Title

Combining Data-driven models with Physics-based Approaches for Computational Molecule Characterization and Generation

Permalink

https://escholarship.org/uc/item/0zn838xz

Author

Li, Jie

Publication Date

2023

Peer reviewed|Thesis/dissertation

Combining Data-driven models with Physics-based Approaches for Computational Molecule Characterization and Generation

by

Jie Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Teresa Head-Gordon, Chair
Professor Mark van der Laan
Professor K. Birgitta Whaley

Summer 2023

Combining Data-driven models with Physics-based Approaches for Computational Molecule Characterization and Generation

Abstract

Combining Data-driven models with Physics-based Approaches for Computational Molecule Characterization and Generation

by

Jie Li

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Teresa Head-Gordon, Chair

Theoretical studies of molecules have historically relied on deterministic algorithms, stochastic simulations, and physical models. Recently modern data-driven methods are starting to infiltrate into various fields of molecular science, opening new possibilities for solving problems that are difficult to tackle through traditional approaches. The accumulation of data, advancement of machine learning algorithms and improvement in hardware enables a plethora of data-driven approaches to surpass traditional methods in terms of accuracy and efficiency, but questions remain about how well these data-driven methods can generalize to unseen data to do true prediction. In this dissertation, I will show that when data-driven models are combined with physics-based approaches, through either feature design, or exerting constrains on the machine learning models, new standards can be established in the fields of molecule characterization and generation.

Nuclear magnetic resonance (NMR) chemical shifts (CS) are extremely sensitive to the local atomic environments for different nuclei in a molecule, and therefore is a common technique in molecule characterization. In chapter 2, I focus on the design of the UCBShift predictor for CSs for proteins in aqueous solution. The UCBShift method uniquely fuses a transfer prediction module, which employs sequence and structure alignments to select reference chemical shifts from a database, with a machine learning model that uses carefully curated and physics-inspired features, to predict CSs for proteins with higher accuracy and better reliability compared to all popular methods such as SHIFTX2 and SPARTA+. This chapter further delineates how UCBShift benefits from realistic data that has not been heavily curated, and surpasses existing CS calculators in terms of real-world performance without eliminating test predictions *ad hoc*.

However, in order to achieve rigorous and consistent improvement for an *arbitrary* molecular system, carefully curated feature sets specifically for proteins can be limiting, and we seek features from theoretical calculations. In Chapter 3 I describe the development of a novel

neural network model which employs quantum mechanical (QM) features from affordable Density Functional Theory (DFT) calculations, along with geometric features of the molecular systems, to predict NMR chemical shieldings. The resulting iShiftML model predicts chemical shieldings approaching the highest level of accuracy under the modern theoretical framework of CCSD(T) in the complete basis set limit, but without the computational burden that limits its applicability to large systems. Not only does the iShiftML model demonstrate excellent predictive performance when compared with small molecule gas phase experimental CSs, but it also offers a capability to predict chemical shifts for much more complex natural products, and can be used for differentiating diasteromers based on chemical shift assignments. This chapter unveils new possibilities for integrating machine learning and QM calculations for accurate and transferable molecular characterization.

In Chapters 4 an 5, my research addresses fundamental issues for large and small molecule generation relevant to proteins and drug molecules. Chapter 4 describes the Int2Cart method that uses a recurrent neural network to predict the correlations between bond lengths, bond angles and backbone torsion angles and amino acid sequence of a protein. By incorporating these correlations, proteins reconstructed from just torsion angles display not only physically more accurate bond lengths and bond angles, but the reconstructed proteins are closer to their crystal structures than under the common assumption that bond lengths and bond angles are fixed, or that coming from a static library that only relies on local residue geometries. I have also shown potential applications of this method in estimating model quality for AlphaFold2 predicted structures, and reconstructing intrinsically disordered protein (IDP) ensembles with decreased steric overlap. Chapter 5 describes the combination of deep generative networks trained by reinforcement learning and physical docking study. I developed the iMiner method, which generates *de novo* drug-like molecules with an augmented binding potency towards specific protein targets, facilitating the discovery of potential new therapeutic targets. SARS-COV-2 Main Protease was used as an example to show that our generated molecules cover a broader chemical space than crowdsourcing efforts, and the newly generated molecules exert optimized interactions and correct shape for the compatibility with the binding pocket.

To summarize, this dissertation contains multiple methods that harmonize data-driven models and physics-based approaches in the area of NMR spectroscopy, protein structure modelling and *de novo* drug discovery, which provides a new perspective for researchers striving to leverage computational methods in molecular science and chemical biology.

