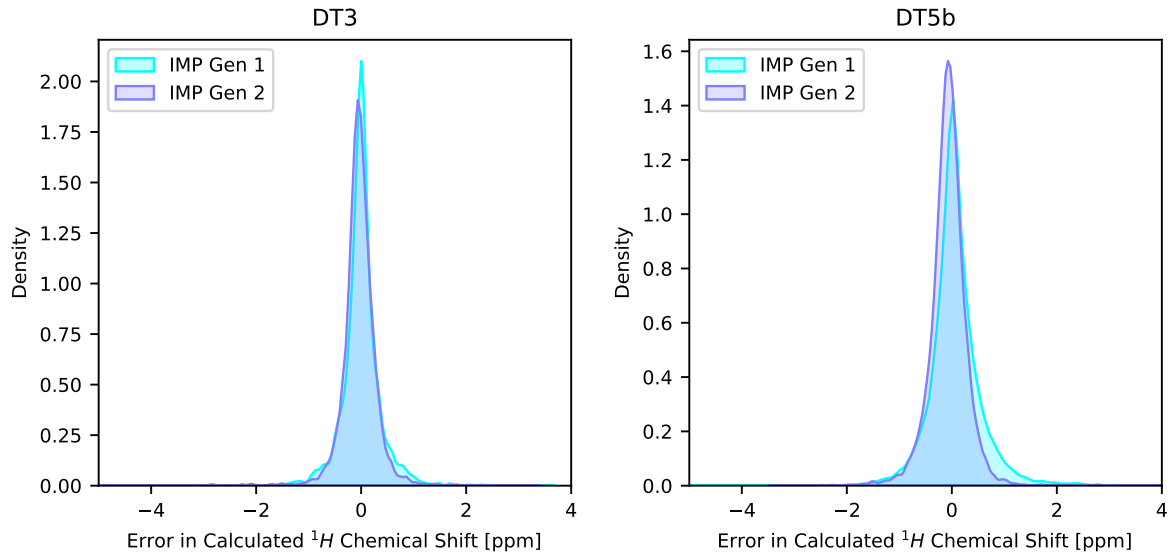

Figure 4.29: Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for δ^1H prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 0.24 ppm MAE, 0.39 ppm RMSD, 4.27 ppm MaxE, DT5b: 0.34 ppm MAE, 0.54 ppm RMSD, 8.78 ppm MaxE. Fit statistics for Generation 2, DT3: 0.22 ppm MAE, 0.36 ppm RMSD, 8.01 ppm MaxE, DT5b: 0.27 ppm MAE, 0.36 ppm RMSD, 5.96 ppm MaxE.

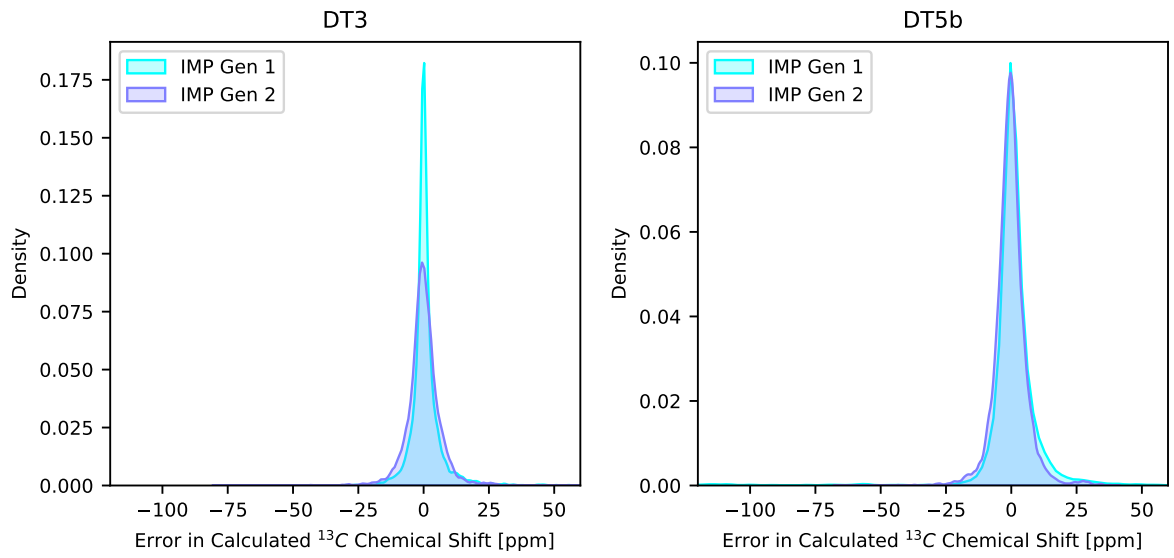

Figure 4.30: Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for $\delta^{13}C$ prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 3.50 ppm MAE, 7.05 ppm RMSD, 106.5 ppm MaxE, DT5b: 6.34 ppm MAE, 17.1 ppm RMSD, 271.7 ppm MaxE. Fit statistics for Generation 2, DT3: 4.41 ppm MAE, 6.71 ppm RMSD, 90.12 ppm MaxE, DT5b: 4.31 ppm MAE, 6.31 ppm RMSD, 64.1 ppm MaxE.

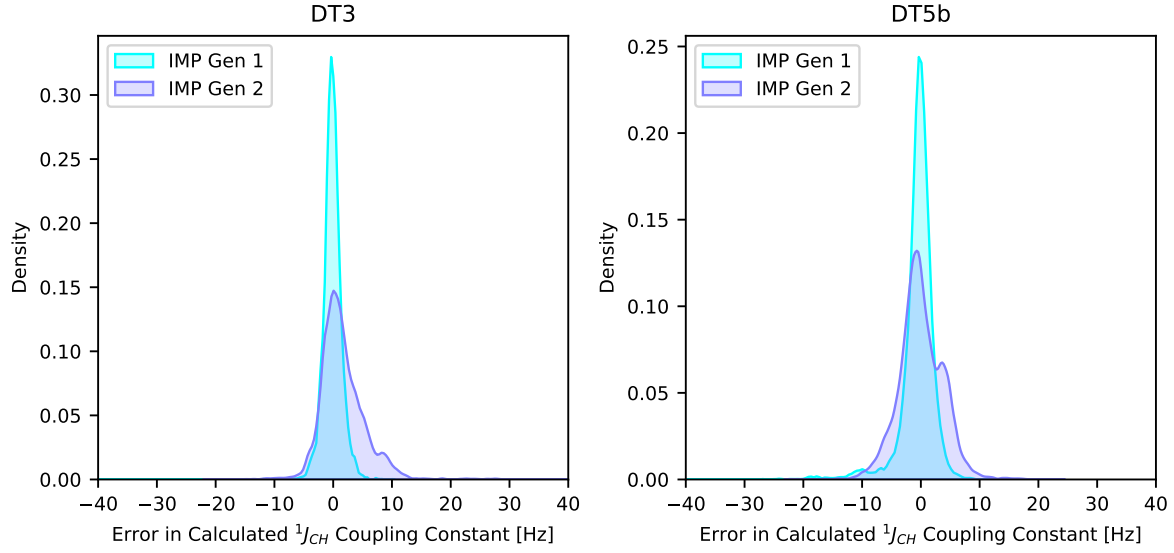

Figure 4.31: Comparison between IMPRESSION generation 1 model (trained using DT4) and IMPRESSION generation 2 model (trained using DT45) for $^{1}J_{CH}$ prediction. Tested against DT3 and DT5b. Fit statistics for Generation 1, DT3: 1.12 Hz MAE, 1.71 Hz RMSD, 60.9 Hz MaxE, DT5b: 1.83 Hz MAE, 3.20 Hz RMSD, 19.3 Hz MaxE. Fit statistics for Generation 2, DT3: 3.41 Hz MAE, 4.79 Hz RMSD, 51.6 Hz MaxE, DT5b: 2.92 Hz MAE, 3.8 Hz RMSD, 23.4 Hz MaxE.

4.5 QM9 models and overfitting

The results in this section for models trained using molecules taken from QM9 dataset [77] are intended to test a specific hypothesis, namely whether the incredibly impressive NMR parameter prediction accuracy reported in recent literature is a genuine result, and easily expandable to molecules of more genuine scientific interest, or a result of overfitting on a relatively simple dataset.

Overfitting is discussed in general terms in Section 2.1.1, and there is significant evidence in the results produced to suggest that the QM9 models reported suffer from overfitting. The QM960k trained generation 2 model analysed in this chapter produces accuracy similar to that of several recently published models [80][67][65] when tested against the QM91k dataset (1000 molecule subset of QM9, discussed in section 2.4.5). However, this accuracy does not generalise to the two other datasets (DT3 and DT5b, discussed in sections 2.4.2 and 2.4.4 respectively). The prediction accuracy is over 10x worse across the three main parameters investigate (δ^1H, $\delta^{13}C$, and ^1JCH). This combined with the fact that the same architecture and training procedure is capable of producing good quality predictions on datasets DT3 and DT5b when trained on appropriate molecules, strongly suggests that the QM9 trained models produced here and in the literature are examples of significant overfitting, and are of little use beyond extremely simple molecules. It should be noted here that the work into QM9 trained models both here and in other work, still serves to demonstrate the potentially highly significant improvements possible in chemical property prediction through the use of more complex machine learning architectures. It would not be possible to demonstrate the hypothetical possibility of δ^1H prediction accuracy better than 0.1 ppm without producing such models, and more realistic training datasets are unlikely to be produced (due to the high cost) unless their potential benefit is demonstrated. The analysis provided here merely suggests that the work needed to produce a genuinely useful model capable of that level of accuracy is potentially greater than suggested in other recent work.

4.6 Conclusion

Overall, the success of the generation 2 IMPRESSION models is largely in the expansion of accurate predictions to a greater variety of NMR parameters, whilst marginally improving the accuracy for most of parameters currently predicted by the generation 1 models. The GTN

architecture also has distinct advantages over the KRR method used in generation 1, making significantly faster predictions and removing the memory restrictions which limit the potential training dataset size for kernel ridge regression models.

The notable exception to the improvement in accuracy is the prediction of $^1J_{CH}$ coupling constants, where the IMPRESSION generation 1 model is significantly better (2-3x smaller MAE) than the generation 2 model, including on ChEMBL derived testing dataset DT5b. This clearly demonstrates some advantage in the generation 1 architecture and or molecular representation in predicting $^1J_{CH}$ coupling.

The pre-prediction variance results for generation 2 are disappointing, and the loss of this tool greatly diminishes the utility of the models in general application, some work is needed to either adjust the training of the model or calculation of this parameter so that it does provide some uncertainty estimation, or other ways of obtaining this information investigated.

The generation 2 models trained using DT45, as presented above, are likely the most accurate machine learning NMR prediction models available for the types of molecules found in DT5b, i.e. large (30+ atoms), drug like molecules. This is likely due to the fact that the equivalent model trained using QM960k outperformed several of the current-best prediction models in the literature on predicting parameters for QM9 molecules, and the DT45 model far outperformed the QM960k trained model in the prediction of DT5b molecules. Whilst a more accurate comparison would be preferable, that would require all of the published models to be recreated in order to make predictions on DT5b. This is beyond the scope of the work for this thesis and so is left as potential future work, though it is hoped that more relevant testing datasets will be adopted in the literature in future, making this work unnecessary.

CHAPTER 5

STRYCHNINE PREDICTION TASK

Three of the machine learning models presented in this thesis will be compared further in this section: the generation 1 kernel ridge regression model trained using dataset 4 (Chapter 3 describes the model architecture and performance, section 2.4.3 in Chapter 2 describes the training dataset), the generation 2 graph transformer network trained using datasets 4 and 5, as well as the generation 2 graph transformer model trained using the QM960k dataset (Chapter 4 describes the model architecture and performance, sections 2.4.3 and 2.4.4 in Chapter 2 describe the training datasets). The three models will be assessed in their performance of a prediction task similar to one in which these models may be used in practice.

The polycyclic alkaloid strychnine has a naturally occurring stereoisomer (structure 1 in Figure 5.1). A set of 12 other energetically viable diastereomers can be constructed (structures 2-13 in Figure 5.1), and a hypothetical situation imagined in which a strychnine sample had been obtained but the stereoisomeric form of the sample is unknown. The ^{13}C and ^{1}H chemical shifts, along with ^{13}C-^{1}H and ^{1}H-^{1}H coupling constants have been obtained from experimental NMR spectra. It is then the task of the prediction models developed to identify which diastereomer is in the sample, by making predictions of the NMR parameters for all 13 structures and finding those which match the experimentally obtained values the closest. It should be noted that this is a particularly difficult test of the models' performance, considering none of the models have been trained using multiple diastereomers of the same structure, and so the ability to distinguish

between such subtle variations in structure is hoped for but not necessarily expected.

Beuvich et al previously demonstrated the ability of DFT calculated $^{1}J_{CH}$ values to perform this task, identifying the correct diasteromer from 12 other diasteromers along with a second (less populated) conformer of the correct diasteromer [133]. The dominant conformer of the correct diastereomer is labelled structure 1, then the 12 other diasteromers labelled 2-13, the non-dominant conformer of the correct diastereomer is labeled structure 14. The 'dominant' conformer here refers to the fact that this conformer represents 97% of the population in solution, and so best represents the structure in the sample, whereas the less-favourable conformer represents 3%. This labelling system will be used for the analysis in this chapter.

In an ideal solution, the structure with the lowest mean absolute error relative to the experimental values would be assumed to be the correct structure. This can be quantified through the mean absolute error itself, but also through applying a 'Softmin' function to the mean absolute errors, e, across all structures i:

$$\text{Softmin}(e_i) = \frac{\exp(-e_i)}{\sum_{i=0}^{j} \exp(-e_i)} \tag{5.1}$$

The values returned by the softmin function sum to 1, and so the values can be interpreted in this case as a crude percentage probability for each structure. This approach emphasises the difference in how closely the parameters match for each structure, which is helpful, but can be potentially misleading. It is important in problems such as this to not over-rely on a binary classification of the correct structure, this will become clearer through examples in this analysis, but it is of more scientific utility to accept that the chosen prediction method does not provide a clear identification that any structure is correct, than it is to allow the model to incorrectly assign the wrong structure as the correct structure.

Further to this point, using models such as those produced in this work as a method of identifying the single correct structure of a compound is likely to produce incorrectly assigned structures more often than not, largely due to the current error in NMR prediction models being significantly greater than the difference in the value of parameters between very similar structures. The use of models in this way is detailed in other work [53], however the intention in this work is to produce models which act as a screening tool for structural assignment, as a complementary source of information to DFT calculations. As such the relative errors on different compounds have not been converted to a concrete recommendation of a given structure over another, and it is strongly recommended that models such as these should not be used in this

NMR Parameter	Gen1 DT4 Cutoff	Gen2 DT45 Cutoff	Gen2 QM960k Cutoff
δ^1H	0.1	1.0	1.0
$\delta^{13}C$	5.0	10,000	1,000
$^1J_{CH}$	1.0	20	500
$^2J_{CH}$	∞	200	50
$^2J_{HH}$	∞	1.0	20
$^3J_{CH}$	∞	20	20
$^3J_{HH}$	∞	50	10

Table 5.1: Variance cutoff values used for each model for each parameter in the strychnine prediction task.

way unless it could be demonstrated that the error in predicting the given NMR parameter was far lower than the difference in that parameter between the structures being analysed.

5.0.1 Uncertainty estimation

In both the generation 1 and generation 2 machine learning models, a 5-fold variance across drop-out models has been used as a measure of uncertainty in the core model prediction. The pre-prediction variance correlates with prediction error across different datasets in some cases however there still remain many environments which are poorly predicted, but which retain low pre-prediction variance values, this is discussed in detail in Section 3.1.4.

In the strychine prediction task, the pre-prediction variance provides a much more meaningful advantage than it does in the simple comparison between prediction error of different models. The discrimination between structures in problems such as these can be clouded by environments with high error, and so if any of these can be identified and removed from the analysis this presents an advantage. Limiting the number of environments involved in the comparison will potentially make it harder to discriminate between structures, and so a balance must be struck between removing unreliably predicted environments, whilst retaining enough environments to make meaningful distinctions between structures.

For the purpose of this task, variance cutoffs for the parameters predicted in each case were set in advance of the task being performed, so as to not bias the results. The maximum allowed variance for an environment for each parameter was set through analysis of the distributions in variance across testing datasets 3 and 5b, and attempting to balance the priorities mentioned above. The choice of variance cutoffs is intrinsically arbitrary to some extent. The chosen cutoffs are shown in Table 5.1.

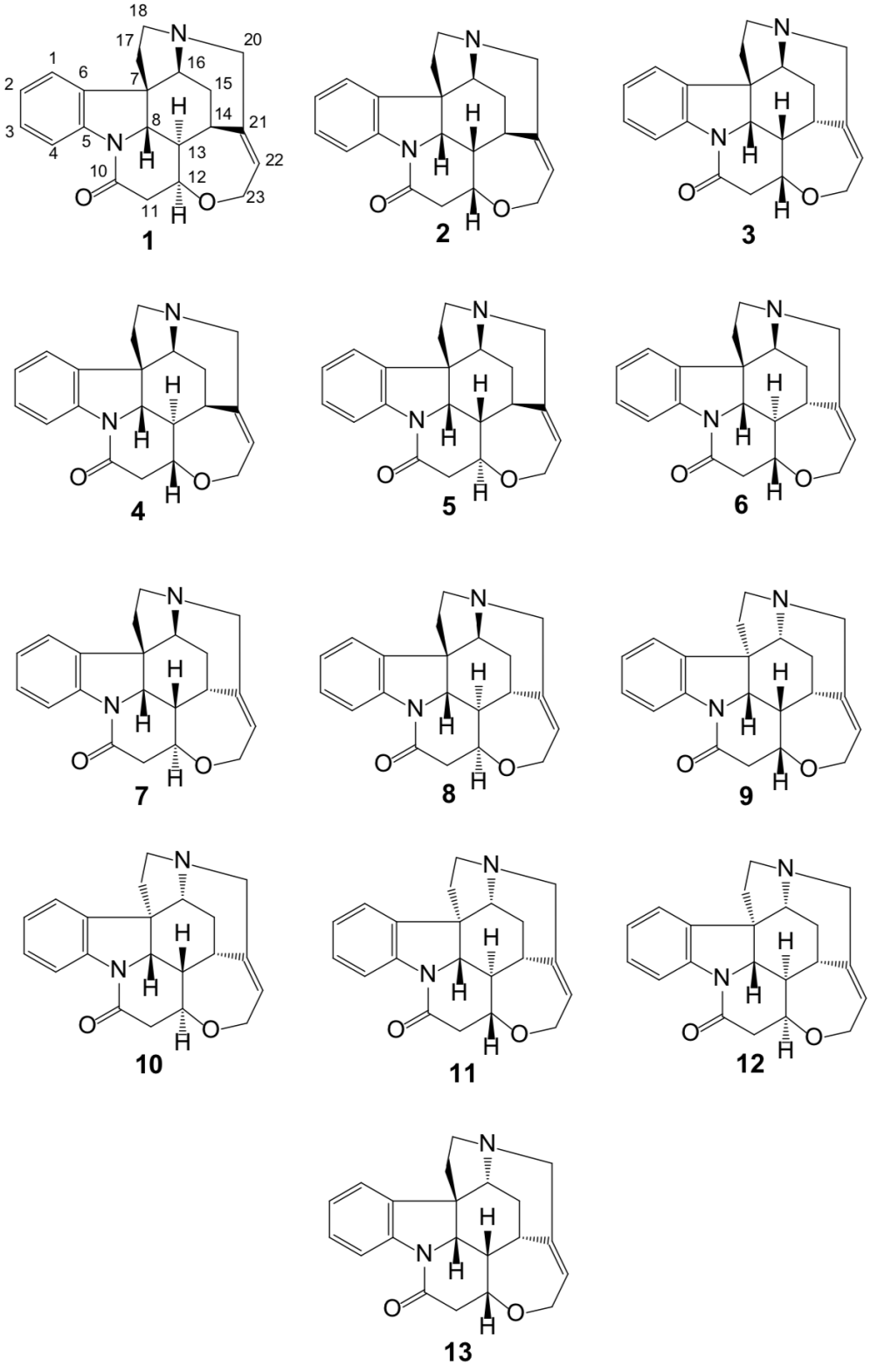

Figure 5.1: The structure of the natural occurring structure of strychnine (1), along with 12 energetically viable diastereomers (2-13).

5.1 $^1J_{CH}$ comparison

DFT provides the best discrimination between the structures as expected, identifying the correct structure (1) with a mean absolute error in $^1J_{CH}$ prediction of 1.25 Hz relative to the next closest match (6) of 2.01 Hz. The second best performing prediction method is the generation 1 model trained using DT4, identifying the correct structure with an MAE of 2.43 Hz relative to the next closest match (structure 2) of 2.63 Hz, a smaller difference than for the DFT predictions. Neither of the generation 2 trained models correctly identified the correct structure in this case, with structure 1 being ranked second for both models with MAEs of 6.73 Hz and 11.81 Hz for the DT45 and QM960k trained models respectively. Figures 5.2 and 5.3 show the results across all 14 structures, and highlights that for all models there is a correlation between the DFT prediction error and the machine learning prediction error for each structure. There is a clear difference in the MAE across the three models, with the generation 1 model producing predictions which match the experimental values to within 4.5 Hz for all structures, not just the correct diastereomer. In contrast the DT45 trained generation 2 model produces MAEs of 6 Hz and 10 Hz, and the QM960k trained generation 2 model produces MAEs between 11 Hz and 16 Hz. Whilst the DT4 trained generation 1 model obtains the correct result and the other 2 models do not, the difference in the discriminatory power between the models is minor. This is particularly clear in the errors adjusted using the softmin function (Figure 5.3), where the relative rankings of the structures appear similar, with the exception of the incorrectly chosen structures for the generation 2 models.

These results highlight two important factors in the task; the absolute prediction accuracy of each model on the target parameter being used to discriminate, and the degree of discrimination achieved between the structures. The gen 2 DT45 model is hampered by the fact it reports an even closer match for the predicted NMR parameters for structure 2. The ideal model for tasks such as this would provide predictions that are only accurate for the correct structure.

Despite these issues, there is clear utility in the predictions from these models. It is perhaps easier to imagine this utility in cases where the cost of the DFT predictions may become prohibitive, such as where the pool of possibly structures contains thousands of molecules. In this scenario the machine learning models could rapidly narrow the selection down to a handful of structures for which DFT calculations could be performed to obtain the final result. For all three models, if the machine learning results were used to select just 2 of the structures to take

for further analysis through DFT, the correct final assignment would be found. In a practical application the number of molecules to select would be dependent on the results of further testing of these models and the available CPU time, however it would be easy to see how reducing a potential dataset to the best 5-10% of structures as determined by machine learning could significantly improve a workflow, either through reducing overall computational cost, or allowing significantly higher quality of DFT calculation on the final structures.

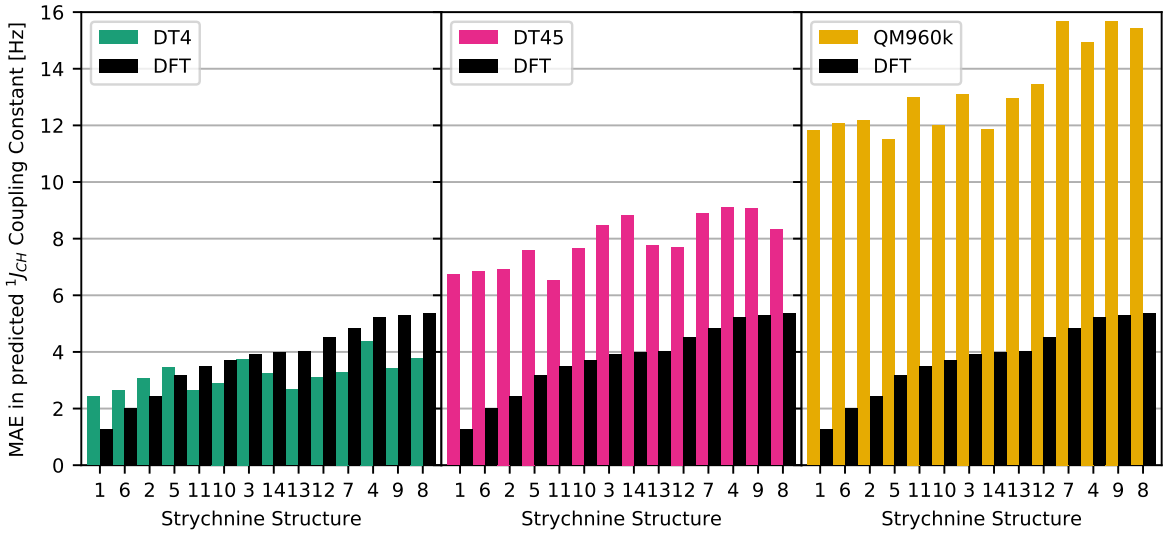

Figure 5.2: Mean absolute error between experimentally measured $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by mean absolute error in DFT prediction

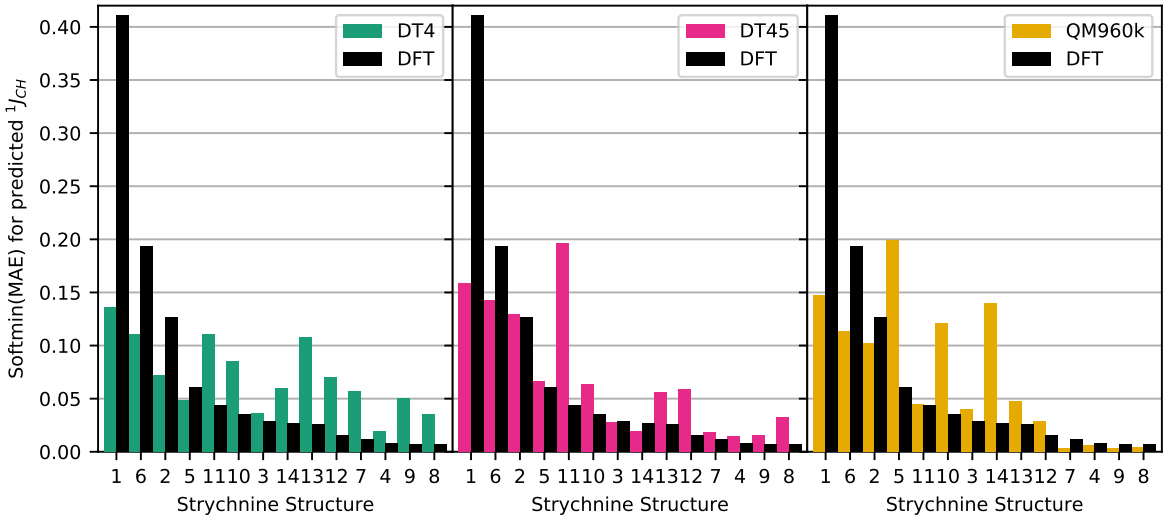

Figure 5.3: Mean absolute error, adjusted using a softmin function, between experimentally measured $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by softmin of the mean absolute error in DFT prediction

5.2 Geometric mean of δ^1H, $\delta^{13}C$, and $^1J_{CH}$ comparison

Improvements to the discrimination between the structures can be made by included multiple NMR parameters in the analysis. In this case the geometric mean across the mean absolute error in the parameters δ^1H, $\delta^{13}C$, and $^1J_{CH}$ is used instead of just the mean absolute error in $^1J_{CH}$.

With the inclusion of the additional NMR parameters, the DFT predicted values for structure 1 have a geometric mean mean absolute error (GMMAE) relative to experiment of 0.95 Hz, whereas the values for structure 2 have a GMMAE of 1.62 Hz (the second lowest GMMAE), this is a smaller difference than for $^1J_{CH}$ alone. Furthermore the difference in softmin probability between structures 1 (0.18) and 2 (0.12) is smaller in this case than it is for $^1J_{CH}$ alone (0.41 and 0.19 respectively). The inclusion of the extra NMR parameters therefore provides no benefit to the discriminatory power of the DFT method, though the results still provide a clear identification of the correct structure. For the DFT predictions the inclusion of δ^1H appears to have a negative effect on the accuracy, whilst the inclusion of $\delta^{13}C$ has a positive effect, the optimal results for the DFT method are obtained by using a combination of the two NMR parameters: $\delta^{13}C$ and $^1J_{CH}$. This is discussed at the end of this chapter.

For the machine learning models the inclusion of extra NMR parameters does show an improvement. For the gen 1 DT4 trained model, structure 1 is again identified as the correct structure with a GMMAE of 1.65 Hz, the next closest match is structure 6 in this case with a GMMAE of 1.94 Hz. This is a bigger difference than the 0.2 Hz difference between the two lowest MAE values for $^1J_{CH}$ alone. This also produces a clearer difference in the softmin probability with structure 1 now having a softmin value of 0.130, as opposed to 0.136, and the second best structure having a value of 0.10, as opposed to 0.11 for $^1J_{CH}$ alone. The improvement is more clearly seen in Figure 5.4 where the machine learning GMMAE values more strongly correlate with the DFT GMMAE values than they do in the equivalent $^1J_{CH}$ Figure (5.2.

The results for the generation 2 DT45 trained model are also improved in that now structure 1 is identified as the correct structure, though the relative discrimination between the structures is now marginally worse (Figure 5.2). The QM960k trained generation 2 model results change in the opposite way, with a stronger discrimination between different structures (Figure 5.2), correlating with the DFT results in terms of the order of closest match for most structures, but an incorrect structure is now assigned with even greater confidence (Figure 5.3).

For all three models the inclusion of further NMR parameters presents an improvement in the quality of predictions, especially if used in a situation where the cost of performing DFT calculations on the entire molecule candidate pool was prohibitive. All three models would include the correct structure in the top two structures by correlation to the experimental results. If the DFT calculations in this case were taken to be prohibitively costly, and it was desirable to limit them to only two or three structures, the machine learning results presented here would enable the selection of the most likely structures.

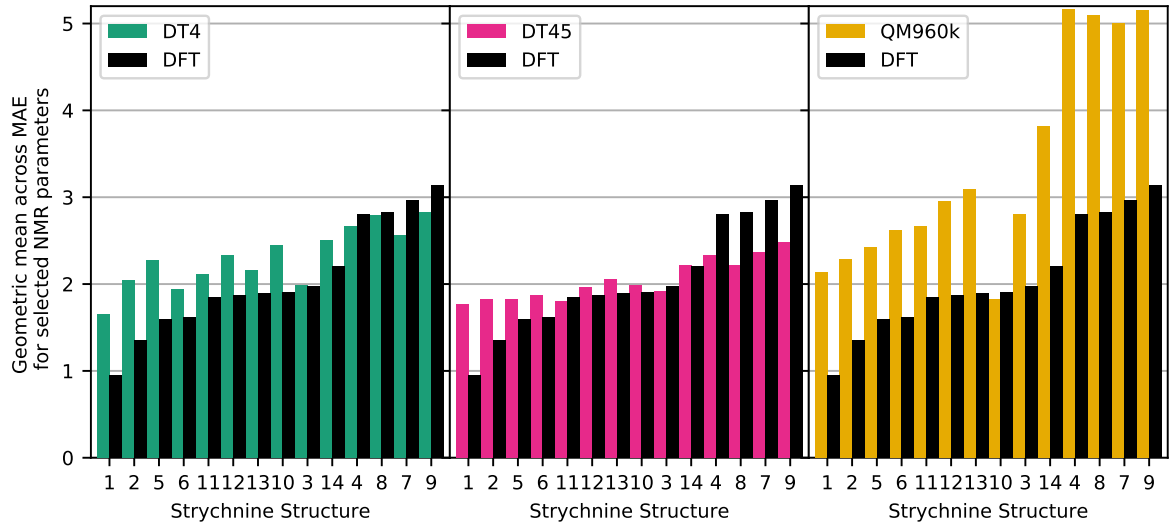

Figure 5.4: Geometric mean across the mean absolute error between experimentally measured δ^1H, $\delta^{13}C$, and $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by mean absolute error in DFT prediction

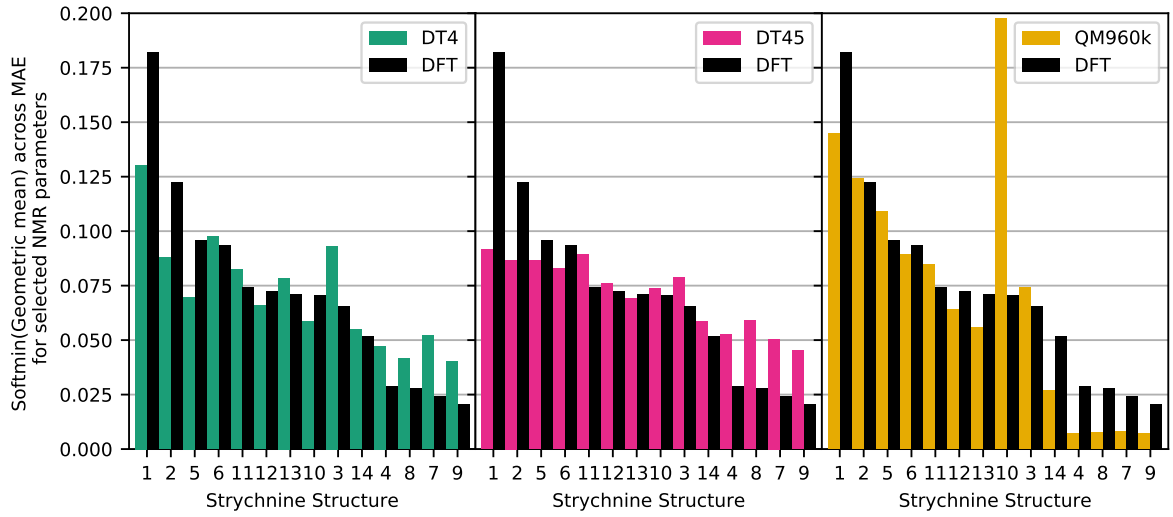

Figure 5.5: Geometric mean across the mean absolute error, adjusted using a softmin function, between experimentally measured δ^1H, $\delta^{13}C$, and $^1J_{CH}$ for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green), impression generation 2 trained on DT45 (labeled DT45, pink), impression generation 2 trained on QM960k (labeled QM960k, yellow), and DFT (labeled DFT, black) for all structures. Structures ordered by softmin of the mean absolute error in DFT prediction

5.3 Inclusion of further NMR parameters

The generation 2 models also possess the ability to predict further NMR parameters, and so the comparison between structures could be performed using any combination of δ^1H, $\delta^{13}C$, $^1J_{CH}$, $^2J_{CH}$, $^2J_{HH}$, $^3J_{CH}$, $^3J_{HH}$, those being the values for which experimental data is available for strychnine in this case. This yields a total of 127 possible metrics for the generation 2 models, making a detailed comparison between all possible metrics unfeasible.

To simplify this analysis the comparison was reduced to two key metrics which indicate the performance of the model and metric combination in the strychnine prediction task. Both metrics used here rely on the softmin populations of the mean or geometric mean, calculated through equation 5.1, the potential pitfalls of using these values having already been discussed. Firstly the difference in softmin population between the correct structure and the next closest matching structure is used to indicate the magnitude of the identification of the correct structure, and to highlight cases where the correct structure is not identified. Secondly the difference between the softmin population of the correct structure and the mean population of the remaining structures, to highlight how well the correct structure is identified relative to the majority of the incorrect structures.

Using these two metrics, all possible combinations of parameters for all models and DFT were assessed (the generation 1 model has 7 possible metrics due to the limited range of parameters, the generation 2 models and DFT have 127 possible metrics) and optimal combinations identified. Across all models no benefit was found in including more than 3 parameters in the comparison metric,

For DFT predicted values, the best performing single metrics are $\delta^{13}C$ and $^1J_{CH}$ (Figure 5.6), and the best metric overall is the combination of these two parameters (Figure 5.10). The additional inclusion of $^2J_{CH}$, $^3J_{CH}$, or $^3J_{HH}$ produced a similar result in terms of both the separation from the second closest matching structure and the mean of the incorrect structures.

For the generation 1 model trained using dataset 4, the best metric is $\delta^{13}C$ with $^1J_{CH}$ providing minimal benefit and the inclusion of δ^1H reducing the quality of the results (Figure 5.7). The single metric $\delta^{13}C$ is therefore used in the final comparison in Figures 5.10 and 5.11. The best metrics for the generation 2 models were a combination of three NMR parameters including $\delta^{13}C$ and $^1J_{CH}$, with $^3J_{HH}$ providing further benefit for the DT45 trained model, and $^2J_{HH}$ providing the same for the QM960k trained model. These metrics were used in the

final comparison in Figures 5.10 and 5.11. The best singular metric for the DT45 trained model was $^3J_{HH}$, with no other single metric assigning the correct structure (Figure 5.8). For the QM960k trained model, the best singular metric was $\delta^{13}C$, again with with no other single metric assigning the correct structure (Figure 5.9).

Overall it is unsurprising that each method of prediction performs best using a different set of metrics, considering the different training datasets and architectures being used, but the scale of the difference in performance on certain metrics between models is notable. The results for using δ^1H alone are similarly poor for all prediction methods, even DFT, and $\delta^{13}C$ is one of the best performing metrics across all the methods except the machine learning predictions from the generation 2, DT45 trained model, where it performs poorly. The MAE using $^3J_{HH}$ alone provides poor discrimination for the QM960k trained model whereas it is the best performing metric for the DT45 trained model. The complex differences between the performance of different combinations of NMR parameters in the comparison metric makes assigning an optimal metric difficult, and there is no reason to assume that the performance of these metrics in this task will generalise to other, even similar, tasks.

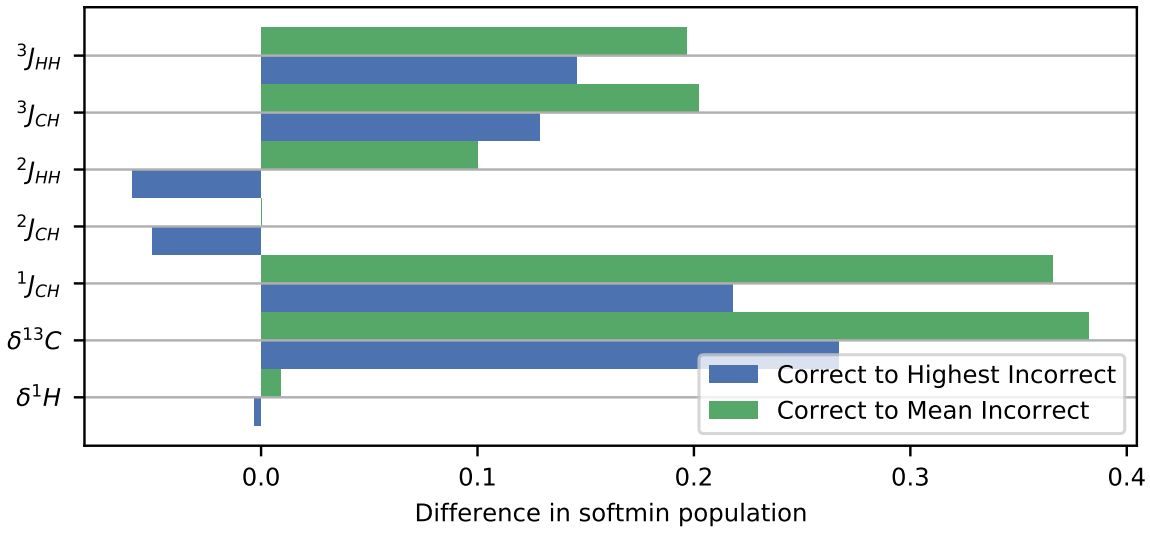

Figure 5.6: For the NMR parameters calculated by DFT. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures.

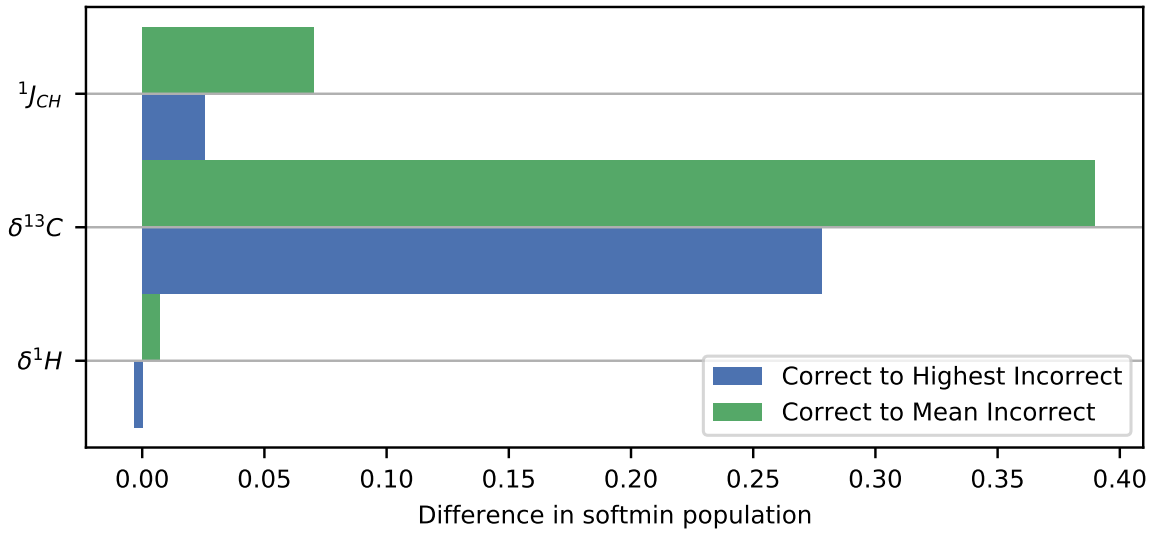

Figure 5.7: For the NMR parameters predicted by the generation 1 model trained using dataset 4. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures.

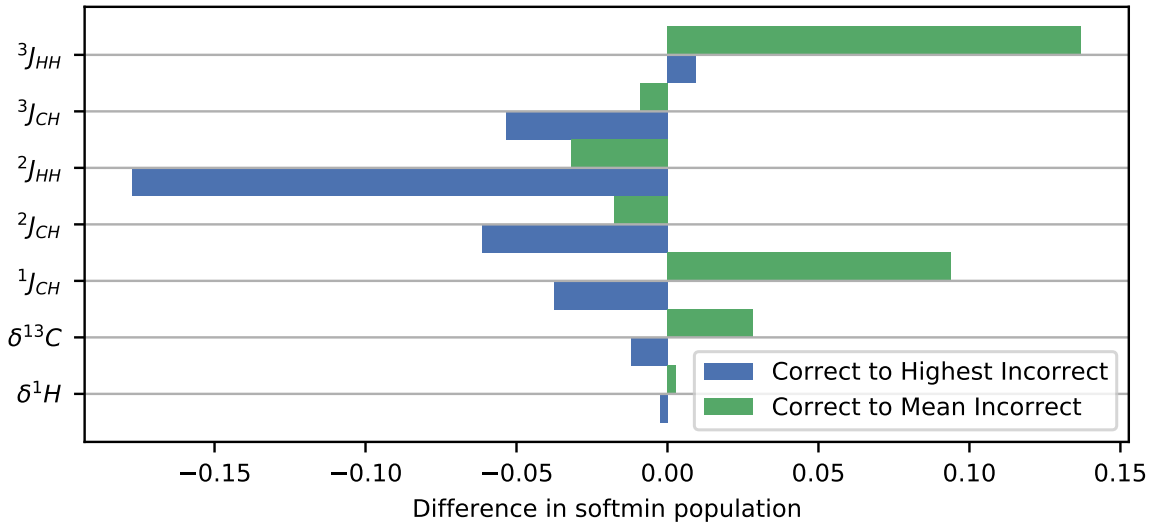

Figure 5.8: For the NMR parameters predicted by the generation 2 model trained using datasets 4 and 5. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures.

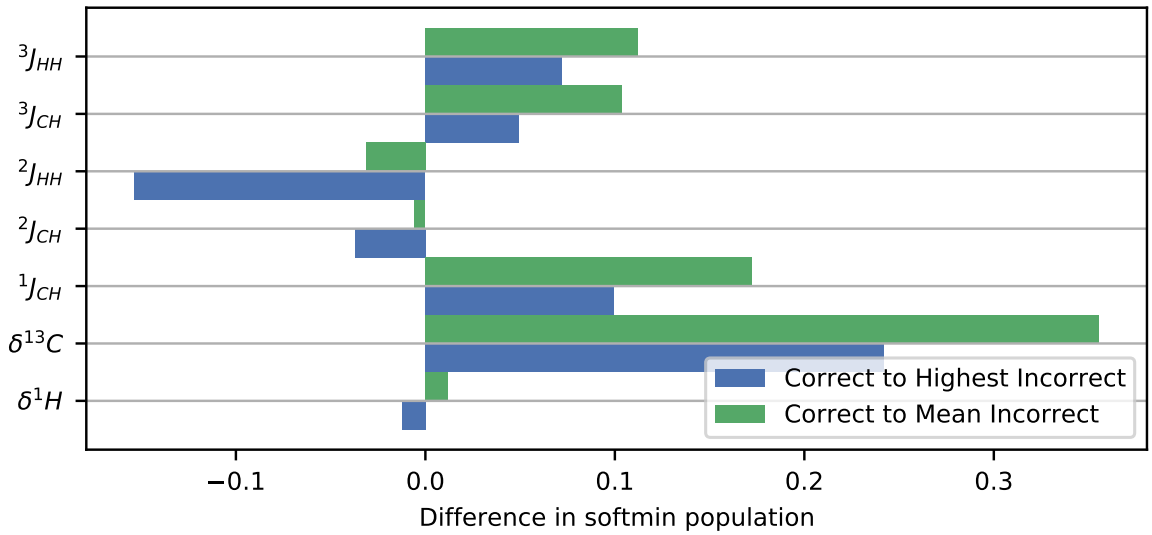

Figure 5.9: For the NMR parameters predicted by the generation 2 model trained using dataset QM960k. Difference in softmin calculated population, for different single NMR parameter metrics, between the correct structure (1) and the highest population incorrect structure, and between the correct structure and the mean population of the incorrect structures.

The final comparison graphs in Figures 5.10 and 5.11 give an indication of the optimal performance of each of the prediction methods in a task such as this, as in each case the combination of NMR parameters used in the comparison was selected based on performance on the same task. Qualitatively the generation 1 model trained on dataset 4 performs the best in this task relative to the DFT results, providing nearly the same degree of certainty in the identification of structure 1 as the correct structure. Surprisingly, for the generation 2 models, the QM960k trained model outperforms the DT45 trained model in terms of discrimination between the correct and incorrect structures as well as the correlation to DFT. The fact the DT45 trained generation 2 model performs the worst in this task is an unexpected result considering it achieved the best accuracy across datasets 3 and 5b (Chapter 4) across all parameters compared to the QM960k trained model, and outperformed, or performed very similarly to, the generation 1 model in terms of prediction accuracy on DT3 and DT5b for δ^1H, $\delta^{13}C$, and the experimental datasets for δ^1H, $\delta^{13}C$, and $^1J_{CH}$.

Figures 5.10 highlights the fact that good performance in the benchmark prediction task does not correlate to good performance in this task, as the DT45 trained generation 2 model provides accurate predictions, but is clearly less sensitive to the small changes in structure than the other two models. Conversely the DT4 trained generation 1 model accuracy is the worst of the three models in terms of overall GMMAE, but the predictions are clearly sensitive to the subtle differences in structure between the diastereomers and off-equilibrium structure.

It is unclear why the generation 1 model performs worst in terms of the mean absolute error on the correct structure, but best in terms of the differentiation between the correct and incorrect structures. It should perform worse due to the fact it has a smaller training dataset than either of the other molecules, and that that dataset suffers from the same limitations as the training datasets for the generation 2 DT45 trained model. The remaining factor which must therefore be responsible for this success is the model architecture and representation. The kernel ridge regression architecture with the FCHL representation is clearly providing an advantage to this model in this prediction task.