Contents

Acknowledgments

Reflecting on the five years spent in the Chemistry Ph.D. program, there are countless people and experiences I wish to express my gratitude towards. First and foremost, my deepest appreciation goes to Teresa Head-Gordon, my Principal Investigator and mentor. Her infectious passion for science has significantly influenced my approach towards research, guiding me towards the exhilaration of scientific discovery. I am profoundly grateful for the freedom in choosing my research topics, which made my PhD experience both challenging and enjoyable. Beyond academia, Teresa's warmth and care provided a sense of home in an unfamiliar country.

I would also like to thank the many colleagues in the lab who supported me throughout my journey. Dr. Kochise Bennitt and Dr. Shuai Liu helped me starting machine learning research in the lab. Collaborating with the knowledgeable Dr. Mojtaba Haghighatlari and Dr. Farnaz Heidar-Zadeh on numerous fascinating projects was a great privilege. Furthermore, I would like to extend my many thanks to Nancy Guan, Oufan Zhang, Oliver Sun, Eric Wang and many other group members for the exciting projects we have been working together and the enjoyable discussions we have had in the lab. In addition, I have made quite some fruitful collaborations with Dr. Martin Head-Gordon, Dr. Rommie Amaro, Dr. Julie Forman-K. and experimental collaborators from the Anti Viral Drug Discovery Center. These great PIs and their enthusiastic lab members have also gave me power to work on the most ambitious research projects together.

I truly want to thank all the friends and roommates I had during my PhD journey: Daisong Pan, Yang Lyu, Zhibo Yang, Xingzhi Wang, Yuchen Liu, Haoran Liao, Taige Wang, Jianbo Jin, Yehao Qiu, and many other people I have met. Our shared experiences—cooking, unforgettable trips, gaming, and discussions on an array of topics from science and technology to politics, history, art, and philosophy—infused my Ph.D. life with vibrancy and unforgettable memories.

Finally I want to say thank you to my parents and grandparents. You kindled my passion in science in my childhood, and supported my decision to pursue PhD abroad. Your full support on my life and your encouraging words on my research progress have gave me strength to overcome all the difficulties I encountered during my PhD, and allow me to think big and think deep.

Despite the challenging times we have encountered, such as the pandemic and academic interruptions, I remain grateful. I consider myself fortunate to be in an environment where the primary focus can be on scientific discoveries. The pandemic, despite its hardships, presented opportunities, particularly in the realm of drug discovery. I sincerely believe I am living in an extraordinary age.

Lastly, to future readers of this dissertation, thank you for your interest and time. I hope the research within this document enriches your understanding of our field and inspires you going forward.

CHAPTER 1

Introduction

Computational chemistry has undergone a transformation under modern developments of data and machine learning (ML) algorithms, illuminating new paths to tackle theoretical chemistry problems that previously did not have a straightforward solution.[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] Unlike the clear layout and formulation of model equations to describe a molecular system, and a well-defined progression of numerical steps to reach solution (deductive reasoning), ML instead uses a "non-algorithmic" formulation to "train" parametric or non-parametric models from a great abundance of data[16], so that new predictions can be made from data inference (inductive reasoning). Because the fundamentals of modern ML are statistics and pattern matching, the generalizability of a ML model depends on both the amount of data used for training models, and the formulation of data representations and/or model architectures that better exploits that data.

Given the accumulation of more and better quality data, along with improvements in theoretical methodologies and hardware that can generate high quality data with unprecedented speed, modern ML methods are often based on large neural network (NN) models with a huge amount of parameters, which distinguishes them from early-day ML methods that rely on much simpler model architectures.[17] However, that does not mean modern ML methods are universally better than simple models and statistical approaches for chemical applications, especially when existing data are scarce, or generation of new data is time consuming.[16] Depending on the amount of data available, some of the most successful and popular ML approaches used for chemical science applications today also include statistical learning methods that are based on decision trees together with ensemble learning and gradient boosting, exemplified by random forest[18] and XGBoost[19] approaches; kernel based methods that have evolved from support vector machines (SVMs)[20] to kernel ridge regression (KRR)[21] and Gaussian process regression (GPR)[22]; and deep learning models spanning from fully-connected NNs to more advanced architecture designs, including convolutional neural networks (CNNs)[17], recurrent neural networks (RNNs)[23], graph neural networks (GNNs)[24], and most recently large pretrained transformers like ProteinBERT[25], TAPE[26] and ESM[27].