Further work is required to identify the extent to which these results generalise to similar tasks, however as an initial assessment of performance these results highlight both the potential utility of machine learning models as a cost-effective alternative to large-scale DFT calculations in tasks like this, at least in the initial stages, and the fact that performance in accuracy benchmark

tasks does not translate to performance in tasks like this in a straightforward way.

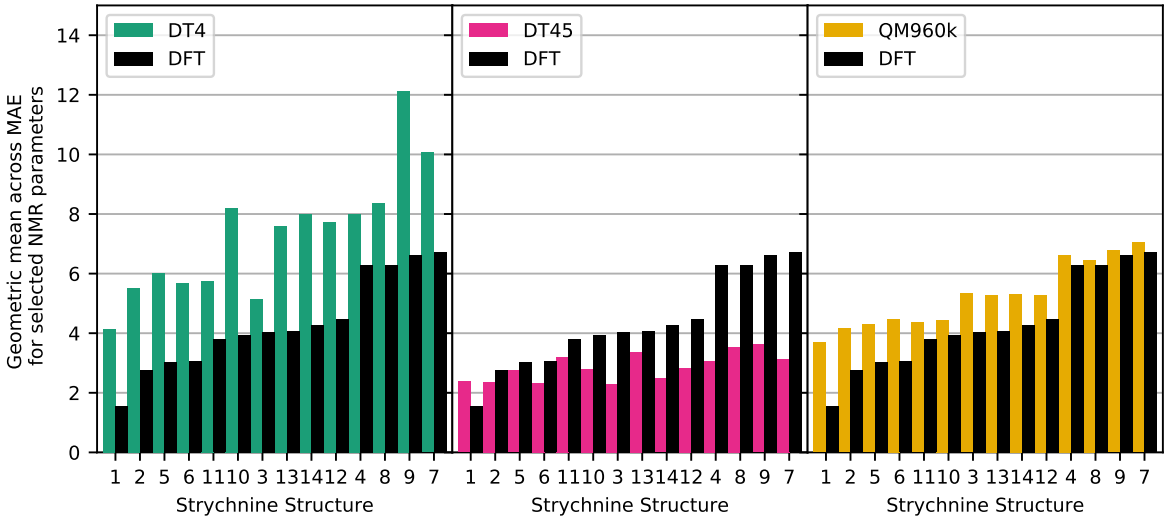

Figure 5.10: Score metric between experimentally measured NMR parameters for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green, using the MAE in $\delta^{13}C$ prediction), impression generation 2 trained on DT45 (labeled DT45, pink, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{3}J_{HH}$), impression generation 2 trained on QM960k (labeled QM960k, yellow, using the geometric mean across $\delta^{13}C$, $^{1}J_{CH}$, and $^{2}J_{HH}$), and DFT (labeled DFT, black, using the geometric mean across $\delta^{13}C$ and $^{1}J_{CH}$) for all structures. Structures ordered by mean absolute error in DFT prediction

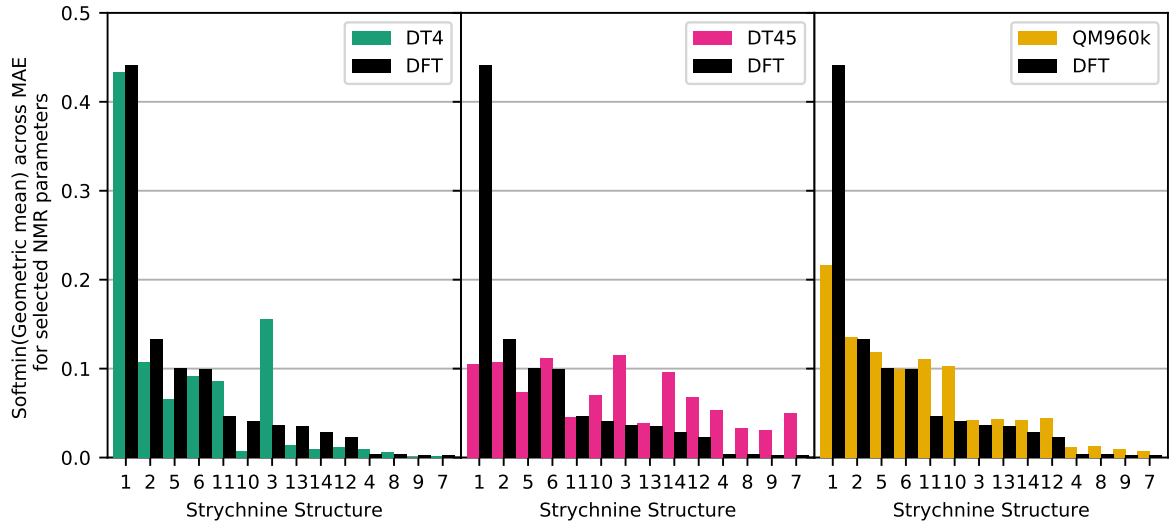

Figure 5.11: Relative softmin populations between experimentally measured NMR parameters for structure 1 and those predicted by impression generation 1 trained on DT4 (labeled DT4, green, using the MAE in $\delta^{13}C$ prediction), impression generation 2 trained on DT45 (labeled DT45, pink, using the geometric mean across $\delta^{13}C$, ${}^{1}J_{CH}$, and ${}^{3}J_{HH}$), impression generation 2 trained on QM960k (labeled QM960k, yellow, using the geometric mean across $\delta^{13}C$, ${}^{1}J_{CH}$, and ${}^{2}J_{HH}$), and DFT (labeled DFT, black, using the geometric mean across $\delta^{13}C$ and ${}^{1}J_{CH}$) for all structures. Structures ordered by mean absolute error in DFT prediction

CHAPTER 6

IMPRESSION FOR BINDING AFFINITY PREDICTION

6.1 Predicting Binding Affinity

One of the most important tasks in the development of novel drug compounds is the evaluation of how well the new compound binds to the target. Binding affinity assays over large compound libraries are regularly carried out to find promising compounds which bind well to a given target molecule. Performing binding assays in vivo in such a situation is very expensive and so cheaper, computational methods are often used [134–136]. Computational techniques can accurately predict the binding affinity of compounds to a given target, in particular free-energy perturbation [137] calculations which rely on molecule dynamics simulations is one of the most popular and widely used methods. Like DFT NMR calculations, these techniques are computationally expensive, and so machine learning methods again offer a potentially significant improvement to the process [138]. Machine learning techniques to predict binding affinity often generalise to the prediction of binding affinity for any ligand-target pair [138–140], however the focus of this application is to predict the binding affinity relative to a specific target, where the target is not a part of the input representation.

Free energy perturbation (FEP) methods predict the free energy difference between the unbound and bound states of a given ligand-target complex using molecular dynamics and computational force fields to simulate the movement of the molecules and calculate the energy of the system. [141–143] The improvement of these calculations both in terms of speed and accuracy

is a desireable aim, however a far simpler problem to solve is how to quickly obtain accurate predictions for the binding affinity for a given set of compounds to a specific target, using binding affinity calculations for a subset of these compounds to the desired target. This approach provides a balance between the expense of binding affinity calculations and the inaccuracy of pure machine learning models.

In an industrial setting, the purpose of a given study is to identify the compounds in a large pool of potential candidates which have the highest binding affinity to the chosen target. To do this, sets of compounds are chosen from the compound pool and the binding affinity to the target calculated through an FEP calculation. These FEP calculations are then used to train a model to predict the value for the remaining compounds in the pool. These predictions are used to select the next batch of compounds for calculations, usually selecting the best predicted binders, typically up to 5 rounds are performed, at which point most of the best binding molecules should have been identified. The results of this are that the binding affinity values have been calculated for the best binding compounds in the pool. Work was undertaken as a part of this project to identify whether improvements could be made to the selection process of each batch of compounds through active learning, and whether an adapted IMPRESSION model could provide better FEP predictions for each round than a reference model.

6.1.1 pChEMBL

The target values which the machine learning models will be trained to predict are the pChEMBL values obtained from the ChEMBL database. The pChEMBL value is the negative logarithm of one of several values (IC50, EC50, Kd, etc), and makes comparisons across these measures of half-maximal response concentration/potency/affinity possible. This means that throughout this chapter the values presented as pChEMBL are comprised by any number of these quantities, the purpose of the conversion to pChEMBL is to make them comparable for the purposes of evaluating relative binding affinity of different compounds. [101]

6.2 Model Architecture

6.2.1 ECFP4 neural network reference model

In order to provide a representative example of current techniques in FEP prediction [144, 145], a model was constructed based on a simple neural network architecture with Extended Connectivity Fingerprints (ECFPs, [146]) as the input. The model was designed and the code written in part by Calvin Yiu, a PhD student in the Butts research group.

The model consists of 3 linear feed-forward network layers with trainable weights, each being followed by the rectified linear unit function [147]. The input features for the model are extended connectivity or Morgan fingerprints [148], a 2-dimensional representation, representing the surrounding chemical structure through circular atom neighborhoods, with variable diameter (which affects the length of the fingerprint). The most commonly used diameter is 4, and is referred to then as the ECFP4 fingerprint [149]. This model is referred to as the ECFP4 model for the remainder of this chapter.

6.2.2 IMPRESSION for molecular properties

A relatively straightforward adaptation of the IMPRESSION Generation 2 architecture yields a model which can learn one property per molecule rather than per atom or per bond. A global attention pooling step is added to the end of the graph transformer network model architecture (outlined in Chapter 4), which allows information to pass between all nodes/edges of the molecular graph. Two linear model layers then converge to a single output value for each graph. This model is referred to as the IMPRESSION model for the remainder of this chapter.

6.3 Active Learning

In order to obtain the best predictions for the majority of the dataset, different strategies can be used to select molecules for the training set. Active learning selection strategies use predictions from the current or previously trained models to inform the selection of new molecules for the training set. This is similar to the method used to select the DFT NMR training dataset 4 (section 2.4.3). Several strategies were investigated, labelled F1-5. The selection of molecules at random is labelled scheme A, and acts as the reference scheme in this case:

- Select molecules at random (A)
- Select molecules with the lowest predicted pChEMBL values (F1(low))
- Select molecules with the highest predicted pChEMBL values (F2(high))
- Select molecules such that the range of predicted pChEMBL values over the entire dataset is covered evenly by the predictions for the selected molecules (F3(range))
- Select molecules such that the distribution of predicted pChEMBL over the entire dataset is matched by the predictions for the selected molecules (F4(distribution))
- Select molecules such that the inverse of the distribution of predicted pChEMBL over the entire dataset is matched by the predictions for the selected molecules (F5(inverse))

Selection schemes F1(low) and F2(high) are straightforward in their implementation, the first or last n molecules in a dataframe sorted by predicted pChEMBL value are selected (where n is the number of molecules to be selected in each round). Scheme F2(high) is equivalent to what is referred to as an enrichment scheme, where molecules likely to be good binders are preferentially selected, if the initial predictions are relatively accurate this results in more, better predicted good binding molecules in the final dataset. For selection scheme F3(range), n evenly spaced values are generated between the minimum and maximum predicted pChEMBL values, the n molecules with pChEMBL values closest to these values are selected. For scheme F5(inverse) the probability of selection for each molecule is calculated as the sum of the differences between its pChEMBL value and all other predicted pChEMBL values, for scheme F4(distribution) the inverse of this probability is used. For both F4(distribution) and F5(inverse), n molecules are then chosen at random, weighted by the calculated probability.

6.3.1 Identification of Binders

In the practical application of these models, the ability of the model to accurately predict any one value is secondary to its ability to distinguish the few molecules with the highest binding affinity from the rest of the molecule pool. As such the model performance is evaluated as a classification task. The most common way to evaluate classification models such as this is through a Receiver Operating Characteristic (ROC) curve [150], whereby the discriminatory power of the model is highlighted, and is comparable across different models, datasets and applications. The ROC

curve is constructed by evaluating the model predictions for a continuous series of classification cutoff values, where the cutoff in this case is the difference between a pChEMBL value being assigned as a high binding affinity or a low binding affinity. For each value of the classification cutoff, the true positive rate is calculated, given by:

$$\text{True Positive Rate} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{6.1}$$

and the false positive rate is calculated, given by:

$$\text{False Positive Rate} = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}} \tag{6.2}$$

The area under the curve (AUC) is a common metric used to compare classification models [151], calculated as the area under the ROC curve. A model with no discriminatory power, i.e. it predicts all molecules to have a pChEMBL of 0.0 would have an AUC value of 0.5, as it would present a straight line in an ROC plot: $y = x$. In order to simplify the comparison between each model and each selection scheme, a set of binary labels were assigned to the molecules based on the pChEMBL binding affinity obtained from ChEMBL. Molecules with pChEMBL greater than 7 were assigned as 'good' binding molecules, and molecules with pChEMBL less than or equal to seven assigned as 'bad' binding molecules. This simplifies the analysis of the ROC curves as the in this case the number of True and False positives does not change with the classification cutoff.

The two different binding targets will naturally produce a different binding affinity from different compounds to each other. The best binding molecules (according to pChEMBL score) are shown for each target in Figures 6.1 and 6.2. There are clear differences in the molecules which possess the highest binding affinity for each structure. Furthermore this makes clear another aspect of the ChEMBL database, namely that whilst it contains a very wide range of molecules, it also contains many molecules which are extremely similar to one another, many of the best binding molecules are merely the same molecule with one or two atoms substituted. The only significant difference in these sets of molecules which may make one dataset harder to predict than the other, is the prevalence of sulphur and fluorine nuclei in the best binding molecules for CDK2. Assuming the pattern from previous work is extended to the models trained to predict binding affinity, the model will be worse at making predictions for compounds which contain nuclei that are otherwise rare in the dataset. This could potentially lead to the CDK2 trained models having a poorer prediction accuracy on the highest binding molecules, as these are more likely to contain rarer nuclei.

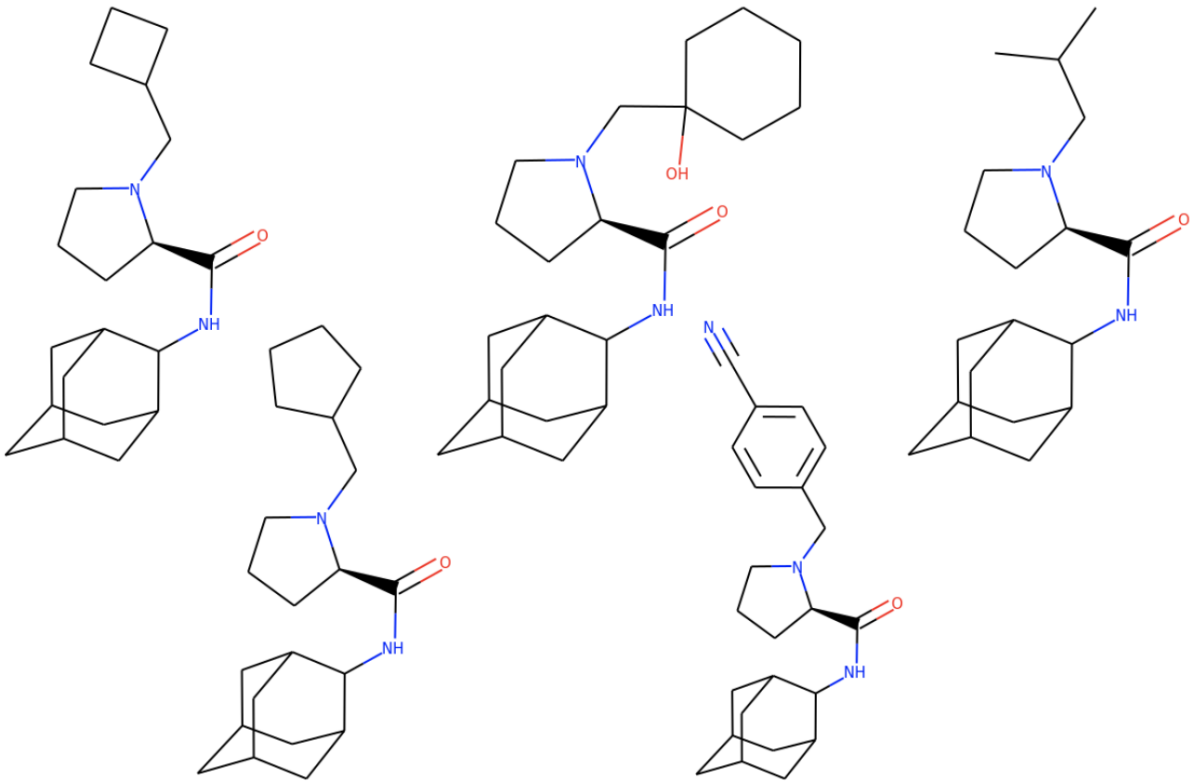

Figure 6.1: The 5 best binding molecules for the HSD11 target, as ranked by pChEMBL value. Molecule IDs: CHEMBL1098145 CHEMBL1096451 CHEMBL1098130 CHEMBL1096870 CHEMBL1098131

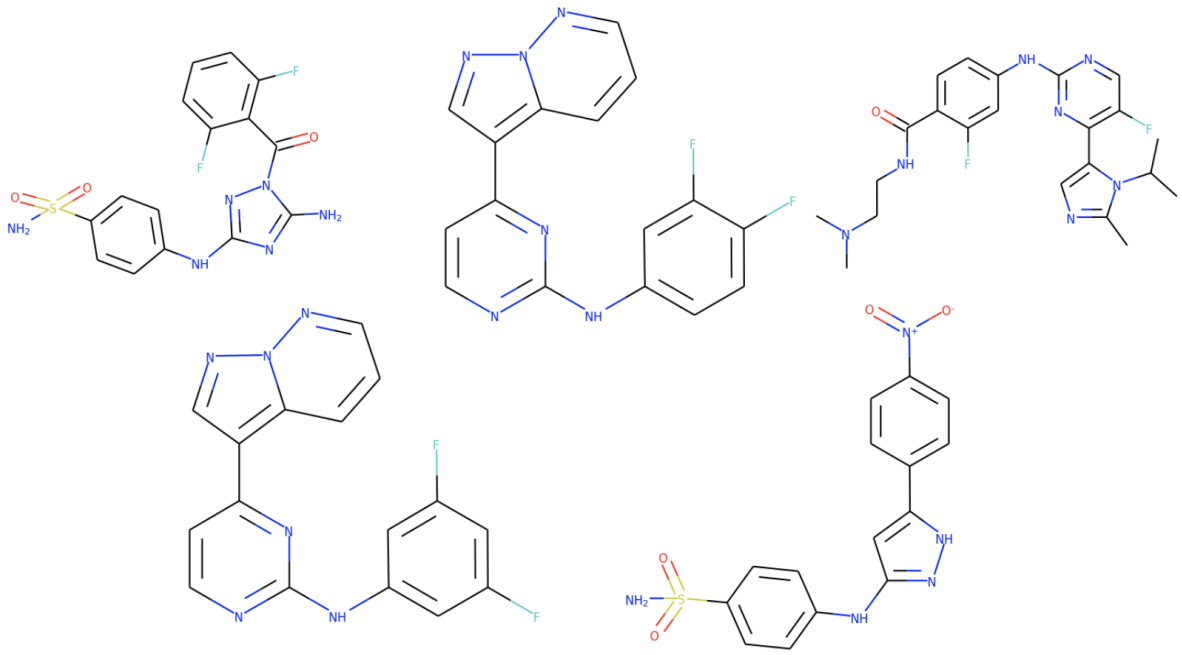

Figure 6.2: The 5 best binding molecules for the CDK2 target, as ranked by pChEMBL value. Molecule IDs: CHEMBL462385 CHEMBL191003 CHEMBL364370 CHEMBL184510 CHEMBL317703

6.4 Results

6.4.1 Training and Testing Datasets

Two test datasets were obtained from ChEMBL with the associated pChEMBL binding affinities: a set of 2,698 molecules targeting the enzyme 11-beta-dehydrogenase HSD11, and a set of 1,362 molecules with associated binding affinity to the cyclin-dependent kinase CDK2. The size of molecule in the datasets vary between 8 and 77 atoms, the mean molecule size for both datasets is between 26 and 28 atoms, and there are very few molecules with more than 45 atoms (Figure 6.4, left). The distribution of pChEMBL values is similar between the two datasets, though the values are higher on average in the HSD11 dataset than in the CDK2 dataset. Both datasets contain a large number of molecules with pChEMBL greater than 7.0 which will be used as the classification threshold, discussed below.

To identify whether the HSD11 and CDK2 datasets are biased towards a particular class of molecules the distribution of Tanimoto similarity (based on the ECFP4 fingerprint) between each pair of molecules in the datasets was compared to those from a random selection of molecules (of similar molecular weight) taken from ChEMBL. Figure 6.3 clearly shows that there are only a very small number of pairs of molecules in the HSD11 and CDK2 datasets with similarity higher than any pairs in the random set. This can be seen from the very similar tails on the right side of all three distributions. This means that the molecules in the HSD11 or CDK2 datasets are no more similar to each other than a random pair of molecules from the ChEMBL dataset, therefore these are suitable test sets and results established for these datasets should hold reasonably well for other molecules across the entire ChEMBL database.

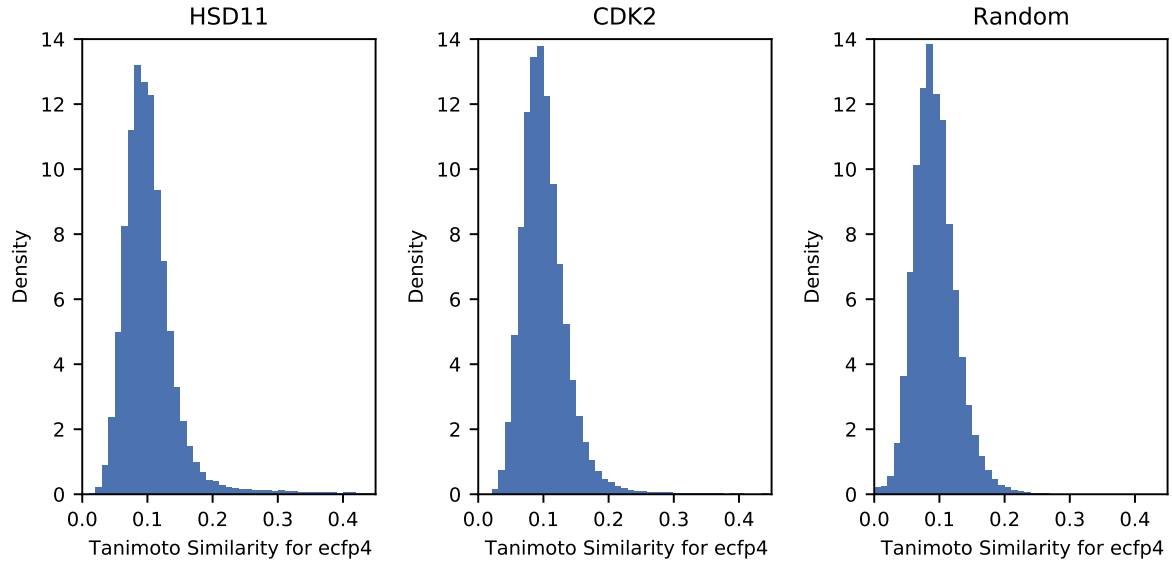

Figure 6.3: Plots of the tanimoto similarity for HSD11, CDK2 and random datasets from CHEMBL

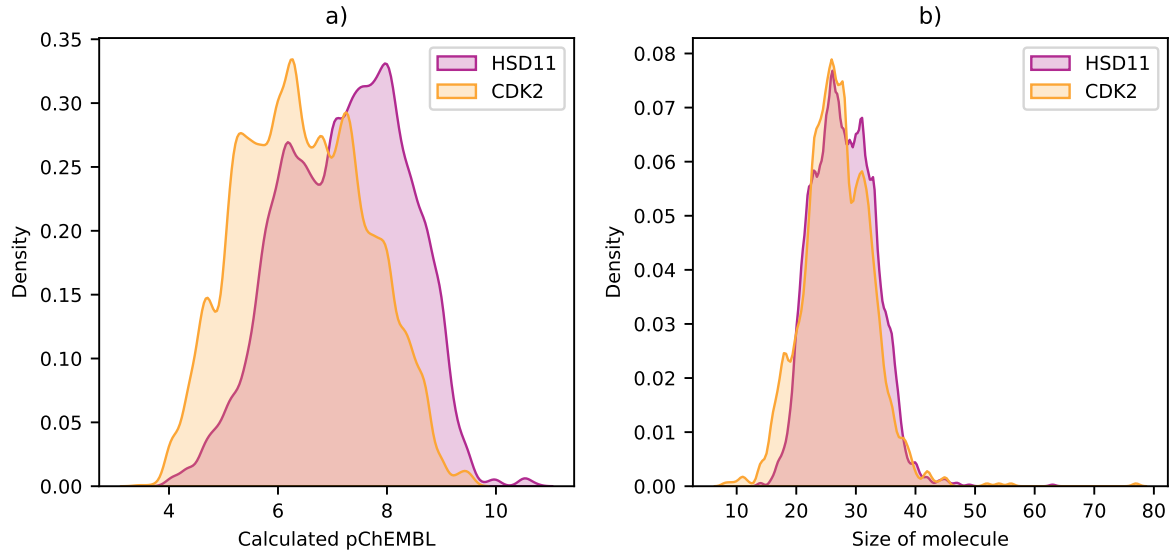

Figure 6.4: Distribution of pChEMBL values (a), and distribution of molecules sizes (b) for the HSD11 and CDK2 datasets

6.4.2 Model Training and regression performance

The initial validation for both models was performed by training on 1000 molecules selected at random, this was performed separately for both the HSD11 and CDK2 datasets. The models were trained for 100 epochs, however the prediction accuracy on molecules outside of the training set saw minimal improvement beyond 40 epochs (Figure 6.5) for the IMPRESSION based model, and beyond 20 epochs for the ECFP4 model. Models were only trained to 40 epochs in all subsequent training runs, in order to save computation time. Each training epoch for the 1000 molecule training set for the IMPRESSION model takes around 10 seconds to perform, for the ECFP4 model it is around 2 seconds (both run on a single Quadro K620 nvidia GPU). The time saving for training a single model is therefore modest but for two models, across two datasets for 6 different selection schemes this adds up to hours of unnecessary computational time.

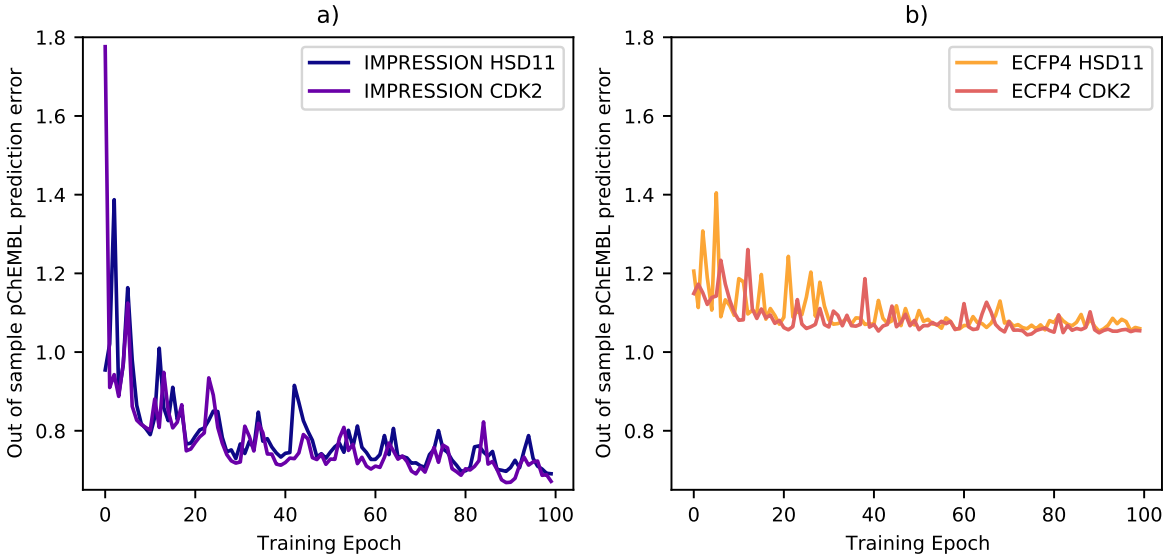

Figure 6.5: Out of sample learning curves for the IMPRESSION model (a) and ECFP4 neural network (b), for datasets HSD11 and CDK2. The out of sample error is the mean absolute error in prediction of pChEMBL for molecules not in the training dataset (1000 molecules).

The performance of the models is assessed firstly as a regression task. The size of the training set is 1000 molecules, which leaves 1698 molecules in the testing set for HSD11 and 362 molecules in the testing set for CDK2. The IMPRESSION model achieves a mean absolute error (MAE) of 0.78 and 0.75 for the HSD11 and CDK2 testing datasets respectively. The ECFP4 model achieves an MAE of 1.05 and 1.13 on the HSD11 and CDK2 testing datasets, considerably worse than the IMPRESSION model. The root mean squared deviation (RMSD) and maximum error (MaxE) are also higher across both datasets for the ECFP4 model. A summary of these results is shown in Table 6.1.

The IMPRESSION model shows a clear advantage in prediction accuracy over the ECFP4 model, and this is further supported by the shape of the error distributions (Figures 6.6a and 6.8a), with the IMPRESSION model showing a sharper peak better centered towards zero than the equivalent error distribution for the ECFP4 model. The ECFP4 model predictions contain a far greater number of large errors than the IMPRESSION model predictions, the difference being most clear in the scatter plots in Figures 6.6b and 6.8b. Furthermore the 2D histograms in Figures 6.7 and 6.9 highlight how the IMPRESSION predictions show a much better correlation, many more predictions lie on or near the $x = y$ line for both datasets, however this is especially clear for the CDK2 dataset.

The ECFP4 model appears to suffer most through the under prediction of pChEMBL values in both datasets, however it is clear that the IMPRESSION model fails to predict any pChEMBL value below 5.5 for the HSD11 dataset, despite a considerable number of these values being present in the dataset. This will have an affect on the models performance in the classification tasks as the bias towards higher predicted pChEMBL values will increase the number of molecules identified with high binding affinity, but equivalently increase the false positive rate. The consistent under prediction of values will have the opposite effect.

Model	Dataset	Train/Test	MAE	RMSD	MaxE
IMPRESSION	HSD11	1000/1698	0.78	0.99	3.38
IMPRESSION	CDK2	1000/362	0.75	0.93	3.32
Base Model	HSD11	1000/1698	1.05	1.29	4.39
Base Model	CDK2	1000/362	1.13	1.41	4.29

Table 6.1: Model performance for IMPRESSION and the base model in the prediction of pChEMBL for the HSD11 and CDK2 datasets

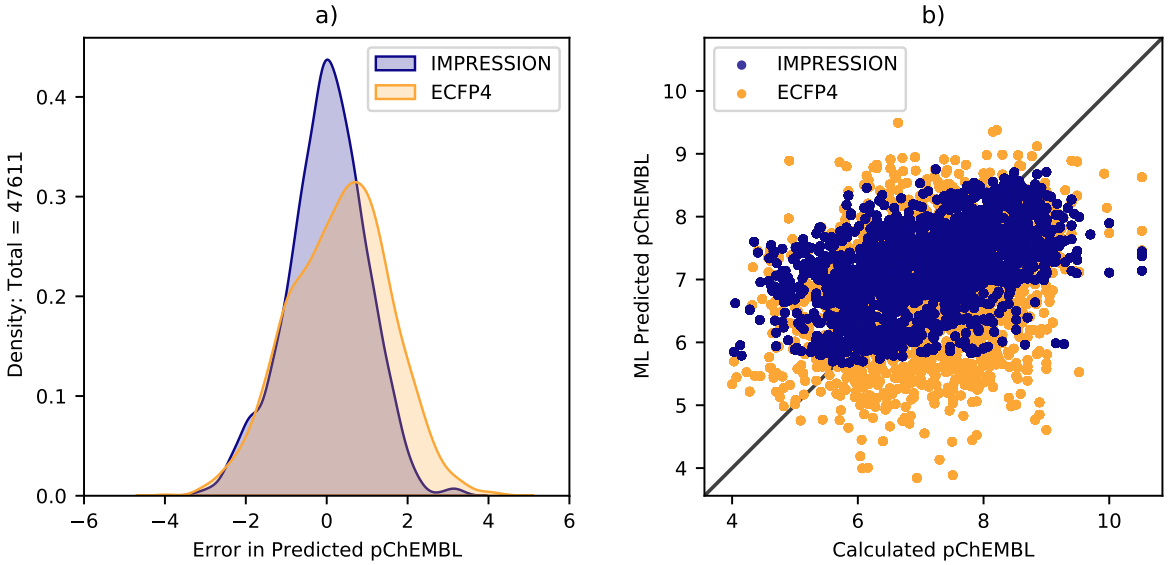

Figure 6.6: Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the HSD11 dataset. Errors displayed as error distributions (a) and scatter plots (b). Fit statistics for IMPRESSION: 0.78 MAE, 0.99 RMSD, 3.38 MaxE, fit statistics for ECFP4: 1.05 MAE, 1.29 RMSD, 4.39 MaxE.

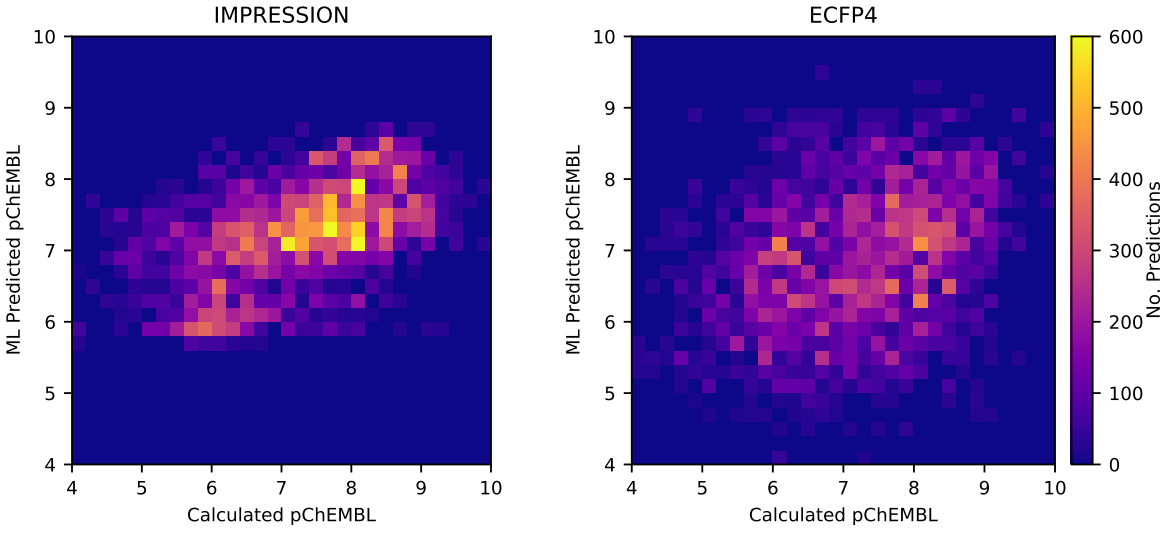

Figure 6.7: Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the HSD11 dataset. Errors displayed as 2D Histograms. Fit statistics for IMPRESSION: 0.78 MAE, 0.99 RMSD, 3.38 MaxE, fit statistics for ECFP4: 1.05 MAE, 1.29 RMSD, 4.39 MaxE.

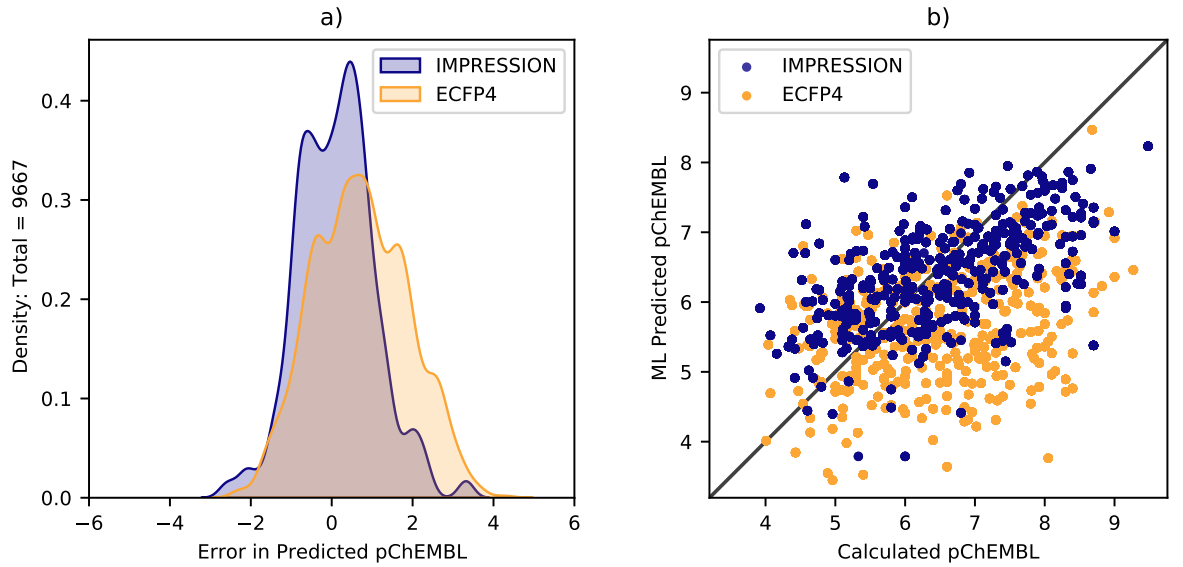

Figure 6.8: Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the CDK2 dataset. Errors displayed as error distributions (a) and scatter plots (b). Fit statistics for IMPRESSION: 0.75 MAE, 1.29 RMSD, 4.39 MaxE, fit statistics for ECFP4: 1.13 MAE, 1.41 RMSD, 4.29 MaxE.

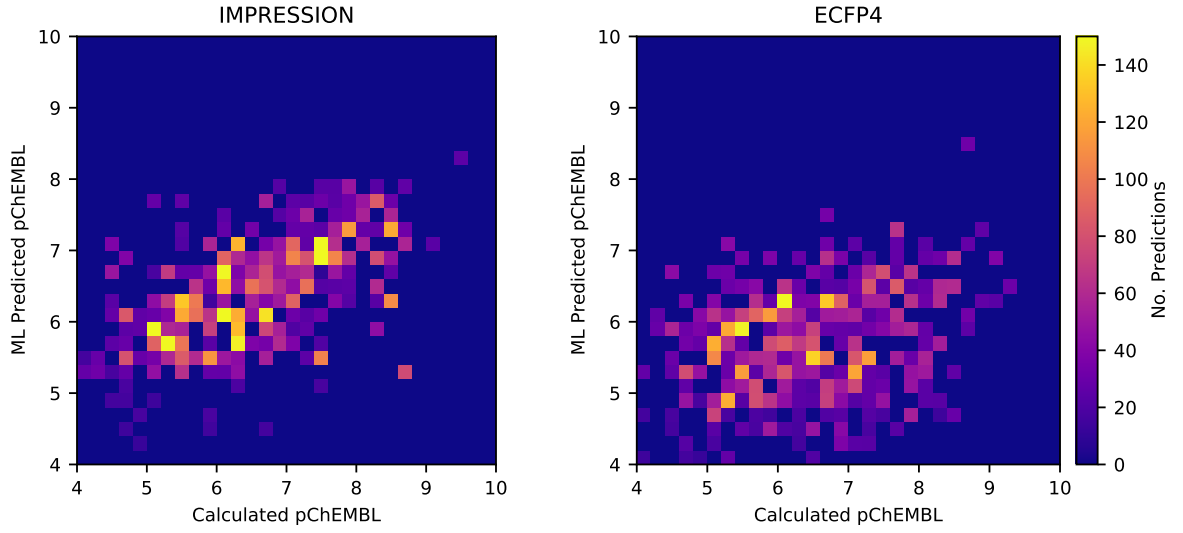

Figure 6.9: Prediction error on the remaining 1698 molecules in the dataset for the IMPRESSION and ECFP4 models trained on 1000 randomly selected molecules, for the CDK2 dataset. Errors displayed as 2D Histograms. Fit statistics for IMPRESSION: 0.75 MAE, 1.29 RMSD, 4.39 MaxE, fit statistics for ECFP4: 1.13 MAE, 1.41 RMSD, 4.29 MaxE.

6.4.3 Classification performance

The IMPRESSION model also performs better than the ECFP4 model when the task is converted to a classification task (i.e. a binary assignment of good or bad binding, based on a cutoff value). The receiver operating characteristic (ROC) curve for the IMPRESSION model shows a significantly better True to False positive ratio than the ECFP4 model across the entire range of classification cutoff values, for both datasets (Figure 6.10). The area under the curve (AUC) values for the IMPRESSION model are significantly higher than for the ECFP4 model, 0.76 and 0.82 (IMPRESSION) compared with 0.65 and 0.65 (ECFP4) for the HSD11 and CDK2 datasets respectively.

The IMPRESSION model therefore demonstrates significantly better discriminatory power in identifying higher or lower binding affinities against the reference model. This translates to an ability to select better candidate molecules for binding to a specific target, with higher binding affinity and fewer false positives.

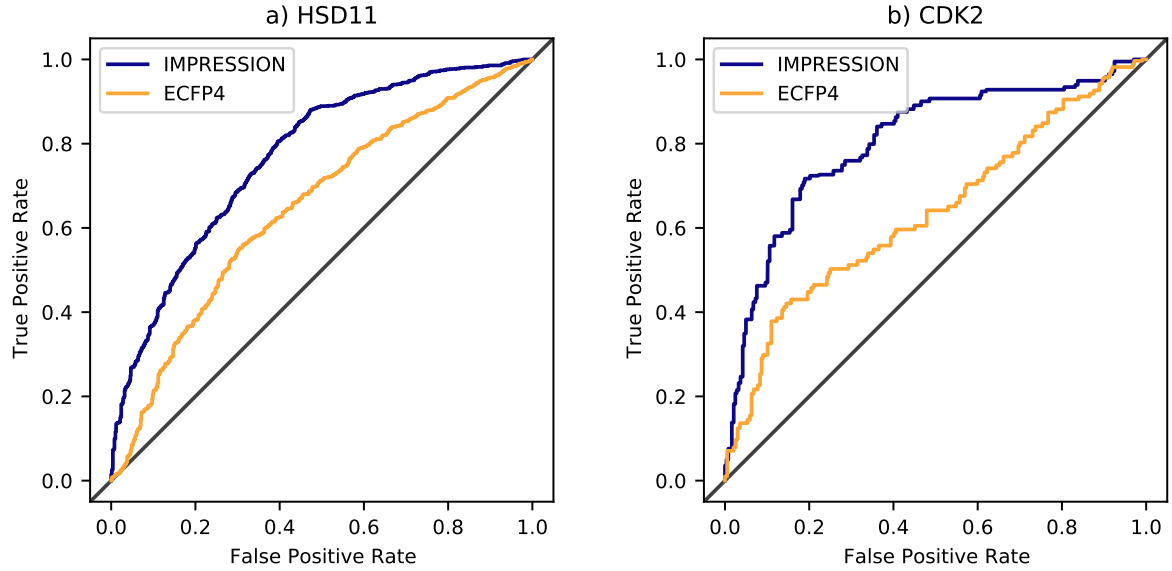

Figure 6.10: Receiver operating characteristic (ROC) plots for the IMPRESSION and ECFP4 models, for the HSD11 (a) and CDK2 (b) datasets.

Model	Dataset	Train/Test	AUC Score
IMPRESSION	HSD11	1000/1698	0.7648
IMPRESSION	CDK2	1000/363	0.8181
ECFP4	HSD11	1000/1698	0.6511
ECFP4	CDK2	1000/365	0.6494

Table 6.2: AUC scores for both models trained using 1000 molecules from each dataset, tested on the remaining molecules.

6.4.4 Active learning molecule selection

In order to improve the model performance several strategies in active learning were investigated. Each selection scheme (A, F1-5) is evaluated based on its performance over 9 selection rounds selecting 100 molecules in each round according to the selection scheme criteria. For each round the predictions generated on the remaining, unseen molecules, are used to generate an ROC curve, and the AUC value is calculated for that curve. Both models were used for each selection scheme, with the ECFP4 model showing little difference between the selection schemes, and little improvement in performance with further training rounds for the HSD11 dataset (Figure 6.11b). The same pattern is observed for the CDK2 dataset with the exception of the F4(distribution) selection scheme which shows a significant improvement over the 9 selection rounds, and outperforms the random selection scheme from round 4 onwards. Scheme F4(distribution) selected molecules according to the distribution of predicted pChEMBL values, molecules with commonly predicted pChEMBL values were more likely to be chosen. Scheme F4(distribution) is also the top performing scheme for the ECFP4 model in the HSD11 test case, achieving the highest AUC score in rounds 1-8. For the IMPRESSION model, a much greater improvement is seen over the 9 selection rounds for selection schemes A, F4(distribution) and F5(inverse).

The active learning selection schemes provide no clear advantage over the random selection of molecules, however the manner in which molecules are selected does appear to have a significant effect on the model performance, as the poor performance of selection schemes F1(low) and F2(high) demonstrate (Figure 6.11a). It is possible therefore that the success of the random selection scheme is an artifact of the use of a small selection pool, and the schemes F4(distribution) and F5(inverse) may yet present an advantage in selection molecules from selection pools of 10,000 or more molecules as is commonly the case in real binding affinity studies. To perform the

above analysis for such a pool would require the calculation of 10,000 binding affinities to a single target, in order to verify the results, and so is not feasible in the timescale of this thesis work.

The IMPRESSION model shows a clear advantage over the reference model in all cases, demonstrating a significantly improved ability to identify higher binding molecules than the ECFP4 model. Models similar to the ECFP4 model are used frequently in industrial applications, and so these results suggest improvements to these processes could be made with the use of a graph transformer network based model.