Along with ML methods that were mainly developed by statisticians and computer scientists, more domain-specific ML research in chemistry includes better representation of the molecular systems in terms of more physical descriptors of molecules and model architectures that encompass constraints based on physical laws.[28] For example, physically inspired feature representations may find less discrepancy between molecules used for training and those that extrapolate to real-world predictions. The recognition of fundamental symmetries of a certain molecular property, such as invariance or equivariance under translation, rotation or permutation operations may also greatly help reducing data requirements for training, and improving performance of the final model.[29, 30, 31] Finally, ML models designed with physics in mind are usually more interpretable in regards the theories behind them compared to "black-box" ML models that we neither fully understand nor have close control over in how they work.[32]

While ML methods have achieved better accuracy and improved efficiency for molecular characterization applications, molecule generation is a whole new area of research that has been fully empowered by the accumulation of data and advancement of deep learning generative models.[5] ML models can now "learn" from distributions of data with properties of interest, and propose new molecules with similar properties. When combined with an evaluation metric, a feed-back loop can be established that automates the generation of new molecules with even better properties[33, 34, 35, 36], revolutionizing areas where designing new molecules with better characteristics is the main purpose, such as materials science, catalyst design and drug discovery.

In this introduction chapter of my dissertation, I will describe my research work at the frontline of combining data-driven models with physics-based approaches for molecule characterization and generation. First, I will give theoretical background about machine learning and how it has been broadly applied in the molecular sciences. I will then introduce physics-inspired ML models for molecule characterization, exemplified by the prediction of experimental nuclear magnetic resonance (NMR) chemical shifts for aqueous proteins [37] and computed chemical shifts from theoretical models for arbitrary molecular system with H, C, N and O atoms.[38] I will also briefly discuss how ML can be used for automatically analyzing transmission electron microscope (TEM) images for nanoparticles and identifying morphologies of synthesized products without human intervention.[39] Next I discuss how proteins can be rebuilt with higher quality using only torsion angles, by predicting more accurate bond lengths and bond angles from learning their correlations with torsion angles and amino acid sequence.[40] Finally, I present how a deep learning generative model when combined with physical docking simulations and empirical drug-likeliness metrics, can generate chemically diverse molecules that have optimized shape and non-bonded interactions with a target protein pocket of interest.[36]

1.1 Machine learning models for chemistry[†]

Many modern ML models are based upon artificial neural networks (ANN), architectures that mimic the neuronal connections in a mammalian brain, that perform successive linear transformations and non-linear activations of input data to approximate an arbitrary function. In the context of molecular characterization, this architecture best describes a supervised learning task of predicting molecular properties from input representations of a molecular system, such as a protein, a crystal structure, or a drug molecule, etc. The term "supervised learning" corresponds to a set of labelled data on which the model is trained in order to minimize the error for the mappings between input features and the target outputs.

The most basic computing element of an ANN, the simple perceptron[41], is capable of performing linear or logistical regression and classification with appropriate activation functions (Figure 1.1), and can perform Boolean operations such as the simple OR and AND functions. A slightly more complex architecture is needed when executing the exclusive XOR function that requires a pre-processing "hidden" layer between the input and output layers to appropriately define the linear decision boundaries that separates its solution space. Such early shallow ANN architectures, using everything from hand-crafted features to molecular structures, have successfully predicted more than 20 different types of physiochemical properties of a molecule, such as water solubility, Henry's law constant, heats of formation, and crystal packing.[42]

The universal approximation theorem states that a single hidden layer with many simple perceptrons and suitable activation functions can represent any function of $\{\mathbf{x}\}$ to predict $f(y|\{\mathbf{x}\})$, regardless of complexity or how non-linear is its solution space. However what is not guaranteed is that there is a universal procedure for how to *learn* the transformation $\{\mathbf{x}\}\rightarrow f(y|\{\mathbf{x}\})$ using a single layer architecture, nor what is the best feature representation of $\{\mathbf{x}\}$ to ensure that it will perform well on previously unobserved target function data. The DL network learns the input-output representations by minimizing a loss function through adjustments of the weights that connect the neuronal nodes of its architecture. Hence most of the recent excitement in ML is centered around deep learning (DL) architectures, an approach that replaces a single hidden layer with many, many hidden layers, each composed of many artificial neurons, and the rapidly evolving meta-heuristics used to calculate with them.

One of the most classical examples of a DL architecture are the CNNs that were originally introduced and popularized by LeCun for handwriting and other image recognition tasks[43]. CNNs are neural networks that use convolution operations (in actuality a cross-correlation operation) in place of general matrix multiplication (as in standard ANNs) in at least one of their layers. During the learning process the convolutional layers typically generate multiple feature maps that when aggregated together represent new formulations of the input data.

[†] Partly reproduced with permission from: Haghighatlari M*, Li J*, Heidar-Zadeh F*, Liu Y, Guan X, Head-Gordon T. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. *Chem.* **2020**, 6(7):1527-1542. (* denotes equal first authors)

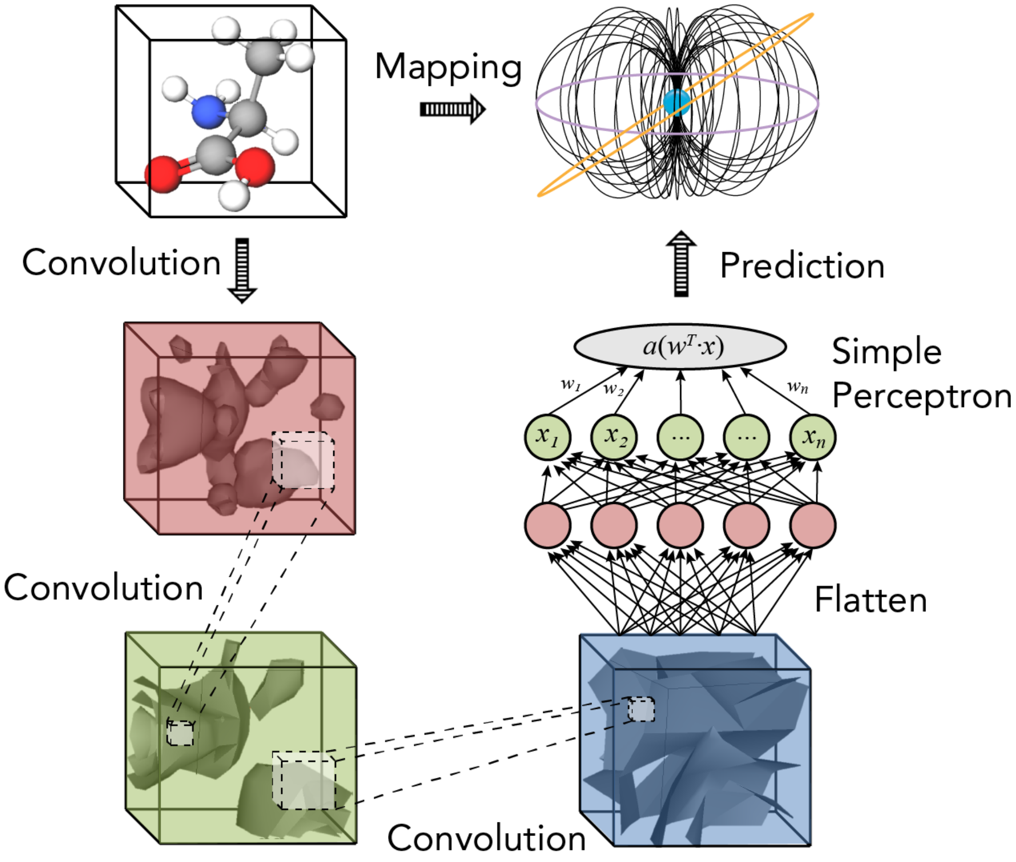


Figure 1.1: *Illustration of using convolution neutral networks to predict magnetic properties of a molecule.* The use of a simple perceptron of an ANN as part of the transformation of a 3D representation of a molecule with convolutions accumulated through layers of a CNN to yield atomic magnetic properties in a molecular framework, such as a chemical shift prediction.

Figure 1.1 pictorially displays how the input data is "transformed" by the processing units of the convolution through many layers. In order to aid the learning strategy of a CNN, the sparser L connections between L convolutional layers have been recently replaced by a "denser" network of L(L+1)/2 direct connections, also known as a "DenseNet"[44]. In this case the feature maps of all preceding layers are used as inputs to a current convolution layer, and its own resulting feature maps are then used as inputs into all subsequent layers of the deep layered architecture.

In addition, many sequential models that have been designed for natural language or temporal data processing have also proven useful for chemistry applications. Small molecules can be represented as 1-dimensional strings using SMILES[45], InChi[46] or SELFIES[47], and protein sequences can be resembled as another "language" using the 20 amino acids as its alphabet. Therefore, sequential models like RNNs and transformers that can easily capture spatial-temporal correlations are highly appropriate for molecular systems. The latent representations of the molecules after encoding by an NN can be used for predicting atom or molecule level properties. Alternatively, the latent representations can be used to decide which token is the next one by sampling from a predicted token distribution using partial strings as input, and predict (or sample) tokens in an auto-regressive manner. Most recently, large language models that have human-level understanding of the physical world are created using combinations of un-supervised pre-training and fine-tuning techniques such as reinforcement learning. Many of these techniques are also applicable in the chemical sciences, enabling sequential generative models to design new molecules with certain properties that we need.