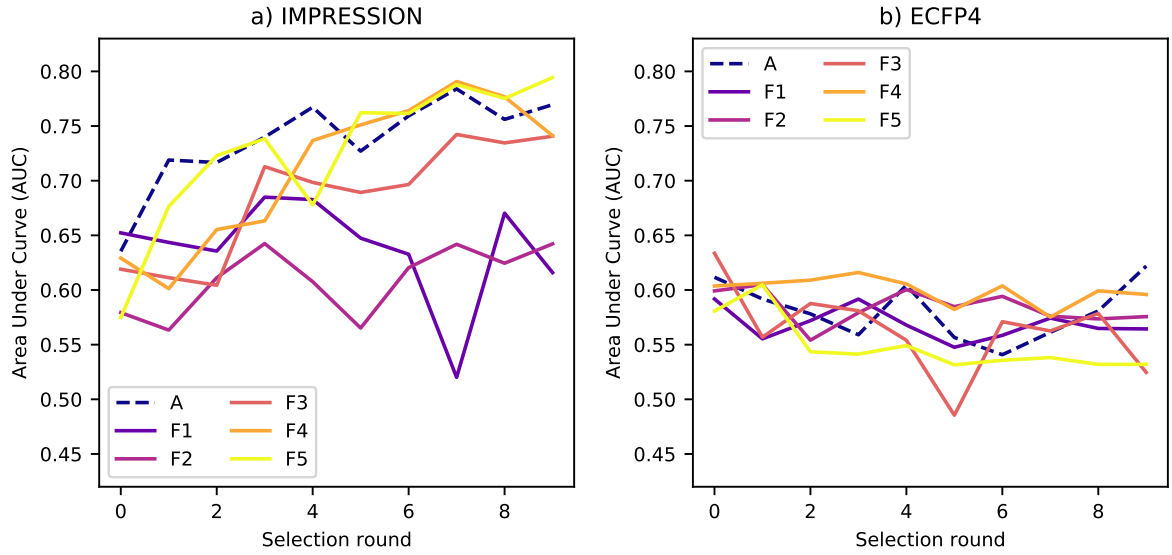

Figure 6.11: AUC at each selection round for the IMPRESSION (a) and ECFP4 (b) models, for each selection scheme, for the HSD11 dataset. Select schemes: F1(low), F2(high), F3(range), F4(distribution), F5(inverse)

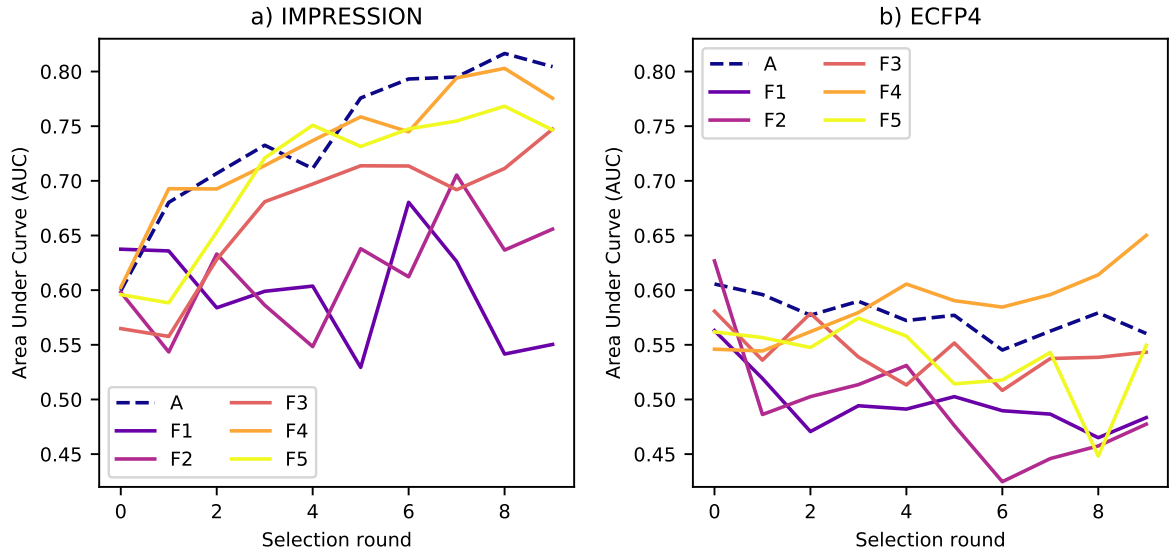

Figure 6.12: AUC at each selection round for the IMPRESSION (a) and ECFP4 (b) models, for each selection scheme, for the CDK2 dataset. Select schemes: F1(low), F2(high), F3(range), F4(distribution), F5(inverse)

CHAPTER 7

SUMMARY AND FUTURE WORK

The prediction of several NMR parameters has been investigate through the use of two architectures, and several different training and testing datasets. The efficacy of the predictions from each model was demonstrated in a prediction task similar to those in which machine learning models are commonly applied in practice. Finally the adaptation of the second generation machine learning framework to the prediction of binding affinity was explored. Across these areas of research, several common themes warrant further discussion.

7.1 Training and Testing datasets

One of the most important aspects of the work in this thesis is the creation of datasets for use in training and testing the machine learning models. The goals and considerations which informed the choices around the selection of databases to draw molecules from, the method of DFT NMR calculation used, and other factors are discussed in Chapter 2. The results in later chapters have emphasised the need to select suitable molecules for the intended application, especially when looking at the comparison between QM9 molecules and those from the CSD or ChEMBL.

The prediction accuracy of models trained using QM9 data perform exceptionally well in the prediction of molecules also taken from the QM9 dataset. In this thesis the model trained using 60,000 molecules chosen at random (QM960k, Section 2.4.5) achieved an accuracy on the 1,000 further randomly selected QM9 molecules (QM91k, Section 2.4.5) of up to 10 times better than

the accuracy of any other model presented on the same dataset (Section 4.2.2). The accuracy of the QM960k trained model presented in Chapter 4 also surpasses several of the leading NMR prediction models in the literature in terms of prediction accuracy against QM9 data [65, 67]. The issue with the QM960k trained model is the equally exceptionally poor accuracy they achieve against the molecules from the CSD (dataset 3, DT3, 2.4.2) or ChEMBL (dataset 5b, DT5b, 2.4.4), as is also demonstrated in Section 4.2.2. The molecules in DT3 and DT5b are much larger, having been drawn from X-ray crystal structures (DT3) or drug-like molecules (DT5b), than the molecules in QM9, which were constructed algorithmically to fully cover chemical space for up to 9 heavy atoms (H,C,N,O,F). There is clear evidence that models trained on QM9 in this case, and in the other work referenced above, are significantly overfitted to small molecules, and do not generalise very well. It is suggested here that in the application of machine learning models for NMR prediction, accuracy on the type of molecules found in the CSD and ChEMBL is of far more relevance to the practical application of these models, than the accuracy on the QM9 structures.

The caveat to this suggestion is found in the final analysis of the Strychnine prediction task in Chapter 5, where the QM960k trained generation 2 model demonstrates a better sensitivity to smaller structural changes than the model with the same architecture trained using datasets 4 (DT4, Section 2.4.3) and 5a (DT5a, Section 2.4.4), collectively referred to at DT45. The QM960k trained model achieves a worse mean absolute error in the prediction of the experimental NMR parameters than the DT45 trained model, supporting the argument that QM9 trained models do not generalise well to larger molecules, however clearly the QM9 training dataset presents an advantage in learning the relationship between NMR parameter values and very small structural changes. Future work in this field could attempt to replicate the dense coverage of a small chemical space provided by QM9 in larger datasets by making small modifications to the larger molecules and training using multiple similar molecules for each structure.

7.2 Model Architecture

Two distinct model architectures are presented in this thesis: the Kernel Ridge Regression (KRR) framework (Chapter 3, and the Graph Transformer Network (GTN) (Chapter 4). When the same training dataset is used (DT4), the KRR model performs similarly or better than the GTN model, suggesting that there is little advantage to the more complex GTN architecture purely in terms of the extraction of the relationship between chemical features and NMR parameters.

GTN models do however allow the use of much larger training datasets than the KRR models. This is a result of the fundamental architecture in each case, the KRR model relies on the calculation of kernel distances between all environments (feature vectors) in the training dataset, which requires all training feature vectors to be held in memory at the same time, and the kernel matrix (of size <number of training environments> by <number of training environments>) stored to make further predictions. The GTN architecture on the other hand exists as a set of learnable weight vectors, and a theoretically infinite amount of training data can be fed through the model, adjusting the values of these weights, without increasing the size of the model in memory.

The use of further training data, in this case through the addition of the DT5a dataset, significantly improves the prediction accuracy across some NMR parameters relative to the KRR model trained using a smaller dataset. The GTN architecture therefore presents a major advantage over the KRR architecture and the improvement available by using larger training sets in the GTN models is unlikely to be outperformed by modifying the KRR algorithm or obtaining different datasets of the same size. As such future work in this field is likely to focus on neural network style architectures over kernel based methods.

7.3 Estimation of uncertainty

The use of pre-prediction variance through Chapters 3 and 4 presents an important, if variable, advantage to the prediction models. The pre-prediction variance in this case is calculated by training 5 drop-out models using 80% subsets of the total training set for any given model. Predictions are then made for each environment using each drop-out model, and the variance calculated across them. The pre-prediction variance shows a consistent correlation with prediction error across the generation 1 models for most NMR parameters, however the effectiveness of the metric in improving prediction quality varies between parameters, and does not appear to function very well for the generation 2 models.

The pre-prediction variance is most useful in situations like the Strychnine prediction task described in Chapter 5, where a comparison is being made across multiple sets of similar data and the important outcome is a classification rather than the regression of specific values. In situations such as this poorly predicted environments can negatively impact the outcome of the task, and so identification of these environments at the point of prediction is advantageous. Removing environments from the comparison in such cases reduces the quality of the comparison, by

reducing the dimensionality, but this is outweighed by the advantage of removing exceptionally high errors, which will dominate the comparison if left. For this identification to be useful therefore, the correlation between prediction error and pre-prediction variance must be good for the worst predicted environments (usually much less than 10% of the available data), however in several examples in Chapter 4 this was not the case.

The most effective use of the pre-prediction variance therefore is to set a value of the variance for each model for each parameter where the associated prediction is highly likely to be inaccurate. This was done in Chapter 5 using the data presented in Chapters 3 and 4. Setting the value of the variance to a higher value limits the impact it will have on a given task, but avoids the removal of too much data in the comparison.

APPENDIX A

DATASET STRUCTURES REFERENCE

A.1 CSD and ChEMBL Structure Reference Names

Training Data CSD Reference Names

HIYHAY	KADDIE	MAMKAO	NEFHOY	PENBUH
HMCNSP	KAGZIE	MAPLIZ01	NEMZAG	PENTYN
HNOBCH	KAMROH	MAQWIM16	NEPXIR06	PEPGEW
HOCPUL	KATKIA	MATGOG	NEPXOX	PEXFUT
HOMCOD	KAVCOC	MATPEC	NESZOB	PEXLAH
HOPKUT	KAYHIE	MATVAE	NETIND01	PEZFEG01
HOQSIQ	KEDRER	MAXDUL	NEWREN	PHTHAC02
HOVFUT	KEMHAL	MECZID	NEXMOT	PHTHAC06
HOWWOH	KESTAD	MEDLEN	NIFBEJ	PIBGOX
HOZBII	KIBKAJ	MEGNES	NIFJOB	PIGROM01
HOZGAG	KIGQIA	MEHPIB	NIFRAX	PIGTAC
HURLAI	KIHXUW	MELVAA	NIHNEY	PINVOX
HXMTAM10	KIMSUU01	MENNAV	NIJKEX	PINYIW
HXOCTM	KINGUJ	MENSEE	NINWEO	PIPINE01
IBUYIQ	KIXROA	MEQFAS	NIPYAZ	PIPINE11
ICAPOR07	KIZVEV	MESYIS	NISMAD	PITQIS01
ICEMIO01	KOCKET01	METAMI02	NIVJAE	POBDER
ICOYEE	KOKLIH	MEWROX	NIVMIQ	POBSAB
IDILUD01	KONTIQ01	MEYCIC	NIYWID	POQVUO
IGENOZ	KOPBAS	MEYTUH	NOFYEM	POQWOJ
IHANAG	KOTJAE	MEYWOC	NOQBUQ	PORROE
IHOQUT	KOVFUW	MEZHEG	NOVDOR	POSJAI
IJIHOA	KOWCAC	MIDXIH	NUBLOL	POVJAL
ILAJIQ	KOXBEE	MIHZUZ	NUHFEB	POZWUW
ILIMEV02	KUGKAZ	MIMREG	NUKJIO	PUDDUP
IMUXOF	KUKCUP	MIMTAE	NUKXEX	PUQNUK
INACET03	KUQFUY	MINGAR	NUPQEU	PUQTAW
IQIDIV	KUVBEI	MIPYAL	NUQHIR	PUYTAE
IQIZAK	KUVKES	MIQNEF	NUYWIP	QACVAT
IQOROW	KUVWON01	MIVTUG	OBOWOU	QAHSOI
IQUFUX01	KUWZOS	MIWQIS	OCEHIP01	QAJBUZ
ITAFEP	KUXJIY	MIXWEX	OCOPOL	QAKJUJ
ITIKEB	KUYNOH	MNPYDO10	OGOXEP	QAKMOG
ITUVOI	LACVAM	MOBXAC	OHIWUX	QALZUA
IVAKAS	LAFHEH	MOFCOA	OJAQOH	QANQUR
IVEREH	LAVCET	MOGYIR	OLOJAB	QAPJIA
IVIDAS	LEGXUS	MOLQUB	OLOREM	QAPNAZ
IVIHAY	LEHJAM	MOYKUG	OMCHDO	QAPVOT
IXOYEA	LEMVEH	MTHPRG	OMOMOS	QATVIS
IYASUW	LEPPIF	MTYROS01	ONILAZ	QAZMIP
JABKUV	LERJAV	MUGDID	OPOZAW	QEBBUW
JAPBIO	LESCET	MUHZUM	OQUHEP	QECHEO
JAWCIW	LEZJUV	MUKBUR	ORIDAW	QEPNUW
JAXHEW	LGLUAC13	MULBIE	OTAKEB01	QEYRER
JECNUD	LIHMOG	MUNWUP	OWOHAL01	QIKJIF
JEDTIV	LILDEP	MUVCAI	OXOFMB	QIMKIG03
JEGTUN	LILJOG	MVAHIV	OZICAC	QIQYIA
JEXBOE	LIWFEC	NACGOP	PABBIF	QIRLUA
JINHET	LIYPEO	NADVIX	PADTIX	QIWGEJ
JOCDAG	LOCVEE	NAFHOR	PADXOJ	QIWMUG
JONQOU	LOKDEW	NAMZAC	PAFGUA	QOVREZ01
JOTBAV	LOMHOK	NAMZEG	PAGLEO	QUDREM
JOYGEJ	LOMNUY	NAPHTA23	PAGWIG	QUVPOO
JOZYUU	LOSMOW	NAPTYR11	PAJDOU	QUWJOJ
JUMCEB	LOVCAC	NASRUV	PAJVOO	QUYJUQ
JUNJIN	LUPGAG	NATNAA	PARHAR	RAFINO01
JUPJAH	LUQSOG	NAXRUC	PAXCEX	RALQUR
JUSQUL	LUQYIG	NAYPAF	PAYJEH	RAMZEL
KABHED	LURVUR	NAYZOD	PEFSID	RAYXEU
KACNIN	LUVPEX	NEDYEA	PEGLUL	RAYXOH

Training Data CSD Reference Names

REBXON	SUPKET	URAWEQ	WOBLAA	YEXZIM01
REDYAB	SUSYAI01	URESOB	WOBWUF01	YIDPEG
REGFER	SUVCUJ	USUZUF	WOGQEO	YIDPIM
REGKIX01	SUXCAQ	UTAGAZ	WOJGUX	YIFWAM
REGYEJ	SUXROS	UTEJIO	WOJHAG	YIGSUE
RELCUH	SUZJAZ	UTIHOV	WOKPER05	YIHHON16
REYCII	TABBOQ	UVIMES	WOLNIW	YILYOJ
REZJUC	TABNIV	UWACEB	WOZPUW	YOGSIY
RIFBUE	TACRIB02	UXICAH	WUCJOV	YOKYOO
RIGVEJ	TAHMOE	UYIREB	WUKLAP	YONBOT
RIQWIZ	TALHAR	UYUDUO	WUSQUY	YOPLIY10
RIWNEQ	TALNAV01	UZUHED	WUWMEG	YOWRAF
RIXXOM	TAMLID	VACLAM02	WUYMUZ	YOXGIB
RIZWUS	TANBEP	VAJVOU	XAKLUR	YUCQUJ
ROLVEV	TANTEK	VAPCEW	XAVMUE	YUDLAM
ROSLAO	TAPCIW	VAWJAG	XAVZOJ	YUDMOZ
RUCFAX	TARGEB	VAXLAJ	XAXHOW	YUDPAQ
RUGCED	TARGUO	VEBWEH	XAYDIK	YUFYED
RUGQOA	TATNEI	VECSAZ	XAZQOF	YUHTEA03
RUJQOE	TECQEX	VEFPIF	XAZROH	YUHTOK
RUJSAS	TEGVUW	VESHUX	XEBYUA	YUNTOR
RURRAY	TEJREG	VEXCUW	XEDNAX	YUNYIR
RUVSAC	TEKSOR	VEZNOF	XEDTEG	YUQCUJ
RUWJAU	TELKAZ	VIBZUB	XEHTUZ	YUQMED
RUWMAX	TENMIK	VIDFEV	XEMDAX	ZAJHOH
RUWQIK	TEPHME02	VIGWOY	XENLAE	ZAJVAK
RUZXIU	TEVLIQ	VIGXAK	XETMAL	ZETHUD
SADJEM	TICBUD	VIHBIZ	XEVCEH	ZEWPUM
SADXOL	TIHBAO	VOFSEP	XEWNES	ZIFKEG
SAGQUO	TIMHED	VOKXOJ	XEXQOH01	ZILQOA01
SAHCOV	TIQNIQ	VOLKIS	XEYRIE	ZIYSIL
SAHZAF	TIQWOG	VUDKIP	XEZYIK	ZODXEV
SAKJUM	TIXPOF	VUFGEI01	XIJFEB	ZOFCUU
SANWEJ	TMXSTQ10	VUFSEU	XIMCOL	ZOLBUX
SAPHAU	TOHVIW	VUFWAV	XIMJAE	ZONYUY
SARJED	TOPROG	VUKFOY	XINJIN	ZOZTOX
SAWHUV	TOPSEW	VUNFUF	XISHOY	ZUPGIA10
SAWJUX	TOVSUS02	VUPHIZ	XIVVAA	ZUPGUM
SAYTAN	TPHETY01	VUTBUI	XIWREA02	ZUPHAT
SAYWOG	TUCJEI	VUTNAB	XOBGAY	ZUQVOY
SAZLAH	TUCNUC	VUZQOX	XOGWAR	ZZZLUK05
SECTIF	TUJJEP	WABTAU	XOGXEX	ZZZMBS02
SEDMOD	TULDAH	WACZUX	XOMJIS	
SEHNAW01	TUNCOW	WADGEO01	XOWDAQ	
SEJWOT	TUNTUT	WADQID	XUHPIB	
SELKEB	TUSQUU	WAGBEO	XUPYIR	
SEQREN	TUWCEU	WALNEC	XUVSUE	
SIQQEP	UBEBAG	WANVEP	XUYZIC	
SITCUU	UBUPEM	WAQNUZ01	YAGJEX	
SIVJOY	UCOMOO	WAZMAL	YAMHID01	
SIWDEH	UCOQAE	WECXUZ	YAPBUO	
SIYYUU	UCUZOJ	WESVIZ	YAPZEU	
SOPLEO	UDEHER	WEWTUP	YAQWAR	
SOXHAQ	UFAGOY	WIBWIN	YARDUQ	
SUCACB12	UHADOX	WIBXUA	YAWWAU01	
SUCANH12	UKUTUP	WIFZOC	YAYDIN	
SUCROS47	UPACUK	WIPHAG	YEJPAG	
SUCTAN	UPADOG	WIVYUV	YEJZES	
SUFGAB	UQIMUE	WIYDUF	YEKVEQ	
SUHYIE	URAHIF	WIZZAI	YENLAF	

Testing Dataset CSD Reference Names

AFIQUC	COYBOJ	FAJDEC	JIPCUG10	NEZFON	RIMHEC	VEQMUA	ZATDOP
AHATEK	CUTCUQ	FELDOR	JOQTUE	NIQTAJ	RIZBAF	VEZCUY	ZAYPOE
AHOWOL	CXMTUN	FEPTID	JULGOO	NORFUW	ROGRIQ	VIDDAO	ZEMNAG
AHOXOL	DAFTAF	FEZLUT	KAHJEK	NUKSAO	ROHJED	VIDMAX02	ZIGBAS
AJIXUM	DASNIV	FIHLEO	KEMFIS	NURZOP	ROJHOP	VILPUB	ZIKQIT
AKUBIT	DENXUP02	FOSLEG	KOFKAR	OCATOC	ROJXOD	VOCHUR	ZIWMOJ
ALOSEZ	DILDUZ	FUPWES	KOGWUZ	OCAWOF	RUCNOU	VOGDIE	ZOFNUD
AMEXOH	DILKIT	GADVAJ	KOJTOT	OCIPAR	RUKTAU	VONNOB	ZOSVEI
ANAHII	DITZOX	GASXON	KOTMUB	OFEVOL	RULDAF	VOXNOL	ZOXYOA
ANOSAY	DIWWEN	GAWFEQ	KUJZIY	OGIMIC	RULHOX	VUDDUV	ZZZBPY10
APODUG	DOHPEV	GIDHUW	KUTKAL	OJICUF	SAJCAJ	VUHZEE	ZZZFFY01
APUPIK	DOLBIR10	GIXKOP	KUZJIA	OMABEK	SATPEI02	WAWQUH	
AQAGII	DOMNEY	GIZRUE	KUZQIG	OMSTER01	SATPUZ	WECZEJ	
AQEYAW	DORKOK	GUCJUK	LADNEL	ONBZAM	SAVREN	WEVVEZ	
AROKUN	DOVWAM	GUFYOX	LAVSIL	OPIZAQ	SAWVET	WIFQEI	
ARONOM	DUTKOU	GUJGEX	LEVSIO	OXAROV	SAZFOO	WIHBEW	
AWAVEZ	DUZLUF	GUTZOM	LILDEP	OXUJUN	SEBVAW	WIQZOL	
AXADAF	EABZBU	HABNED	LIXQEO	PACWAU	SEFNOG	WOKJOV	
AXAWIG	EBAXOW	HAMTIZ	LIZHEJ	PANLEZ10	SENKUR	WUCVIB	
AXOSOW03	EBOVEX	HECNOS	LOPLUZ	PEDHAJ	SEYCUU	XAQTUF	
AYUNEO	ECODUV	HIMSUS	LUDZIT	PETRAH	SIGSAD	XASHUW	
AZIDES	EDAXOW	HISNII	LUQDOS	PEXPEN	SIHCES	XAZYIG	
BAJCIY03	EDIZUM	HIWYIV	MALSOH	PIHBOZ	SIHZAM	XIMGAB	
BAPPUF	EFIBAX	HIZHOP	MAQWIM23	PIJREF	SOGCUN	XINHIL	
BAQNEM	EKAHOP	HODKEQ	MATQOO	PILFIB	SORFIQ	XIYTIJ	
BASNOZ	EKAWAQ	HODLOC	MEHLER	POHCAS	SUHFEH	XIZVAD	
BAYPAT	EKOGAO	HOMKIF	MEHNAP	POLJEF	SUKNIW02	XOFFEF	
BEDLEB01	ELAWIX	HOMZUG	MEJDOU	PRMDIN05	SUWKEC	XOHMAI	
BEGDIB01	EMIPUM	HONKEC	MEJQEY	PUMQEV	TAJSOM	XOWJUP	
BEHWER	EMISUQ	HUDHEU	MELAMI05	PUNFAH	TAVJAD	XUJKUK	
BERSOG	EMODUG	HUVWOL	MENDAL01	PUPBAD01	TEMKAZ	XULNOI	
BIKNUE	ENIMET	HUYYOP	MESQOR	PUWNIG	THYDIN05	XUVBAT	
BIXQEF	EPHEDR01	IDUJEW	MISDAT	PYAZAC	TOPRIB	YAZDEI	
BOLGOZ	EVIHUM02	IJEZUS	MOBNUM	QAKDAJ	TOPXUT	YEGGIA	
BUGQUQ	EVINII	INAVIC	MOSLAI	QAMKEW	UBUXOG	YEHWUD	
BUMNOM	EVIQEF	IQIKOI	MOTNUF	QECNAP	UCANIV	YERTIZ01	
BUZJIR	EWOBIB	IQIZEO	MUJGEE	QEPRIO	UJUKIT	YIDTIQ	
BZAMID08	EXEWEJ	IQUBZA	MUTWON	QEXKUA	UMUKUJ	YIMPOB	
BZTROP11	EXEYUD	IQULUC	NAJLUF	QOMVUK	UNAMOL	YIXPUR	
CBMZPN21	EXUVUP	IROZIY	NANJIW	QUFCEZ	UNURIF	YOCWUK	
CIKSAQ	EYASAZ	ITINEG	NASZAJ	QUFJUY	UNUVEF	YODPAJ	
CINCHO10	EZISUC	ITIREI	NBZOAC11	QUWFIZ	UQOLIW	YOFTOE	
COCYAW	FACZIU	IVEZAK	NCUBEB10	RAKTOO	UWOCAM	YOWYOY02	
COLBAG	FAHLAB	IYASUW	NEQPEG	RICTIG	VANFEV	YOXRIO	
COWPUZ	FAHXUH	JESHIZ	NEVDOH	RIHFIY	VASLOR	YUNYUC	