It might be most intuitive to consider molecular structures as graphs, in which each atom defines a node and bonds between atoms defines edges. For 3-dimensional conformations of molecules, they can also be considered as fully-connected graphs with atomic distances as edge features. Therefore, graph neural networks (GNNs) build physical inductive biases in the models to allow them to surpass other NN architectures in various molecular property prediction tasks, including QM properties such as energies and forces[48, 49, 50, 51, 52], physiochemical properties such as hydrophobicity [53, 54, 55] or toxicity [56, 57, 58]. Recently, building equivariance properties into GNNs has been proven to help reduce the amount of data needed for training, allowing models to be trained with more expensive *ab inito* data. Physics-constrained equivarient models have been introduced, such as NequIP that based upon spherical harmonics for convolution filters[29], NewtonNet that borrows NN design from Newton's equations[30], and PaiNN that directly builds equivariance into the model[31].

The primary distinction of a DL architecture is its much greater network capacity relative to early ANN's, and thus its greater advantage in handling much larger data sets than previously possible. The DL approach has also advanced through learning heuristics that are better established relative to early ANNs[59]: regularization through choice of appropriate loss functions, back-propagation and back-propagation through time, data augmentation using noise injection or non-linear transformations, and the use of dropout and batch normalization; adaptive learning strategies that bear strong equivalence to a Newton step using preconditioners that are combined with stochasticity in the gradients as per methods like RMSProp[60] and Adam[61]; and finally the finetuning of the "hyperparameters" in all of these learning choices through formulations of validation data sets and through methods such as early stopping and ensemble prediction.

As such, DL is ready for prime time in the chemical sciences as their architectures can be adapted to many types of problems, their hidden layers reduce the need for feature engineering, and they have benefited from several important regularizations that allows them to efficiently learn from high-dimensional data. At the same time DL approaches are not always suitable as general-purpose ML methods because they have orders of magnitudes more parameters to optimize and thus require much more expertise to tune (i.e. to set the architecture and optimize the hyperparameters), and especially because they require a very large amount of well-curated labelled data. We note that a DL model is characterized as being overfit when the test error increases from the minimum of the bias-variance trade-off curve[59], reaching a maximum when the DL model is merely interpolating on the training data. However, very recent work has shown that increasing model capacity beyond the point of interpolation results in improved performance for reasons that are not well understood.[62]

Alternatively, machine learning methods such as GPR and KRR can be traced back to the advent of Support Vector Machines (SVMs), which formulate a clever choice of kernel to capture the similarities of a collection of data points. If the optimal kernel is found, the simplest linear regression is sufficient to predict the target value from its input data using similarity to the input features of the training dataset. As such kernel methods are powerful supervised classifiers that optimize non-linear decision boundaries directly. They have been found to be superior to multiple linear regression and radial basis function neural networks when applied to chemical toxicity prediction for example[63]. More recently, KRR has realized excellent performance on regression prediction for molecular properties such as NMR chemical shifts for small molecules either in solution[64, 65] or in the solid state[66]. In this case the physical understanding of a chemical system helped in the creation of a reasonable kernel function. Specifically the SOAP kernel[67] is explicitly designed to faithfully represent an atomic environment of a molecule with uniqueness. Furthermore kernel methods naturally incorporate symmetry functions for which it is often desirable to enforce translational or rotational invariances that may be relevent to the chemical prediction[67, 68].

While kernel methods work very well in practice, and are robust against overfitting even in high-dimensions, they are tricky to tune due to the importance of picking the right kernel, and if the kernel function is not smooth enough in the space of the atomic environment, the resulting kernel-based method will suffer from outliers in the training dataset that will degrade prediction performance. They also require the storage of and operation on all of the support or feature vectors, which can be prohibitive for application to large datasets. Especially in the case of KRR and GPR, because the similarity kernel needs to be applied between the pairwise features with all data examples in the training dataset, its unfavorable scaling with the number of training examples prevents it from benefiting from large datasets, although a number of strategies including parallelization can mitigate their cost[69].

Often statistical models such as decision trees are preferred over kernel methods as they are more robust to outliers, are much more computationally scalable, and do not require the luck of finding the kernel function as they quite naturally model non-linear decision boundaries thanks to their hierarchical structure.[59] In a statistical learning model such as decision trees, training comprises the optimal splitting of the features driven by a decrease in the maximum entropy loss function from information theory. Decision tree models are equally suited for big or small datasets because once the cutting points have been identified, the application of the algorithm to new data is just a constant of time. The classification or regression prediction from a statistical model are also easier to interpret compared to other parametric models, because the splitting reveals causal relationships which are easy to understand and explain. For example, by analyzing the number of times each feature is used in a node to split data in a decision tree, we can understand the relative importance of different features and to determine those that are most influential for the predicted property[70]. But of all machine learning techniques, decision trees are amongst the most prone to overfitting because we cannot know *a priori* how to formulate the smallest tree that completes the learning task, and all practical implementations must mitigate this challenge. This has led to specialized approaches such as pruning or bagging and boosting to prevent overfitting, as well as other regularization techniques also developed in deep learning such as early stopping and ensemble learning for which decision trees benefit from becoming “random forests”[59]. Statistical learning models have been successfully applied to molecular property predictions, as in the example of modeling of different quantitative structure-activity relationships with a decision tree based on random forest optimization[71], and are starting to replace the use of SVMs in classification tasks more broadly.

1.2 Physics-inspired machine learning models for chemical shifts prediction

NMR chemical shifts are highly sensitive observables dependent on 3-dimensional atomic details and environment of a molecular system. Due to its high accuracy from experimental measurements and high sensitivity to molecular compositions, detailed geometries, short-range atomic environments, and long-range ring currents, NMR chemical shifts have been widely used by experimentalists to model structures or validate structure correctness in systems ranging from small organic molecules to large biopolymers, both in crystalline form and in solution phase.[72, 73, 74, 75, 76]

However, it is not always straightforward to directly utilize the information contained in an NMR spectrum. Atoms that have similar atomic environments will be hard to distinguish using chemical shifts alone, and it is usually a painstaking step to assign various peaks from an NMR spectrum to individual atoms in a molecule. Furthermore, measured NMR signals only reflect highly averaged properties from rapidly evolving dynamic conformations of the molecules, which makes it more challenging to elucidate the relationship between measured chemical shifts and a static molecular picture. Therefore, computational methods are indispensable tools to help scientists utilize experimental NMR chemical shifts more efficiently and easily.

When a molecule with atoms that have non-zero nuclear spin is under a uniform magnetic field, the orbital motions of the electrons create an electric current which generates an induced magnetic field according to the formula[77]

$$\boldsymbol{B}_{in}(\boldsymbol{r}) = \frac{1}{c}\int d^3r'\boldsymbol{j}(r') \times \frac{(\boldsymbol{r}-\boldsymbol{r'})}{|\boldsymbol{r}-\boldsymbol{r'}|^3} \tag{1.1}$$

where $\boldsymbol{j}(r)$ is the electric current that scales with the external magnetic field $\boldsymbol{B}_{ext}$. The relationship between $\boldsymbol{B}_{in}$ and $\boldsymbol{B}_{ext}$ can therefore be described as:

$$\boldsymbol{B}_{in}(\boldsymbol{r}) = -\hat{\boldsymbol{\sigma}}(\boldsymbol{r})\boldsymbol{B}_{ext} \tag{1.2}$$

The magnetic shielding tensor elements σ_{ab} describe the influence of the external field on the induced field at different Cartesian directions, and can be evaluated by[78]:

$$\sigma_{ab} = \left.\frac{\mathrm{d}^2E(\boldsymbol{M}^A, \boldsymbol{B}_{ext})}{\mathrm{d}M_a^A\mathrm{d}B_b}\right|_{\boldsymbol{B}_{ext}=0,\boldsymbol{M}^A=0} \tag{1.3}$$

where E is system energy, $\boldsymbol{M}^A$ is the nuclear spin of nucleus A, "d" stands for total derivative, and a, b correspond to Cartesian indices.