A.2 ChEMBL Structures

Training Dataset ChEMBL Reference Names

CHEMBL1075643	CHEMBL1075666	CHEMBL1075668	CHEMBL1075679	CHEMBL1075723
CHEMBL1075761	CHEMBL1075834	CHEMBL1075839	CHEMBL1075878	CHEMBL1075896
CHEMBL1075993	CHEMBL1076030	CHEMBL1076072	CHEMBL1076074	CHEMBL1076077
CHEMBL1076081	CHEMBL1076088	CHEMBL1076098	CHEMBL1076101	CHEMBL1076110
CHEMBL1076128	CHEMBL1076131	CHEMBL1076212	CHEMBL1076214	CHEMBL1076217
CHEMBL1076221	CHEMBL1076265	CHEMBL1076268	CHEMBL1076286	CHEMBL1076290
CHEMBL1076294	CHEMBL1076357	CHEMBL1076408	CHEMBL1076419	CHEMBL1076477
CHEMBL1076491	CHEMBL1076555	CHEMBL1076577	CHEMBL1076587	CHEMBL1076637
CHEMBL1076668	CHEMBL1076712	CHEMBL1076741	CHEMBL1076755	CHEMBL1076769
CHEMBL1076775	CHEMBL1076790	CHEMBL1076792	CHEMBL1076799	CHEMBL1076803
CHEMBL1076814	CHEMBL1076905	CHEMBL1076906	CHEMBL1076925	CHEMBL1076999
CHEMBL1077002	CHEMBL1077081	CHEMBL1077125	CHEMBL1077257	CHEMBL1077261
CHEMBL1077272	CHEMBL1077276	CHEMBL1077282	CHEMBL1077287	CHEMBL1077333
CHEMBL1077336	CHEMBL1077343	CHEMBL1077360	CHEMBL1077368	CHEMBL1077381
CHEMBL1077407	CHEMBL1077476	CHEMBL1077485	CHEMBL1077528	CHEMBL1077609
CHEMBL1077648	CHEMBL1077686	CHEMBL1077805	CHEMBL1077878	CHEMBL1078134
CHEMBL1078223	CHEMBL1078396	CHEMBL1078513	CHEMBL1078741	CHEMBL1078833
CHEMBL1079031	CHEMBL1079125	CHEMBL1079127	CHEMBL1079305	CHEMBL1079618
CHEMBL1080664	CHEMBL1081781	CHEMBL1082274	CHEMBL1082278	CHEMBL1082437
CHEMBL1082532	CHEMBL1082636	CHEMBL1082898	CHEMBL1082938	CHEMBL1083224
CHEMBL1083240	CHEMBL1083248	CHEMBL1083268	CHEMBL1083537	CHEMBL1083541
CHEMBL1083543	CHEMBL1083581	CHEMBL1083593	CHEMBL1083863	CHEMBL1083881
CHEMBL1084042	CHEMBL1084425	CHEMBL1084477	CHEMBL1084482	CHEMBL1084510
CHEMBL1084974	CHEMBL1085225	CHEMBL1085285	CHEMBL1085658	CHEMBL1085713
CHEMBL1085980	CHEMBL1086063	CHEMBL1086190	CHEMBL1086220	CHEMBL1086312
CHEMBL1086439	CHEMBL1086459	CHEMBL1086527	CHEMBL1086694	CHEMBL1086957
CHEMBL1087095	CHEMBL1088145	CHEMBL1088172	CHEMBL1088204	CHEMBL1088217
CHEMBL1088309	CHEMBL1088337	CHEMBL1088344	CHEMBL1088591	CHEMBL1088641
CHEMBL1088782	CHEMBL1088958	CHEMBL1088966	CHEMBL1088978	CHEMBL1088980
CHEMBL1088996	CHEMBL1089152	CHEMBL1089238	CHEMBL1089317	CHEMBL1089574
CHEMBL1089576	CHEMBL1089583	CHEMBL1089585	CHEMBL1089592	CHEMBL1089598
CHEMBL1089667	CHEMBL1089674	CHEMBL1089958	CHEMBL1089969	CHEMBL1090006
CHEMBL1090111	CHEMBL1090248	CHEMBL1090356	CHEMBL1090597	CHEMBL1090649
CHEMBL1090651	CHEMBL1090660	CHEMBL1090668	CHEMBL1090695	CHEMBL1091275
CHEMBL1091285	CHEMBL1091319	CHEMBL1091334	CHEMBL1091340	CHEMBL1091359
CHEMBL1091391	CHEMBL1091402	CHEMBL1091420	CHEMBL1091501	CHEMBL1091561
CHEMBL1091683	CHEMBL1091740	CHEMBL1091753	CHEMBL1091763	CHEMBL1092024
CHEMBL1092041	CHEMBL1092059	CHEMBL1092090	CHEMBL1092134	CHEMBL1092443
CHEMBL1092453	CHEMBL1093288	CHEMBL1093316	CHEMBL1093521	CHEMBL1093605
CHEMBL1093624	CHEMBL1093640	CHEMBL1093669	CHEMBL1093690	CHEMBL1093989
CHEMBL1094237	CHEMBL1094324	CHEMBL1094354	CHEMBL1094608	CHEMBL1094634
CHEMBL1094663	CHEMBL1094671	CHEMBL1094701	CHEMBL1094708	CHEMBL1094711
CHEMBL1094870	CHEMBL1094912	CHEMBL1094998	CHEMBL1095005	CHEMBL1095033
CHEMBL1095192	CHEMBL1095336	CHEMBL1095378	CHEMBL1095653	CHEMBL1095821
CHEMBL1095900	CHEMBL1096263	CHEMBL1096298	CHEMBL1096328	CHEMBL1096330

Training Dataset ChEMBL Reference Names

CHEMBL1096337	CHEMBL1096579	CHEMBL1096583	CHEMBL1096640	CHEMBL1096673
CHEMBL1096822	CHEMBL1096890	CHEMBL1097210	CHEMBL1097293	CHEMBL1097360
CHEMBL1097634	CHEMBL1097664	CHEMBL1097899	CHEMBL1097906	CHEMBL1097916
CHEMBL1097992	CHEMBL1098259	CHEMBL1098265	CHEMBL1098309	CHEMBL1098866
CHEMBL1099227	CHEMBL1099299	CHEMBL1099329	CHEMBL1159434	CHEMBL1159451
CHEMBL1159458	CHEMBL1159483	CHEMBL1159511	CHEMBL1159526	CHEMBL1159610
CHEMBL1159637	CHEMBL1159641	CHEMBL1160112	CHEMBL1160274	CHEMBL1160275
CHEMBL1160322	CHEMBL1160429	CHEMBL1160667	CHEMBL1160680	CHEMBL1160711
CHEMBL1160741	CHEMBL1160771	CHEMBL1161178	CHEMBL1161220	CHEMBL1161558
CHEMBL1161567	CHEMBL1161839	CHEMBL1161852	CHEMBL1161924	CHEMBL1161944
CHEMBL1162038	CHEMBL1162080	CHEMBL1162102	CHEMBL1162104	CHEMBL1162169
CHEMBL1162174	CHEMBL1162204	CHEMBL1162209	CHEMBL1162315	CHEMBL1162403
CHEMBL1162462	CHEMBL1163060	CHEMBL1163069	CHEMBL1163095	CHEMBL1163139
CHEMBL1163140	CHEMBL1163145	CHEMBL1163148	CHEMBL1163149	CHEMBL1163153
CHEMBL1163186	CHEMBL1163197	CHEMBL1163200	CHEMBL1163203	CHEMBL1163213
CHEMBL1163214	CHEMBL1163216	CHEMBL1163233	CHEMBL1163234	CHEMBL1163239
CHEMBL1163275	CHEMBL1163390	CHEMBL1163392	CHEMBL1163408	CHEMBL1163427
CHEMBL1163473	CHEMBL1164909	CHEMBL1164921	CHEMBL1164922	CHEMBL1164952
CHEMBL1165018	CHEMBL1165220	CHEMBL1165401	CHEMBL1165404	CHEMBL1165423
CHEMBL1165499	CHEMBL1165530	CHEMBL1165739	CHEMBL1165741	CHEMBL22
CHEMBL1165801	CHEMBL1169603	CHEMBL1170005	CHEMBL1170046	CHEMBL1170054
CHEMBL1170234	CHEMBL1170644	CHEMBL1170659	CHEMBL1170662	CHEMBL1170828
CHEMBL1171068	CHEMBL1171111	CHEMBL1171133	CHEMBL1171139	CHEMBL1171146
CHEMBL1171471	CHEMBL1171525	CHEMBL1171643	CHEMBL1171705	CHEMBL1171794
CHEMBL1171953	CHEMBL1172056	CHEMBL1172204	CHEMBL1172222	CHEMBL1172235
CHEMBL1172569	CHEMBL1172581	CHEMBL1173577	CHEMBL1173726	CHEMBL1179555
CHEMBL1179567	CHEMBL1179704	CHEMBL1180192	CHEMBL1180347	CHEMBL1181431
CHEMBL1181953	CHEMBL1181959	CHEMBL1183203	CHEMBL1183536	CHEMBL1184248
CHEMBL1184883	CHEMBL1184894	CHEMBL1185211	CHEMBL1185294	CHEMBL1185408
CHEMBL1185871	CHEMBL1186057	CHEMBL1186068	CHEMBL1186096	CHEMBL1186159
CHEMBL1186168	CHEMBL1186195	CHEMBL1186245	CHEMBL1186290	CHEMBL1186537
CHEMBL1186606	CHEMBL1187270	CHEMBL1187588	CHEMBL1187616	CHEMBL1187750
CHEMBL1187868	CHEMBL1187911	CHEMBL1187952	CHEMBL1187970	CHEMBL1188172
CHEMBL1188238	CHEMBL1188474	CHEMBL1188581	CHEMBL1189068	CHEMBL1189598
CHEMBL1189963	CHEMBL1190414	CHEMBL1191207	CHEMBL1195116	CHEMBL1195488
CHEMBL1195988	CHEMBL1196830	CHEMBL1198796	CHEMBL26565	CHEMBL1199124
CHEMBL1199159	CHEMBL1199171	CHEMBL1199236	CHEMBL1199618	CHEMBL1199633
CHEMBL1199671	CHEMBL1199724	CHEMBL1200037	CHEMBL1200285	CHEMBL1200656
CHEMBL1200714	CHEMBL1201295	CHEMBL1201356	CHEMBL1201843	CHEMBL1203155
CHEMBL1204461	CHEMBL1204470	CHEMBL1204471	CHEMBL1204670	CHEMBL1205279
CHEMBL1205372	CHEMBL1205595	CHEMBL1205641	CHEMBL1205646	CHEMBL1205733
CHEMBL1205831	CHEMBL1206474	CHEMBL1206530	CHEMBL1206588	CHEMBL1206631
CHEMBL1206762	CHEMBL1207158	CHEMBL1207360	CHEMBL1207397	CHEMBL1207533
CHEMBL1207835	CHEMBL1207867	CHEMBL1207937	CHEMBL1207940	CHEMBL1207979
CHEMBL1207983	CHEMBL1207998	CHEMBL1208034	CHEMBL1208314	CHEMBL1208419

Training Dataset ChEMBL Reference Names

CHEMBL1209290	CHEMBL1209501	CHEMBL1209731	CHEMBL1209733	CHEMBL1209811
CHEMBL1209991	CHEMBL1210111	CHEMBL1210341	CHEMBL1210347	CHEMBL1210356
CHEMBL1210364	CHEMBL1210578	CHEMBL1210789	CHEMBL1210852	CHEMBL1213094
CHEMBL1213147	CHEMBL1213185	CHEMBL1213309	CHEMBL1213497	CHEMBL1213498
CHEMBL1213533	CHEMBL1213701	CHEMBL1213767	CHEMBL1213854	CHEMBL1214058
CHEMBL1214092	CHEMBL1214096	CHEMBL1214259	CHEMBL1214400	CHEMBL1214443
CHEMBL1214455	CHEMBL1214513	CHEMBL1214515	CHEMBL1214531	CHEMBL1214569
CHEMBL1214612	CHEMBL1214724	CHEMBL1214847	CHEMBL1214854	CHEMBL1214894
CHEMBL1215356	CHEMBL1215485	CHEMBL1215684	CHEMBL1215831	CHEMBL1215838
CHEMBL1221917	CHEMBL1222016	CHEMBL1222580	CHEMBL1222581	CHEMBL1222611
CHEMBL1222706	CHEMBL1222790	CHEMBL1222867	CHEMBL1223009	CHEMBL1223111
CHEMBL1223703	CHEMBL1223923	CHEMBL1223924	CHEMBL1223976	CHEMBL1223978
CHEMBL1223980	CHEMBL1224265	CHEMBL1224292	CHEMBL1224447	CHEMBL1224732
CHEMBL1224734	CHEMBL1224737	CHEMBL1224854	CHEMBL1224864	CHEMBL1229215
CHEMBL1237043	CHEMBL146675	CHEMBL153086	CHEMBL153534	CHEMBL153717
CHEMBL153812	CHEMBL154192	CHEMBL154209	CHEMBL154228	CHEMBL154288
CHEMBL154314	CHEMBL154341	CHEMBL154360	CHEMBL154392	CHEMBL154414
CHEMBL154463	CHEMBL154556	CHEMBL154609	CHEMBL154714	CHEMBL154771
CHEMBL154789	CHEMBL154818	CHEMBL154832	CHEMBL154837	CHEMBL154940
CHEMBL154997	CHEMBL155331	CHEMBL155374	CHEMBL155439	CHEMBL155451
CHEMBL155459	CHEMBL155478	CHEMBL155537	CHEMBL155636	CHEMBL155780
CHEMBL155896	CHEMBL155979	CHEMBL156224	CHEMBL156363	CHEMBL156555
CHEMBL156735	CHEMBL157042	CHEMBL157232	CHEMBL157236	CHEMBL157367
CHEMBL157397	CHEMBL157446	CHEMBL157450	CHEMBL404	CHEMBL157834
CHEMBL158175	CHEMBL158217	CHEMBL158960	CHEMBL159567	CHEMBL159668
CHEMBL160210	CHEMBL160738	CHEMBL161573	CHEMBL161838	CHEMBL1620719
CHEMBL162280	CHEMBL162786	CHEMBL163579	CHEMBL163819	CHEMBL163871
CHEMBL163965	CHEMBL164375	CHEMBL164518	CHEMBL164817	CHEMBL165039
CHEMBL165476	CHEMBL165532	CHEMBL165949	CHEMBL167929	CHEMBL167990
CHEMBL170493	CHEMBL177285	CHEMBL1782891	CHEMBL180233	CHEMBL180716
CHEMBL182355	CHEMBL196395	CHEMBL199468	CHEMBL2094221	CHEMBL217730
CHEMBL2369103	CHEMBL258525	CHEMBL260909	CHEMBL260970	CHEMBL261102
CHEMBL261894	CHEMBL262244	CHEMBL262267	CHEMBL262399	CHEMBL262664
CHEMBL262819	CHEMBL262924	CHEMBL263193	CHEMBL263329	CHEMBL263555
CHEMBL263614	CHEMBL263810	CHEMBL263956	CHEMBL264055	CHEMBL264137
CHEMBL264472	CHEMBL264812	CHEMBL264899	CHEMBL265024	CHEMBL265130
CHEMBL265174	CHEMBL265216	CHEMBL265362	CHEMBL265564	CHEMBL265732
CHEMBL265763	CHEMBL265830	CHEMBL265900	CHEMBL266873	CHEMBL266902
CHEMBL266960	CHEMBL267740	CHEMBL267832	CHEMBL268086	CHEMBL268339
CHEMBL269191	CHEMBL269289	CHEMBL316793	CHEMBL3309261	CHEMBL3309266
CHEMBL3309270	CHEMBL3309273	CHEMBL3309279	CHEMBL3309284	CHEMBL3309287
CHEMBL3309295	CHEMBL3309302	CHEMBL3309304	CHEMBL3309323	CHEMBL3309324
CHEMBL3309336	CHEMBL3309345	CHEMBL3309352	CHEMBL3309375	CHEMBL3309383
CHEMBL3309396	CHEMBL3309423	CHEMBL3309436	CHEMBL3309440	CHEMBL3309451
CHEMBL3309458	CHEMBL3309460	CHEMBL3309461	CHEMBL3309471	CHEMBL3309480

Training Dataset ChEMBL Reference Names

CHEMBL3309490	CHEMBL3309491	CHEMBL3309496	CHEMBL3309504	CHEMBL3309505
CHEMBL3309508	CHEMBL3309531	CHEMBL3309538	CHEMBL3309541	CHEMBL3309555
CHEMBL3309561	CHEMBL3309654	CHEMBL3309661	CHEMBL3309763	CHEMBL3309767
CHEMBL3309775	CHEMBL3309860	CHEMBL3309867	CHEMBL3309871	CHEMBL3347285
CHEMBL3347288	CHEMBL3347298	CHEMBL3347310	CHEMBL3347320	CHEMBL3348817
CHEMBL360291	CHEMBL383917	CHEMBL385384	CHEMBL386654	CHEMBL403325
CHEMBL405225	CHEMBL405398	CHEMBL405416	CHEMBL405667	CHEMBL406648
CHEMBL408	CHEMBL409208	CHEMBL409297	CHEMBL409812	CHEMBL410788
CHEMBL410790	CHEMBL4115984	CHEMBL4115992	CHEMBL4116000	CHEMBL4116001
CHEMBL4116005	CHEMBL411601	CHEMBL4116085	CHEMBL4116089	CHEMBL4116090
CHEMBL4116091	CHEMBL4116092	CHEMBL4116093	CHEMBL4116095	CHEMBL4116096
CHEMBL4116099	CHEMBL4116100	CHEMBL4116102	CHEMBL4116104	CHEMBL4116107
CHEMBL4116109	CHEMBL4116111	CHEMBL4116112	CHEMBL4116114	CHEMBL4116118
CHEMBL4116119	CHEMBL4116120	CHEMBL4116121	CHEMBL4116122	CHEMBL4116123
CHEMBL4116124	CHEMBL4116129	CHEMBL4116130	CHEMBL4116131	CHEMBL4116132
CHEMBL4116133	CHEMBL4116137	CHEMBL4116140	CHEMBL4116141	CHEMBL4116143
CHEMBL4116146	CHEMBL4116152	CHEMBL4116153	CHEMBL4116155	CHEMBL4116157
CHEMBL4116158	CHEMBL4116162	CHEMBL4116164	CHEMBL4116168	CHEMBL412007
CHEMBL413611	CHEMBL414319	CHEMBL414339	CHEMBL414380	CHEMBL414390
CHEMBL414958	CHEMBL415391	CHEMBL415423	CHEMBL415615	CHEMBL438024
CHEMBL438132	CHEMBL438139	CHEMBL438301	CHEMBL438327	CHEMBL438807
CHEMBL438839	CHEMBL439138	CHEMBL439267	CHEMBL439400	CHEMBL439520
CHEMBL439542	CHEMBL441131	CHEMBL441569	CHEMBL441620	CHEMBL441948
CHEMBL442595	CHEMBL442894	CHEMBL443462	CHEMBL443597	CHEMBL443598
CHEMBL443602	CHEMBL443682	CHEMBL443886	CHEMBL444024	CHEMBL444145
CHEMBL444231	CHEMBL444233	CHEMBL444368	CHEMBL444432	CHEMBL444434
CHEMBL444506	CHEMBL444522	CHEMBL444736	CHEMBL444924	CHEMBL444987
CHEMBL445000	CHEMBL445025	CHEMBL445172	CHEMBL445258	CHEMBL460994
CHEMBL462274	CHEMBL468771	CHEMBL470546	CHEMBL476536	CHEMBL490449
CHEMBL494682	CHEMBL498846	CHEMBL498849	CHEMBL498861	CHEMBL498871
CHEMBL498905	CHEMBL498921	CHEMBL498936	CHEMBL498962	CHEMBL498970
CHEMBL498979	CHEMBL498986	CHEMBL499101	CHEMBL499517	CHEMBL499520
CHEMBL499523	CHEMBL499529	CHEMBL499532	CHEMBL499543	CHEMBL499563
CHEMBL499568	CHEMBL499809	CHEMBL499823	CHEMBL499846	CHEMBL499943
CHEMBL499959	CHEMBL499961	CHEMBL500007	CHEMBL500021	CHEMBL500054
CHEMBL500056	CHEMBL500079	CHEMBL500085	CHEMBL500088	CHEMBL500090
CHEMBL500097	CHEMBL500105	CHEMBL500109	CHEMBL500111	CHEMBL500179
CHEMBL500195	CHEMBL500202	CHEMBL500206	CHEMBL500207	CHEMBL500246
CHEMBL500249	CHEMBL500257	CHEMBL500286	CHEMBL500308	CHEMBL500346
CHEMBL500357	CHEMBL500370	CHEMBL500372	CHEMBL500377	CHEMBL500437
CHEMBL500450	CHEMBL500467	CHEMBL500474	CHEMBL500495	CHEMBL500519
CHEMBL500527	CHEMBL500540	CHEMBL500548	CHEMBL500627	CHEMBL500704
CHEMBL500709	CHEMBL500730	CHEMBL500734	CHEMBL500738	CHEMBL500758
CHEMBL500769	CHEMBL500790	CHEMBL500805	CHEMBL500809	CHEMBL501070
CHEMBL501130	CHEMBL5	CHEMBL501132	CHEMBL501136	CHEMBL501251

Training Dataset ChEMBL Reference Names

CHEMBL501256	CHEMBL501274	CHEMBL501276	CHEMBL501280	CHEMBL501315
CHEMBL501326	CHEMBL501328	CHEMBL501493	CHEMBL501507	CHEMBL501508
CHEMBL501517	CHEMBL501542	CHEMBL501587	CHEMBL501589	CHEMBL501593
CHEMBL501597	CHEMBL501610	CHEMBL501618	CHEMBL501647	CHEMBL501650
CHEMBL501665	CHEMBL501671	CHEMBL501674	CHEMBL501682	CHEMBL501687
CHEMBL501689	CHEMBL501691	CHEMBL501694	CHEMBL501758	CHEMBL501763
CHEMBL501766	CHEMBL501801	CHEMBL501828	CHEMBL501858	CHEMBL501871
CHEMBL501877	CHEMBL501918	CHEMBL501922	CHEMBL501923	CHEMBL501925
CHEMBL501940	CHEMBL501957	CHEMBL501958	CHEMBL501963	CHEMBL501964
CHEMBL501968	CHEMBL502048	CHEMBL502052	CHEMBL502053	CHEMBL502166
CHEMBL502181	CHEMBL502183	CHEMBL502187	CHEMBL502192	CHEMBL502200
CHEMBL502201	CHEMBL502203	CHEMBL502208	CHEMBL502210	CHEMBL502212
CHEMBL502243	CHEMBL502296	CHEMBL502307	CHEMBL502312	CHEMBL502354
CHEMBL502357	CHEMBL502411	CHEMBL502423	CHEMBL502429	CHEMBL502438
CHEMBL502441	CHEMBL502456	CHEMBL502486	CHEMBL502490	CHEMBL502494
CHEMBL502613	CHEMBL502624	CHEMBL502626	CHEMBL502640	CHEMBL502652
CHEMBL502653	CHEMBL502655	CHEMBL502658	CHEMBL502660	CHEMBL502983
CHEMBL6208	CHEMBL502990	CHEMBL503028	CHEMBL503046	CHEMBL503058
CHEMBL503221	CHEMBL503232	CHEMBL503237	CHEMBL503252	CHEMBL503258
CHEMBL503266	CHEMBL503289	CHEMBL503304	CHEMBL503330	CHEMBL503340
CHEMBL503345	CHEMBL503350	CHEMBL503357	CHEMBL503391	CHEMBL503411
CHEMBL503422	CHEMBL503430	CHEMBL503434	CHEMBL503449	CHEMBL503451
CHEMBL503452	CHEMBL503459	CHEMBL503469	CHEMBL503474	CHEMBL503501
CHEMBL503508	CHEMBL503539	CHEMBL503548	CHEMBL503549	CHEMBL503551
CHEMBL503559	CHEMBL503606	CHEMBL503616	CHEMBL503619	CHEMBL503623
CHEMBL503634	CHEMBL503643	CHEMBL503670	CHEMBL503685	CHEMBL503698
CHEMBL503702	CHEMBL503713	CHEMBL503725	CHEMBL503802	CHEMBL503814
CHEMBL503818	CHEMBL503828	CHEMBL503838	CHEMBL503845	CHEMBL503870
CHEMBL503973	CHEMBL503977	CHEMBL503986	CHEMBL503991	CHEMBL504035
CHEMBL504045	CHEMBL504052	CHEMBL504055	CHEMBL504067	CHEMBL504086
CHEMBL504142	CHEMBL504150	CHEMBL504162	CHEMBL504164	CHEMBL504168
CHEMBL504184	CHEMBL504192	CHEMBL504216	CHEMBL6209	CHEMBL504249
CHEMBL504250	CHEMBL504252	CHEMBL504254	CHEMBL504333	CHEMBL504349
CHEMBL504357	CHEMBL504363	CHEMBL504372	CHEMBL504374	CHEMBL504397
CHEMBL504398	CHEMBL504425	CHEMBL504429	CHEMBL504440	CHEMBL504451
CHEMBL504800	CHEMBL505023	CHEMBL505033	CHEMBL505036	CHEMBL505038
CHEMBL505126	CHEMBL505127	CHEMBL505139	CHEMBL505140	CHEMBL505397
CHEMBL505403	CHEMBL505405	CHEMBL505806	CHEMBL505819	CHEMBL505923
CHEMBL505924	CHEMBL505930	CHEMBL505943	CHEMBL506057	CHEMBL506068
CHEMBL506069	CHEMBL506071	CHEMBL506077	CHEMBL506403	CHEMBL506407
CHEMBL506415	CHEMBL506560	CHEMBL506637	CHEMBL506640	CHEMBL506649
CHEMBL506660	CHEMBL507116	CHEMBL507127	CHEMBL507207	CHEMBL507223
CHEMBL507226	CHEMBL507303	CHEMBL507307	CHEMBL507538	CHEMBL507542
CHEMBL507700	CHEMBL507764	CHEMBL507884	CHEMBL507896	CHEMBL507897
CHEMBL508095	CHEMBL508099	CHEMBL508100	CHEMBL508159	CHEMBL508166