Due to the ensemble averaging effect in most experimental NMR measurements, only the isotropic component of the chemical shielding tensor is observed. Therefore, the isotropic chemical shielding value σ is defined as the mean of the diagonal elements from the whole shielding tensor[78]:

$$\sigma(\boldsymbol{r}) = \frac{\mathrm{Tr}[\hat{\boldsymbol{\sigma}}(\boldsymbol{r})]}{3} \tag{1.4}$$

Finally, the most widely used concept of chemical shifts are defined as the chemical shielding offset between the nucleus under investigation and that of a standard substance, usually tetramethylsilicane (TMS) for ^{1}H and ^{13}C chemical shfts. Namely,

$$\delta = \sigma_{ref}(\boldsymbol{r}) - \sigma(\boldsymbol{r}) \tag{1.5}$$

These set of equations completely describe the relationship between a fixed molecular configuration $\{\boldsymbol{r}\}$ and the chemical shifts for each atom in the molecule, with good accuracy assuming the energy of the molecule E can be calculated with high quality theory. However two issues arise that make NMR CS predictions difficult: (1) First, highly accurate calculations of chemical shielding tensors from first principles can be prohibitively expensive for large molecular systems of interest, and (2) the calculations may not account for external factors such as solvation effects, vibrations in the molecular geometries, and the conformational flexibility of the molecules. Various empirical tools have been developed to mitigate these difficulties. [79, 80, 38] Specifically, I will present two physics-inspired data driven approaches to improve the accuracy and efficiency of chemical shift predictions and their applications in real world structure identification.

In the first application we explore how to improve accuracy and robustness for chemical shift calculations for aqueous proteins. Existing methods like SPARTA+[81] and the SHIFTX+ component of SHIFTX2[82] uses backbone geometric features such as bond torsion angles and residue biological similarity properties like block substitution matrix (BLOSUM) numbers to predict chemical shifts using simple feed-forward neural networks or Bagging and Boosting ensemble models. SHIFTX2 also took advantage of existing experimental databases and introduced an alignment-and-transfer technique to fully exploit sequence homology and make more accurate predictions.[82] Specifically, engineered features that are insensitive to the instantaneous conformations of a protein in the thermalized ensemble while still having sufficient discriminatave power for chemical shifts of nuclei in different atomic environments are ideal to build into a model that reproduces experimental chemical shifts in a complex aqueous environment. For example, categorical features like whether a residue is involved in a hydrogen bond, or the secondary structure type of a residue are "stable" features that are consistent among different conformations of the structure in the ensemble while distinct enough to aid the differentiation of chemical shifts for different atoms in the molecule.

Built upon this idea, Chapter 2 describes the UCBShift method that combines a tree-based ensemble regression model using stable features extracted from high quality protein X-ray crystal structures and a structure homology based alignment method.[37] I have used numerical and categorical features extracted from the geometries and biophysical properties of a tripeptide to predict backbone atoms and sidechain β-carbon chemical shifts of the central residue. The features were designed with uniqueness, universality and ease of calculation in mind, which include backbone and sidechain torsion angles, BLOSUM numbers that represent likelihood of residue substitution, secondary structure assignments using DSSP program[83], hydrogen bond geometries, as well as many non-linear transformations and combinations of geometric features inspired by physical models. These features are unique to the protein structure itself, as well as being invariant to the translation and rotation and hence being universal for the molecular fragment of interest. We have also taken into account crystal waters in the evaluation of relevant features like hydrogen bond geometries to allow more faithful reflection of the physical nature of solvation. Furthermore, we have redesigned a structure based alignment module to directly transfer the chemical shifts from the experimental database to the query protein when the structure homology is sufficiently high. The resulting UCBShift algorithm achieved significantly lower root-mean-square-error (RMSE) when compared to SPARTA+ and SHIFTX2 when evaluated on an independently generated test dataset, and has shown superior capability to select the native state from misfolded decoys in two test examples.

Such feature extracted methods are ideal for the molecular systems they have been designed for, but they are not expected to generalize to a different system. For example, the chemical shift predictors for aqueous proteins cannot be used to predict chemical shifts for small organic molecules in the gas phase, for example. By contrast, QM methods can be much more rigorous and predictive regardless of the application domain. The gold standard for QM is coupled-cluster theory with single and double excitation and perturbative-approximated triple excitations [CCSD(T)] when combined with a complete basis set (CBS) or a sufficiently large one for various molecular properties calculations, including magnetic properties such as chemical shieldings.[80, 84, 85] The issues with this method is its extremely unforgiving scaling with system size, which make any calculation using CCSD(T)/CBS essentially impractical for systems with more than 10 heavy atoms using current day computers. Alternatively, one could employ density functional theory (DFT) using gauge-including projector-augmented waves (GIPAW) to calculate chemical shieldings.[86] While these methods have a much more acceptable scaling, their accuracy are far from ideal (likely due to the fact that Kohn-Sham theory does not address magnetization properties), which also limits their applications.[87, 88, 89] Some existing work has tried to bridge these two QM approaches using ML, mostly by exploring the Δ-learning idea of using low level DFT to learn CCSD(T) as a correction. The work by Unzueta, et al. uses the atomic environment vector (AEV) as geometry-dependent features to predict the difference between a cheap DFT calculation with small basis set and one with the same DFT theory but a large basis set[79], while Büning and Grimme take one step further in predicting the difference between DFT and CCSD(T) level chemical shifts for small molecules.[80] In a different route, Guan, et al. employed the idea of transfer learning to train their model first on DFT calculated chemical shifts and then fine tune with experimental data.[90]