Training Dataset ChEMBL Reference Names

CHEMBL508212	CHEMBL508213	CHEMBL508215	CHEMBL6210	CHEMBL508224
CHEMBL509195	CHEMBL509250	CHEMBL509252	CHEMBL509348	CHEMBL520435
CHEMBL522887	CHEMBL523218	CHEMBL524158	CHEMBL524317	CHEMBL525153
CHEMBL525206	CHEMBL525228	CHEMBL525413	CHEMBL525418	CHEMBL525443
CHEMBL525627	CHEMBL525672	CHEMBL525765	CHEMBL525924	CHEMBL526157
CHEMBL526183	CHEMBL526313	CHEMBL526346	CHEMBL526377	CHEMBL526543
CHEMBL526702	CHEMBL526882	CHEMBL527552	CHEMBL527864	CHEMBL529058
CHEMBL529450	CHEMBL530034	CHEMBL530730	CHEMBL531135	CHEMBL532261
CHEMBL533122	CHEMBL534251	CHEMBL534280	CHEMBL538188	CHEMBL538195
CHEMBL538380	CHEMBL538385	CHEMBL538453	CHEMBL538457	CHEMBL538460
CHEMBL538627	CHEMBL538670	CHEMBL538686	CHEMBL538692	CHEMBL538895
CHEMBL538901	CHEMBL539151	CHEMBL539179	CHEMBL539185	CHEMBL539191
CHEMBL539203	CHEMBL539383	CHEMBL539434	CHEMBL539442	CHEMBL539921
CHEMBL540153	CHEMBL540167	CHEMBL540195	CHEMBL540214	CHEMBL540215
CHEMBL540227	CHEMBL540242	CHEMBL540427	CHEMBL540674	CHEMBL540696
CHEMBL540947	CHEMBL540973	CHEMBL541005	CHEMBL541013	CHEMBL541435
CHEMBL541483	CHEMBL541686	CHEMBL541693	CHEMBL6211	CHEMBL541740
CHEMBL541754	CHEMBL541947	CHEMBL541967	CHEMBL541969	CHEMBL542011
CHEMBL546744	CHEMBL547008	CHEMBL547407	CHEMBL548124	CHEMBL548334
CHEMBL549249	CHEMBL550693	CHEMBL552187	CHEMBL552528	CHEMBL552766
CHEMBL552993	CHEMBL553669	CHEMBL553689	CHEMBL553887	CHEMBL553904
CHEMBL553918	CHEMBL553933	CHEMBL554494	CHEMBL554551	CHEMBL554586
CHEMBL555367	CHEMBL555454	CHEMBL555469	CHEMBL555916	CHEMBL556060
CHEMBL557754	CHEMBL557954	CHEMBL557964	CHEMBL558162	CHEMBL558352
CHEMBL558543	CHEMBL558556	CHEMBL558744	CHEMBL558759	CHEMBL558954
CHEMBL559128	CHEMBL559228	CHEMBL559340	CHEMBL559342	CHEMBL562776
CHEMBL563301	CHEMBL563935	CHEMBL564205	CHEMBL564427	CHEMBL564926
CHEMBL568607	CHEMBL568755	CHEMBL568776	CHEMBL568973	CHEMBL569690
CHEMBL569907	CHEMBL569918	CHEMBL570025	CHEMBL570151	CHEMBL570155
CHEMBL570399	CHEMBL570602	CHEMBL570634	CHEMBL570855	CHEMBL571523
CHEMBL571967	CHEMBL572199	CHEMBL572203	CHEMBL572517	CHEMBL572573
CHEMBL572743	CHEMBL572762	CHEMBL572999	CHEMBL573078	CHEMBL573208
CHEMBL573234	CHEMBL573437	CHEMBL573444	CHEMBL573459	CHEMBL573461
CHEMBL6214	CHEMBL573674	CHEMBL573919	CHEMBL574025	CHEMBL574253
CHEMBL574347	CHEMBL574570	CHEMBL574797	CHEMBL574851	CHEMBL574928
CHEMBL574932	CHEMBL575155	CHEMBL575370	CHEMBL575372	CHEMBL575374
CHEMBL575375	CHEMBL576313	CHEMBL576493	CHEMBL576512	CHEMBL576519
CHEMBL576723	CHEMBL577349	CHEMBL577467	CHEMBL577535	CHEMBL577685
CHEMBL577777	CHEMBL577782	CHEMBL578346	CHEMBL578483	CHEMBL578823
CHEMBL579790	CHEMBL580443	CHEMBL580660	CHEMBL581077	CHEMBL581501
CHEMBL581928	CHEMBL582415	CHEMBL582554	CHEMBL582852	CHEMBL582854
CHEMBL582894	CHEMBL583071	CHEMBL583100	CHEMBL583284	CHEMBL583299
CHEMBL583305	CHEMBL583319	CHEMBL583567	CHEMBL583568	CHEMBL583726
CHEMBL583727	CHEMBL583920	CHEMBL583947	CHEMBL584146	CHEMBL584341
CHEMBL584356	CHEMBL584384	CHEMBL584399	CHEMBL584401	CHEMBL584402

Training Dataset ChEMBL Reference Names

CHEMBL584554	CHEMBL584574	CHEMBL584766	CHEMBL584773	CHEMBL584783
CHEMBL584972	CHEMBL584981	CHEMBL584990	CHEMBL585377	CHEMBL588329
CHEMBL589268	CHEMBL589524	CHEMBL589736	CHEMBL589753	CHEMBL589985
CHEMBL590000	CHEMBL590067	CHEMBL590259	CHEMBL6215	CHEMBL590478
CHEMBL590751	CHEMBL590853	CHEMBL590981	CHEMBL591201	CHEMBL591437
CHEMBL591703	CHEMBL591773	CHEMBL591912	CHEMBL592393	CHEMBL592409
CHEMBL592623	CHEMBL592631	CHEMBL592632	CHEMBL592870	CHEMBL592890
CHEMBL592975	CHEMBL593201	CHEMBL593625	CHEMBL593668	CHEMBL593669
CHEMBL593671	CHEMBL593672	CHEMBL593674	CHEMBL593676	CHEMBL593911
CHEMBL593912	CHEMBL593933	CHEMBL594331	CHEMBL595064	CHEMBL595080
CHEMBL595083	CHEMBL595330	CHEMBL595526	CHEMBL595820	CHEMBL596211
CHEMBL600052	CHEMBL600373	CHEMBL600456	CHEMBL601070	CHEMBL601080
CHEMBL601082	CHEMBL601104	CHEMBL601115	CHEMBL601469	CHEMBL601490
CHEMBL602107	CHEMBL602514	CHEMBL602709	CHEMBL603059	CHEMBL603312
CHEMBL603409	CHEMBL603735	CHEMBL604009	CHEMBL604773	CHEMBL605173
CHEMBL605451	CHEMBL606119	CHEMBL606548	CHEMBL606709	CHEMBL606771
CHEMBL607213	CHEMBL607364	CHEMBL607490	CHEMBL607833	CHEMBL608190
CHEMBL608407	CHEMBL608409	CHEMBL608687	CHEMBL608706	CHEMBL608826
CHEMBL608971	CHEMBL609497	CHEMBL609524	CHEMBL609992	CHEMBL610050
CHEMBL610272	CHEMBL610275	CHEMBL610474	CHEMBL6216	CHEMBL610484
CHEMBL610757	CHEMBL610765	CHEMBL610892	CHEMBL611044	CHEMBL611047
CHEMBL611115	CHEMBL611732	CHEMBL611973	CHEMBL612001	CHEMBL612129
CHEMBL612143	CHEMBL612215	CHEMBL6217	CHEMBL6218	CHEMBL6220
CHEMBL6221	CHEMBL6224	CHEMBL6226	CHEMBL6229	CHEMBL6230
CHEMBL6231	CHEMBL6233	CHEMBL6234	CHEMBL6235	CHEMBL6236
CHEMBL6237	CHEMBL6239	CHEMBL6240	CHEMBL6242	CHEMBL6244
CHEMBL6257	CHEMBL6258	CHEMBL6259	CHEMBL6282	CHEMBL6283
CHEMBL6284	CHEMBL6286	CHEMBL6287	CHEMBL6304	CHEMBL6305
CHEMBL6306	CHEMBL6307	CHEMBL6308	CHEMBL6310	CHEMBL6312
CHEMBL6314	CHEMBL6316	CHEMBL6321	CHEMBL6322	CHEMBL6328
CHEMBL6330	CHEMBL6334	CHEMBL6335	CHEMBL6347	CHEMBL6348
CHEMBL6349	CHEMBL6350	CHEMBL6352	CHEMBL6370	CHEMBL6371
CHEMBL6376	CHEMBL6393	CHEMBL6395	CHEMBL6396	CHEMBL6398
CHEMBL6399	CHEMBL6402	CHEMBL6403	CHEMBL6404	CHEMBL6407
CHEMBL6409	CHEMBL6411	CHEMBL6413	CHEMBL6414	CHEMBL6415
CHEMBL6416	CHEMBL6417	CHEMBL6418	CHEMBL6422	CHEMBL6423
CHEMBL6424	CHEMBL6425	CHEMBL6440	CHEMBL6441	CHEMBL6444
CHEMBL6445	CHEMBL6446	CHEMBL6447	CHEMBL6463	CHEMBL6464
CHEMBL6465	CHEMBL6466	CHEMBL6468	CHEMBL6484	CHEMBL6486
CHEMBL6487	CHEMBL6489	CHEMBL6496	CHEMBL6497	CHEMBL6498
CHEMBL6499	CHEMBL6503	CHEMBL6504	CHEMBL6505	CHEMBL6508
CHEMBL6509	CHEMBL6510	CHEMBL6511	CHEMBL6512	CHEMBL6514
CHEMBL6516	CHEMBL6517	CHEMBL6519	CHEMBL6521	CHEMBL6534
CHEMBL6535	CHEMBL6538	CHEMBL6560	CHEMBL6562	CHEMBL6565
CHEMBL6566	CHEMBL6583	CHEMBL6584	CHEMBL6615	CHEMBL6619

Training Dataset ChEMBL Reference Names

CHEMBL6620	CHEMBL6623	CHEMBL6624	CHEMBL6626	CHEMBL6698
CHEMBL6699	CHEMBL6788	CHEMBL6789	CHEMBL6790	CHEMBL6791
CHEMBL6792	CHEMBL6888	CHEMBL6938	CHEMBL6939	CHEMBL6940
CHEMBL6941	CHEMBL6942	CHEMBL78310	CHEMBL8	CHEMBL9
CHEMBL1208484	CHEMBL1208485	CHEMBL1208835	CHEMBL1209153	CHEMBL1209155

Testing Dataset ChEMBL Reference Names

CHEMBL1075841	CHEMBL1076027	CHEMBL1076047	CHEMBL1076054	CHEMBL1076248
CHEMBL1076255	CHEMBL1076341	CHEMBL1076614	CHEMBL1076770	CHEMBL1076895
CHEMBL1076911	CHEMBL1077021	CHEMBL1077089	CHEMBL1077098	CHEMBL1077248
CHEMBL1077270	CHEMBL1077303	CHEMBL1077330	CHEMBL1077410	CHEMBL1077444
CHEMBL1079027	CHEMBL1084170	CHEMBL1084615	CHEMBL1084953	CHEMBL1086372
CHEMBL1086503	CHEMBL1086530	CHEMBL1088207	CHEMBL1089276	CHEMBL1089589
CHEMBL1090397	CHEMBL1090670	CHEMBL16498	CHEMBL1091345	CHEMBL1091702
CHEMBL1091719	CHEMBL1091752	CHEMBL1091854	CHEMBL1092016	CHEMBL1092064
CHEMBL1092426	CHEMBL1093647	CHEMBL1094665	CHEMBL1094672	CHEMBL1095673
CHEMBL1095988	CHEMBL1095997	CHEMBL1096236	CHEMBL1096572	CHEMBL1096781
CHEMBL1096892	CHEMBL1097356	CHEMBL1097917	CHEMBL1098688	CHEMBL1099016
CHEMBL20226	CHEMBL1099322	CHEMBL1159426	CHEMBL1159612	CHEMBL1159639
CHEMBL1159835	CHEMBL1160256	CHEMBL1160450	CHEMBL1160541	CHEMBL1161970
CHEMBL1162071	CHEMBL1162097	CHEMBL1162109	CHEMBL1162123	CHEMBL1162197
CHEMBL1162300	CHEMBL1162317	CHEMBL1162464	CHEMBL1163147	CHEMBL1163207
CHEMBL1163477	CHEMBL1165348	CHEMBL1165737	CHEMBL1169618	CHEMBL1170061
CHEMBL1170642	CHEMBL1170648	CHEMBL1170714	CHEMBL1170815	CHEMBL1171136
CHEMBL1171366	CHEMBL1171716	CHEMBL1172041	CHEMBL1172740	CHEMBL1173565
CHEMBL1178041	CHEMBL1182827	CHEMBL1184967	CHEMBL1185850	CHEMBL1188857
CHEMBL1189498	CHEMBL1189651	CHEMBL1198802	CHEMBL1198965	CHEMBL1199062
CHEMBL1199367	CHEMBL1201754	CHEMBL1205123	CHEMBL1205150	CHEMBL1207005
CHEMBL1207408	CHEMBL1207630	CHEMBL1208039	CHEMBL1209204	CHEMBL1209874
CHEMBL1210336	CHEMBL1213982	CHEMBL2	CHEMBL1215108	CHEMBL1215753
CHEMBL1215806	CHEMBL1222793	CHEMBL1222975	CHEMBL1223621	CHEMBL1224871
CHEMBL1229187	CHEMBL154170	CHEMBL154357	CHEMBL154655	CHEMBL154687
CHEMBL155263	CHEMBL155656	CHEMBL156263	CHEMBL156703	CHEMBL156901
CHEMBL157267	CHEMBL157459	CHEMBL158232	CHEMBL158670	CHEMBL159389
CHEMBL160341	CHEMBL160417	CHEMBL162493	CHEMBL162495	CHEMBL164857
CHEMBL174668	CHEMBL2070335	CHEMBL216546	CHEMBL2310848	CHEMBL259626
CHEMBL260023	CHEMBL260906	CHEMBL262299	CHEMBL264521	CHEMBL265282
CHEMBL305803	CHEMBL405	CHEMBL3309292	CHEMBL3309314	CHEMBL3309327
CHEMBL3309334	CHEMBL3309338	CHEMBL3309390	CHEMBL3309411	CHEMBL3309413
CHEMBL3309427	CHEMBL3309443	CHEMBL3309512	CHEMBL3309524	CHEMBL3309670
CHEMBL3347277	CHEMBL3347280	CHEMBL3347290	CHEMBL385868	CHEMBL402708
CHEMBL407362	CHEMBL4116103	CHEMBL4116106	CHEMBL4116108	CHEMBL4116113
CHEMBL4116117	CHEMBL4116125	CHEMBL4116128	CHEMBL4116136	CHEMBL4116144
CHEMBL4116145	CHEMBL4116147	CHEMBL4116148	CHEMBL4116154	CHEMBL4116156
CHEMBL4116160	CHEMBL4116167	CHEMBL418	CHEMBL437851	CHEMBL438329
CHEMBL438822	CHEMBL443179	CHEMBL443332	CHEMBL443686	CHEMBL443992
CHEMBL444524	CHEMBL444748	CHEMBL445177	CHEMBL454492	CHEMBL498858
CHEMBL498966	CHEMBL499583	CHEMBL499917	CHEMBL499990	CHEMBL500234
CHEMBL500267	CHEMBL500269	CHEMBL500534	CHEMBL501513	CHEMBL501544
CHEMBL501701	CHEMBL501770	CHEMBL501821	CHEMBL501874	CHEMBL501943
CHEMBL501969	CHEMBL502172	CHEMBL502237	CHEMBL502461	CHEMBL502582
CHEMBL502985	CHEMBL502986	CHEMBL503030	CHEMBL503044	CHEMBL503204

Testing Dataset ChEMBL Reference Names

CHEMBL503315	CHEMBL503376	CHEMBL503400	CHEMBL503419	CHEMBL503454
CHEMBL503463	CHEMBL503466	CHEMBL503467	CHEMBL503523	CHEMBL503535
CHEMBL503541	CHEMBL503644	CHEMBL503660	CHEMBL503770	CHEMBL503846
CHEMBL503860	CHEMBL503865	CHEMBL503902	CHEMBL503907	CHEMBL503982
CHEMBL503994	CHEMBL504059	CHEMBL504077	CHEMBL504237	CHEMBL504331
CHEMBL504332	CHEMBL504351	CHEMBL504410	CHEMBL504419	CHEMBL504420
CHEMBL504560	CHEMBL504907	CHEMBL504909	CHEMBL505029	CHEMBL505283
CHEMBL505285	CHEMBL505400	CHEMBL505686	CHEMBL505936	CHEMBL506058
CHEMBL506162	CHEMBL506172	CHEMBL506308	CHEMBL506642	CHEMBL506982
CHEMBL506998	CHEMBL507224	CHEMBL507305	CHEMBL507459	CHEMBL507464
CHEMBL507540	CHEMBL507756	CHEMBL507894	CHEMBL507903	CHEMBL508600
CHEMBL524910	CHEMBL525428	CHEMBL525996	CHEMBL526134	CHEMBL526533
CHEMBL526908	CHEMBL527081	CHEMBL528112	CHEMBL529327	CHEMBL532920
CHEMBL533010	CHEMBL535264	CHEMBL538639	CHEMBL538928	CHEMBL539138
CHEMBL539714	CHEMBL540943	CHEMBL541185	CHEMBL541498	CHEMBL541676
CHEMBL541695	CHEMBL541956	CHEMBL547542	CHEMBL547643	CHEMBL552742
CHEMBL553645	CHEMBL554546	CHEMBL557764	CHEMBL563918	CHEMBL564260
CHEMBL569031	CHEMBL569917	CHEMBL570152	CHEMBL570853	CHEMBL570872
CHEMBL571980	CHEMBL572530	CHEMBL572625	CHEMBL573427	CHEMBL573664
CHEMBL574027	CHEMBL574221	CHEMBL574800	CHEMBL574930	CHEMBL576264
CHEMBL576309	CHEMBL577321	CHEMBL577964	CHEMBL578763	CHEMBL579584
CHEMBL579880	CHEMBL583126	CHEMBL583519	CHEMBL584197	CHEMBL584397
CHEMBL584784	CHEMBL584792	CHEMBL585178	CHEMBL588522	CHEMBL589825
CHEMBL590003	CHEMBL590250	CHEMBL591228	CHEMBL592411	CHEMBL593658
CHEMBL593812	CHEMBL593909	CHEMBL594149	CHEMBL595765	CHEMBL595793
CHEMBL595990	CHEMBL600867	CHEMBL601272	CHEMBL602115	CHEMBL602507
CHEMBL603145	CHEMBL604582	CHEMBL606769	CHEMBL608847	CHEMBL609157
CHEMBL609878	CHEMBL609929	CHEMBL610191	CHEMBL610630	CHEMBL610879
CHEMBL611662	CHEMBL611935	CHEMBL6219	CHEMBL6222	CHEMBL6225
CHEMBL6232	CHEMBL6243	CHEMBL6245	CHEMBL6261	CHEMBL6309
CHEMBL6311	CHEMBL6317	CHEMBL6320	CHEMBL6323	CHEMBL6324
CHEMBL6332	CHEMBL6369	CHEMBL6378	CHEMBL6400	CHEMBL6401
CHEMBL6437	CHEMBL6442	CHEMBL6462	CHEMBL6467	CHEMBL6500
CHEMBL6502	CHEMBL6518	CHEMBL6564	CHEMBL6582	CHEMBL6586
CHEMBL6618	CHEMBL6625	CHEMBL6696	CHEMBL6887	CHEMBL6889

A.3 Full Gaussian Reference

Gaussian 09, Revision D.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.

BIBLIOGRAPHY

[1] C. Dickson and C. P. Butts, "Unpublished work," *University of Bristol*, 2021.

[2] R. Kashaev and N. Faskhiev, "Determination of dispersity of aqueous emulsions of hydrocarbons by nuclear magnetic resonance relaxometry," *Chem. Technol. Fuels Oils*, vol. 47, 2011.

[3] E. VanderHeiden, "Applications of nmr imaging on solid rocket motor materials," in *Review of Progress in Quantitative Nondestructive Evaluation*, Springer, 1992.

[4] S. J. Chang, J. R. Olson, and P. C. Wang, "Nmr imaging of internal features in wood," *For. Prod. J.*, 1989.

[5] P. Rinck, "Magnetic resonance in medicine. the basic textbook of the european magnetic resonance forum.," 2017.

[6] L. H. Antonides, R. M. Brignall, A. Costello, J. Ellison, S. E. Firth, N. Gilbert, B. J. Groom, S. J. Hudson, M. C. Hulme, J. Marron, *et al.*, "Rapid identification of novel psychoactive and other controlled substances using low-field ^{1}H nmr spectroscopy," *ACS Omega*, vol. 4, 2019.

[7] F. Dalitz, M. Cudaj, M. Maiwald, and G. Guthausen, "Process and reaction monitoring by low-field nmr spectroscopy," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 60, 2012.

[8] T. Bartik, B. Bartik, B. E. Hanson, T. Glass, and W. Bebout, "Comments on the synthesis of trisulfonated triphenylphosphine: reaction monitoring by nmr spectroscopy," *Inorg. Chem.*, vol. 31, 1992.

[9] M. V. Gomez and A. de la Hoz, "Nmr reaction monitoring in flow synthesis," *Beilstein J. Org. Chem.*, vol. 13, 2017.

[10] M. Gallo, S. Matteucci, N. Alaimo, E. Pitti, M. V. Orsale, V. Summa, D. O. Cicero, and E. Monteagudo, "A novel method using nuclear magnetic resonance for plasma protein binding assessment in drug discovery programs," *J. Pharm. Biomed. Anal.*, vol. 167, 2019.

[11] J. Feigon, W. A. Denny, W. Leupin, and D. R. Kearns, "Interactions of antitumor drugs with natural dna: proton nmr study of binding mode and kinetics," *J. Med. Chem.*, vol. 27, 1984.

[12] C. Nitsche and G. Otting, "Nmr studies of ligand binding," *Curr. Opin. Struct. Biol.*, vol. 48, 2018.

[13] E. Breitmaier and A. Sinnema, "Structure elucidation by nmr in organic chemistry," 1993.

[14] G. Morris, "Modern nmr techniques for structure elucidation," *Magn. Reson. Chem.*, vol. 24, 1986.

[15] M. Elyashberg, "Identification and structure elucidation by nmr spectroscopy," *Trends. Analyt. Chem.*, vol. 69, 2015.

[16] J. Wu, P. Lorenzo, S. Zhong, M. Ali, C. P. Butts, E. L. Myers, and V. K. Aggarwal, "Synergy of synthesis, computation and nmr reveals correct baulamycin structures," *Nature*, vol. 547, 2017.

[17] A. Navarro-Vázquez, R. R. Gil, and K. Blinov, "Computer-assisted 3d structure elucidation (case-3d) of natural products combining isotropic and anisotropic nmr parameters," *J. Nat. Prod.*, vol. 81, 2018.

[18] A. V. Buevich and M. E. Elyashberg, "Towards unbiased and more versatile nmr-based structure elucidation: A powerful combination of case algorithms and dft calculations," *Magn. Reson. Chem.*, vol. 56, 2018.

[19] W. Proctor and F. Yu, "The dependence of a nuclear magnetic resonance frequency upon chemical compound," *Phys. Rev.*, vol. 77, 1950.

[20] M. Karplus, "Vicinal proton coupling in nuclear magnetic resonance," *J. Am. Chem. Soc.*, vol. 85, 1963.

[21] P. Geerlings, F. De Proft, and W. Langenaeker, "Conceptual density functional theory," *Chem. Rev.*, vol. 103, 2003.

[22] R. Abraham and K. Pachler, "The proton magnetic resonance spectra of some substituted ethanes: the influence of substitution on ch-ch coupling constants," *Mol. Phys.*, vol. 7, 1964.

[23] P. L. Durette and D. Horton, "Conformational studies on pyranoid sugar derivatives by nmr spectroscopy. correlations of observed proton—proton coupling constants with the generalized karplus equation," *Org. Magn. Reson.e*, vol. 3, 1971.

[24] C. Haasnoot, F. A. de Leeuw, and C. Altona, "The relationship between proton-proton nmr coupling constants and substituent electronegativities—i: an empirical generalization of the karplus equation," *Tetrahedron*, vol. 36, 1980.

[25] C. Haasnoot, F. De Leeuw, H. De Leeuw, and C. Altona, "Relationship between proton–proton nmr coupling constants and substituent electronegativities. iii. conformational analysis of proline rings in solution using a generalized karplus equation," *Biopolymers*, vol. 20, 1981.

[26] K. Imai and E. Ōsawa, "An empirical extension of the karplus equation," *Magn. Reson. Chem.*, vol. 28, 1990.

[27] M. Hricovíni and I. Tvaroška, "Conformational dependence of the one-bond carbon–proton coupling constants in oligosaccharides," *Magn. Reson. Chem.*, vol. 28, 1990.

[28] I. Tvaroska and F. R. Taravel, "One-bond carbon-proton coupling constants: angular dependence in α-linked oligosaccharides," *Carbohydr. Res.*, vol. 221, 1991.

[29] A. Perlin and B. Casu, "Carbon-13 and proton magnetic resonance spectra of d-glucose-13c," *Tetrahedron Lett.*, vol. 10, 1969.

[30] K. Bock, C. Pedersen, *et al.*, "Two-and three-bond ^{13}C–^{1}H couplings in some carbohydrates," *Acta Chem. Scand.*, vol. 31, 1977.

[31] R. Aydin and H. Günther, "^{13}C, ^{1}H spin–spin coupling. x—norbornane: a reinvestigation of the karplus curve for 3j (^{13}C, ^{1}H)," *Magn. Reson. Chem.*, vol. 28, 1990.

[32] R. B. Schaller, C. Arnold, and E. Pretsch, "New parameters for predicting ^{1}H nmr chemical shifts of protons attached to carbon atoms," *Anal. Chim. Acta.*, vol. 312, 1995.

[33] E. Pretsch, A. Furst, M. Badertscher, R. Buergin, and M. E. Munk, "C13shift: a computer program for the prediction of carbon-13 nmr spectra based on an open set of additivity rules," *J. Chem. Inf. Model.*, vol. 32, 1992.

[34] W. Bremser, "Hose—a novel substructure code," *Anal. Chim. Acta.*, vol. 103, 1978.

[35] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," in *AIP Conf. Proc.*, vol. 577, AIP, 2001.

[36] A. D. Beck, "Density-functional thermochemistry. iii. the role of exact exchange," *J. Chem. Phys.*, vol. 98, 1993.

[37] C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density," *Phys. Rev. B Condens. Matter*, vol. 37, 1988.

[38] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Phys. Rev. Lett.*, vol. 77, 1996.

[39] R. F. Stewart, "Small gaussian expansions of slater-type orbitals," *J. Chem. Phys.*, vol. 52, 1970.

[40] T. H. Dunning Jr, "Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen," *J. Chem. Phys.*, vol. 90, 1989.

[41] F. Jensen, "Polarization consistent basis sets: principles," *J. Chem. Phys.*, vol. 115, 2001.

[42] R. Ditchfield, W. J. Hehre, and J. A. Pople, "Self-consistent molecular-orbital methods. ix. an extended gaussian-type basis for molecular-orbital studies of organic molecules," *J. Chem. Phys.*, vol. 54, 1971.

[43] R. Ditchfield, "Self-consistent perturbation theory of diamagnetism," *Mol. Phys.*, vol. 27, no. 4, pp. 789–807, 1974.

[44] K. Wolinski, J. F. Hinton, and P. Pulay, "Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations," *J. Am. Chem. Soc.*, vol. 112, 1990.

[45] J. R. Cheeseman, G. W. Trucks, T. A. Keith, and M. J. Frisch, "A comparison of models for calculating nuclear magnetic resonance shielding tensors," *J. Chem. Phys.*, vol. 104, 1996.

[46] I. Alkorta and J. Elguero, "Review on dft and ab initio calculations of scalar coupling constants," *Int. J. Mol. Sci.*, vol. 4, 2003.

[47] T. Helgaker, M. Jaszuński, and P. Świder, "Calculation of nmr spin–spin coupling constants in strychnine," *J. Org. Chem.*, vol. 81, 2016.

[48] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. Petersson, and s. S. S. o. S. I. others (for the full reference, "Gaussian 09 program; gaussian, inc," *Wallingford, CT*, 2016.

[49] M. Frisch, G. Trucks, H. Schlegel, G. Scuseria, M. Robb, J. Cheeseman, G. Scalmani, V. Barone, G. Petersson, H. Nakatsuji, *et al.*, "Gaussian 16," 2016.

[50] H. Schlegel and J. McDouall, "Computational bottlenecks in molecular orbital calculations," *adv. comput. chem.*, 1991.

[51] G. Landrum, "Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling," 2013.

[52] F. Mohamadi, N. G. Richards, W. C. Guida, R. Liskamp, M. Lipton, C. Caufield, G. Chang, T. Hendrickson, and W. C. Still, "Macromodel—an integrated software system for modeling organic and bioorganic molecules using molecular mechanics," *J. Comput. Chem.*, vol. 11, 1990.

[53] S. G. Smith and J. M. Goodman, "Assigning stereochemistry to single diastereoisomers by giao nmr calculation: the dp4 probability," *J. Am. Chem. Soc.*, vol. 132, 2010.

[54] N. Grimblat, M. M. Zanardi, and A. M. Sarotti, "Beyond dp4: an improved probability for the stereochemical assignment of isomeric compounds using quantum chemical calculations of nmr shifts," *J. Org. Chem.*, vol. 80, 2015.

[55] C. Fan, Z. Cui, and X. Zhong, "House prices prediction with machine learning algorithms," in *Proc. 10th Int. Conf. Electron. Comput. Artif. Intell. ECAI 2018*, 2018.

[56] M. Rathi, A. Malik, D. Varshney, R. Sharma, and S. Mendiratta, "Sentiment analysis of tweets using machine learning approach," in *2018 11th Int. Conf. Contemp. Comput. IC3 2018*, IEEE, 2018.

[57] H. M. Zawbaa, M. Abbass, S. H. Basha, M. Hazman, and A. E. Hassenian, "An automatic flower classification approach using machine learning algorithms," in *2018 Int. Conf. Adv. Comput. Commun. Inform. ICACCI 2018*, IEEE, 2014.

[58] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, "Chemical shifts in molecular solids by machine learning," *Nat. Commun.*, vol. 9, 2018.

[59] E. Jonas and S. Kuhn, "Rapid prediction of nmr spectral properties with quantified uncertainty," *J. Cheminformatics*, vol. 11, 2019.

[60] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. Da Silva, and M. G. Quiles, "Machine learning prediction of nine molecular properties based on the smiles representation of the qm9 quantum-chemistry dataset," *J. Phys. Chem. A*, vol. 124, 2020.

[61] M. Tsubaki and T. Mizoguchi, "Fast and accurate molecular property prediction: learning atomic interactions and potentials with neural networks," *J. Phys. Chem. Lett.*, vol. 9, 2018.

[62] T. S. Hy, S. Trivedi, H. Pan, B. M. Anderson, and R. Kondor, "Predicting molecular properties with covariant compositional networks," *J. Chem. Phys.*, vol. 148, 2018.

[63] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," *Chem. Mater.*, vol. 31, 2019.

[64] L. Ruddigkeit, R. Van Deursen, L. C. Blum, and J.-L. Reymond, "Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17," *J. Chem. Inf. Model.*, vol. 52, 2012.

[65] A. Gupta, S. Chakraborty, and R. Ramakrishnan, "Revving up ^{13}C nmr shielding predictions across chemical space: Benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules," *Mach. learn.: sci. technol.*, 2021.

[66] W. Gerrard, L. A. Bratholm, M. J. Packer, A. J. Mulholland, D. R. Glowacki, and C. P. Butts, "Impression–prediction of nmr parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy," *Chem. Sci.*, vol. 11, 2020.

[67] K. Shibata and H. Kaneko, "Prediction of spin–spin coupling constants with machine learning in nmr," *Anal. Sci. Adv.*, 2021.

[68] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[69] T. Charpentier, "The paw/gipaw approach for computing nmr parameters: a new dimension added to nmr study of solids," *Solid State Nucl. Magn. Reson.*, vol. 40, 2011.

[70] R. J. Campello, D. Moulavi, A. Zimek, and J. Sander, "Hierarchical density estimates for data clustering, visualization, and outlier detection," *ACM. T. Knowl. Discov. D.*, vol. 10, 2015.

[71] M. Ceriotti, G. A. Tribello, and M. Parrinello, "Demonstrating the transferability and the descriptive power of sketch-map," *J. Chem. Theory Comput.*, vol. 9, 2013.

[72] A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments," *Phys. Rev. B Condens. Matter*, vol. 87, 2013.

[73] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, *et al.*, "Quantum espresso: a modular and open-source software project for quantum simulations of materials," *J. Condens. Matter Phys.*, vol. 21, 2009.

[74] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[75] S. Kuhn and N. E. Schlörer, "Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2–a free in-house nmr database with integrated lims for academic service laboratories," *Magn. Reson. Chem.*, vol. 53, 2015.

[76] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," *1998 15th Int. Conf. Mach. Learn.*

[77] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Sci. Data*, vol. 1, 2014.

[78] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, "Fast and accurate modeling of molecular atomization energies with machine learning," *Phys. Rev. Lett.*, vol. 108, 2012.

[79] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, "Fchl revisited: faster and more accurate quantum machine learning," *J. Chem. Phys.*, vol. 152, 2020.

[80] L. A. Bratholm, W. Gerrard, B. Anderson, S. Bai, S. Choi, L. Dang, P. Hanchar, A. Howard, G. Huard, S. Kim, *et al.*, "A community-powered search of machine learning strategy space to find nmr property prediction models," *arXiv preprint arXiv:2008.05994*, 2020.

[81] B. Pagenkopf, "Acd/hnmr predictor and acd/cnmr predictor advanced chemistry development, inc.(acd/labs), 90 adelaide street west, suite 600, toronto, on m5h 2v9, canada. www. acdlabs. com. see web site for pricing information.," 2005.

[82] J. Meiler, W. Maier, M. Will, and R. Meusinger, "Using neural networks for ^{13}C nmr chemical shift prediction–comparison with traditional methods," *J. Magn. Reson.*, vol. 157, 2002.

[83] C. Steinbeck and S. Kuhn, "Nmrshiftdb–compound identification and structure elucidation support through a free community-built web database," *Phytochemistry*, vol. 65, 2004.

[84] K. Blinov, M. Kvasha, B. Lefebvre, R. Sasaki, and A. Williams, "NMR prediction accuracy validation," *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.632.2291rep=rep1type=pdf*.

[85] S. Kuhn, B. Egert, S. Neumann, and C. Steinbeck, "Building blocks for automated elucidation of metabolites: Machine learning methods for nmr prediction," *BMC Bioinform*, vol. 9, 2008.

[86] J. Meiler, B. Lefebvre, A. Williams, and M. Hachey, "Addendum to "using neural networks for ^{13}C nmr chemical shift prediction—comparison with traditional methods: [j. magn. reson. 157 (2002) 242–252]," *J. Magn. Reson.*, vol. 1, 2004.

[87] C. Steinbeck, S. Krause, and S. Kuhn, "Nmrshiftdb constructing a free chemical information system with open-source components," *J. Chem. Inf. Model.*, vol. 43, 2003.

[88] T. Helgaker, M. Jaszuński, and M. Pecul, "The quantum-chemical calculation of nmr indirect spin–spin coupling constants," *Prog. Nucl. Magn. Reson. Spectrosc.*, vol. 4, 2008.

[89] S. N. Maximoff, J. E. Peralta, V. Barone, and G. E. Scuseria, "Assessment of density functionals for predicting one-bond carbon- hydrogen nmr spin-spin coupling constants," *J. Chem. Theory Comput.*, vol. 1, 2005.

[90] J. F. San, J. de la Vega García, R. Suardíaz, M. Fernández-Oliva, C. Pérez, R. Crespo-Otero, and R. Contreras, "Computational nmr coupling constants: shifting and scaling factors for evaluating 1jch.," *Magn. Reson. Chem.*, vol. 51, 2013.

[91] V. A. Semenov and L. B. Krivdin, "Dft computational schemes for ^{1}H and ^{13}C nmr chemical shifts of natural products, exemplified by strychnine," *Magn. Reson. Chem.*, 2019.

[92] A. Navarro-Vázquez, "State of the art and perspectives in the application of quantum chemical prediction of ^{1}H and ^{13}C chemical shifts and scalar couplings for structural elucidation of organic compounds," *Magn. Reson. Chem.*, vol. 55, 2017.

[93] M. W. Lodewyk, M. R. Siebert, and D. J. Tantillo, "Computational prediction of ^{1}H and ^{13}C chemical shifts: A useful tool for natural product, mechanistic, and synthetic organic chemistry," *Chem. Rev.*, vol. 112, 2012.

[94] C. Steinmann, L. A. Bratholm, J. M. H. Olsen, and J. Kongsted, "Automated fragmentation polarizable embedding density functional theory (pe-dft) calculations of nuclear magnetic resonance (nmr) shielding constants of proteins with application to chemical shift predictions," *J. Chem. Theory Comput.*, vol. 13, no. 2, pp. 525–536, 2017.
PMID: 27992211.

[95] A. S. Larsen, L. A. Bratholm, A. S. Christensen, M. Channir, and J. H. Jensen, "Procs15: a dft-based chemical shift predictor for backbone and cβ atoms in proteins," *PeerJ*, vol. 3, 2015.

[96] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, 2021.

[97] M. C. Robinson, R. C. Glen, *et al.*, "Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction," *J. Comput. Aided Mol. Des.*, 2020.

[98] V. A. Semenov and L. B. Krivdin, "Dft computational schemes for ^{1}H and ^{13}C nmr chemical shifts of natural products, exemplified by strychnine," *Magn. Reson. Chem.*, vol. 58, 2020.

[99] M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy, and B. Da Mota, "Dataset's chemical diversity limits the generalizability of machine learning predictions," *J. Cheminformatics*, vol. 11, 2019.

[100] C. R. Groom, I. J. Bruno, M. P. Lightfoot, and S. C. Ward, "The cambridge structural database," *Acta. Crystallogr. B. Struct.*, vol. 72, pp. 171–179, Apr 2016.

[101] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, *et al.*, "Chembl: towards direct deposition of bioassay data," *Nucleic Acids Res.*, vol. 47, 2019.

[102] C. Adamo and V. Barone, "Exchange functionals with improved long-range behavior and adiabatic connection methods without adjustable parameters: The m pw and m pw1pw models," *J. Chem. Phys.*, vol. 108, 1998.

[103] A. McLean and G. Chandler, "Contracted gaussian basis sets for molecular calculations. i. second row atoms, z= 11–18," *J. Chem. Phys.*, vol. 72, 1980.

[104] R. Krishnan, J. S. Binkley, R. Seeger, and J. A. Pople, "Self-consistent molecular orbital methods. xx. a basis set for correlated wave functions," *J. Chem. Phys.*, vol. 72, no. 1, pp. 650–654, 1980.

[105] J.-D. Chai and M. Head-Gordon, "Systematic optimization of long-range corrected hybrid density functionals," *J. Chem. Phys.*, vol. 128, 2008.

[106] W. Deng, J. R. Cheeseman, and M. J. Frisch, "Calculation of nuclear spin-spin coupling constants of molecules with first and second row atoms in study of basis set dependence," *J. Chem. Theory Comput.*, vol. 2, 2006.

[107] R. Laskowski, P. Blaha, and F. Tran, *CHESHIRE Chemical Shift Repository*, 2019 (accessed October 2nd, 2019).

[108] P. Gao, X. Wang, and H. Yu, "Towards an accurate prediction of nitrogen chemical shifts by density functional theory and gauge-including atomic orbital," *Adv. Theory Simul.*, vol. 2, 2019.

[109] W. Gerrard, C. Yiu, and M. William, "mol_translator python package," 2021.

[110] T. E. Oliphant, *A guide to NumPy*, vol. 1.
Trelgol Publishing USA, 2006.

[111] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox," *J. Cheminformatics*, vol. 3, 2011.

[112] N. M. O'Boyle, C. Morley, and G. R. Hutchison, "Pybel: a python wrapper for the openbabel cheminformatics toolkit," *Chem. Cent. J.*, vol. 2, 2008.

[113] K. Sato and K. Tsuruta, "Optimization of molecular characteristics via machine learning based on continuous representation of molecules," in *Materials Science Forum*, vol. 1016, Trans Tech Publ, 2021.

[114] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Exploring chemical compound space with quantum-based machine learning," *Nat. Rev. Chem.*, vol. 4, 2020.

[115] C. Venkata, M. J. Forster, P. W. Howe, and C. Steinbeck, "The potential utility of predicted one bond carbon-proton coupling constants in the structure elucidation of small organic molecules by nmr spectroscopy," *PLOS ONE*, vol. 9, 2014.

[116] M. Rupp, R. Ramakrishnan, and O. A. Von Lilienfeld, "Machine learning for quantum mechanical properties of atoms in molecules," *J. Phys. Chem. Lett.*, vol. 6, 2015.

[117] B. Huang and O. A. von Lilienfeld, "The" dna" of chemistry: scalable quantum machine learning with" amons"," *arXiv preprint arXiv:1707.04146*, 2017.

[118] F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, "Alchemical and structural distribution based representation for universal quantum machine learning," *J. Chem. Phys.*, vol. 148, 2018.

[119] C. R. Collins, G. J. Gordon, O. A. von Lilienfeld, and D. J. Yaron, "Constant size molecular descriptors for use with machine learning," *arXiv preprint arXiv:1701.06649*, 2017.

[120] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Phys. Chem. Chem. Phys.*, vol. 18, 2016.

[121] B. Huang and O. A. Von Lilienfeld, "Communication: understanding molecular representations in machine learning: The role of uniqueness and target similarity," 2016.

[122] B. Haasdonk and H. Burkhardt, "Invariant kernel functions for pattern analysis and machine learning," *Mach. Learn.*, vol. 68, 2007.

[123] C. R. Souza, "Kernel functions for machine learning applications," *www.crsouza.com*, vol. 3, 2010.

[124] M. Feurer and F. Hutter, "Hyperparameter optimization," in *Automated Machine Learning*, Springer, 2019.

[125] fmfn, "Bayesianoptimization." `https://github.com/fmfn/BayesianOptimization`, 2019.

[126] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.

[127] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, *et al.*, "Deep graph library: a graph-centric, highly-performant package for graph neural networks," *arXiv preprint arXiv:1909.01315*, 2019.

[128] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[129] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[130] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[131] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh, "Large batch optimization for deep learning: training bert in 76 minutes," *arXiv preprint arXiv:1904.00962*, 2019.

[132] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[133] A. V. Buevich, J. Saurí, T. Parella, N. De Tommasi, G. Bifulco, R. T. Williamson, and G. E. Martin, "Enhancing the utility of 1 j ch coupling constants in structural studies through optimized dft analysis," *Chem. Comm.*, vol. 55, 2019.

[134] V. Kairys, L. Baranauskiene, M. Kazlauskiene, D. Matulis, and E. Kazlauskas, "Binding affinity in drug design: experimental and computational techniques," *Expert Opin. Drug Discov.*, vol. 14, 2019.

[135] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," *Br. J. Pharmacol.*, vol. 162, 2011.

[136] L. M. Mayr and P. Fuerst, "The future of high-throughput screening," *J. Biomol. Screen*, vol. 13, 2008.

[137] R. W. Zwanzig, "High-temperature equation of state by a perturbation method. i. nonpolar gases," *J. Chem. Phys.*, vol. 22, 1954.

[138] G. S Heck, V. O Pintro, R. R Pereira, N. MB Levin, W. F de Azevedo, *et al.*, "Supervised machine learning methods applied to predict ligand-binding affinity," *Curr. Med. Chem.*, vol. 24, 2017.

[139] P. J. Ballester and J. B. Mitchell, "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking," *Bioinform.*, vol. 26, 2010.

[140] M. H. Seifert, "Robust optimization of scoring functions for a target class," *J. Comput. Aided Mol. Des.*, vol. 23, 2009.

[141] F. Deflorian, L. Perez-Benito, E. B. Lenselink, M. Congreve, H. W. van Vlijmen, J. S. Mason, C. d. Graaf, and G. Tresadern, "Accurate prediction of gpcr ligand binding affinity with free energy perturbation," *J. Chem. Inf. Model.*, vol. 60, 2020.

[142] B. J. Williams-Noonan, E. Yuriev, and D. K. Chalmers, "Free energy methods in drug design: prospects of "alchemical perturbation" in medicinal chemistry: miniperspective," *J. Med. Chem.*, vol. 61, 2018.

[143] K. Vanommeslaeghe and A. D. MacKerell Jr, "Automation of the charmm general force field (cgenff) i: bond perception and atom typing," *J. Chem. Inf. Model.*, vol. 52, 2012.

[144] S. Kumar and M.-h. Kim, "Smplip-score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors," *J. Cheminformatics*, vol. 13, 2021.

[145] C. D. Parks, Z. Gaieb, M. Chiu, H. Yang, C. Shao, W. P. Walters, J. M. Jansen, G. McGaughey, R. A. Lewis, S. D. Bembenek, *et al.*, "D3r grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies," *J. Comput. Aided Mol. Des.*, vol. 34, 2020.

[146] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, 2010.

[147] K. Hara, D. Saito, and H. Shouno, "Analysis of function of rectified linear unit used in deep learning," in *2015 Proc. Int. Jt. Conf. Neural Netw.*, IEEE, 2015.

[148] H. L. Morgan, "The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service.," *J. Chem. Doc.*, vol. 5, 1965.

[149] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods*, vol. 71, 2015.

[150] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, 1997.

[151] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Trans. Knowl.*, vol. 17, 2005.