In chapter 3, I present our development of the iShiftML method, a machine learning approach that better connects to the physical nature of chemical shieldings, predicting close to CCSD(T) level chemical shieldings from intermediate QM magnetic tensor elements calculated from a cheap DFT calculation, (i.e. ωB97X-V/pcSseg-1).[38] The diamagnetic (DIA) and paramagnetic (PARA) shielding tensor elements from the DFT calculations, together with geometric-dependent atomic environment vectors (AEVs) were used as input features into a feed forward NN to predict the weights of these low level matrix elements, and the final chemical shielding prediction is given by multiplying these predicted weights with the matrix elements. When trained through a novel active learning workflow that progressively adds most underrepresented data containing more heavy atoms into the training dataset, we can consistently improve model performance while keeping minimal cost on calculating chemical shieldings using high level QM approaches.

Since the features utilized in this model are quite universal, our model illustrated exceptional generalizability in subsequent comparisons to experiments for systems that are much bigger than any molecule included in the training dataset. We demonstrated near 50% error reduction in predicting chemical shifts compared with the same low-level theory used to provide input features for the model on natural products that contain several dozens of heavy atoms and complex ring systems, although our model has been trained with at most 7 heavy atoms without complicated bonding. This superiority was also reflected in our model's capability to recognize the correct structure of a natural product from its diastereomers, by inspecting the difference between calculated chemical shifts for the assumed structures and the experimentally measured chemical shifts.[38]

1.3 Unsupervised characterization of transmission electron microscope images

In addition to performing regression tasks on numerical observables, data-driven models can also enable new applications such as image analysis. Transmission electron microscopy (TEM) is a widely used experimental technique to characterize the morphology of nano-scale materials, such as metal nanoparticles (mNP). The rapid advancement in automated high-throughput electron microscopy enables the collection of TEM images at an unprecedented speed, which allows the scale of NP shape characterization to increase by orders of magnitude, which far exceeds what a human analyst is able to process.[91, 92, 93, 94] Hence, data-driven methodologies for image analysis can play an essential role here.

A fully automated algorithm for TEM image analysis should be able to perform two tasks: particle detection, which includes identification and segmentation of particles of interest, and information extraction, which generally involves the characterization of the shapes of the particles based on their aspect ratios and other such attributes. Multiple algorithms exist for automatic particle detection.[93, 95, 96, 97, 98, 99, 100, 101, 102, 103] However, these algorithms do not explicitly consider the differentiation between particles with different shape attributes, which might limit their application to only analyzing homogeneous samples and not suitable for analyzing TEM images that may contain particles with multiple shape morphologies. Existing algorithms that are capable of classifying particles based on shape attributes usually requires predefined shape categories and is therefore not applicable to unexplored NP systems with the synthesis outcomes unknown *a priori*.[97, 98, 101, 102] Therefore, an unsupervised algorithm that is capable of automatically analyzing TEM images without human intervention would be beneficial to address the pressing needs in greater automation in the material sciences.

In a collaboration with Wang and Alivisatos, we have developed the AutoDetect-mNP algorithm that automatically extracts particle shape features from a collection of TEM images and clustering the particles based on shape attributes, requiring minimum human input in the process.[39] Figure 1.2 describes the process of the algorithm. We have provided flexibility in the particle detection algorithm, because different NP systems might require different particle detection attributes to achieve optimal performance. Multiple shape features were extracted from the recognized particles, including solidity, convexity, area, eccentricity, aspect ratio, and circularity.[104, 105, 97, 106, 107] Based on the convexness of the recognized particles, some irregular particles were further broken down using the ultimate erosion of convex shapes (UECS) algorithm[102], while the other “particles” due to failed detections were discarded.

Particles were then classified based on the four geometrical shape descriptors using K-means clustering followed by naive Bayes classification. The K-means step determines the mean and standard deviations of the descriptors in each category, and each extracted particle was classified into one of the K classes that maximizes the joint probability of all geometric descriptors belonging to that class. To enable fully automatic clustering of particles into the

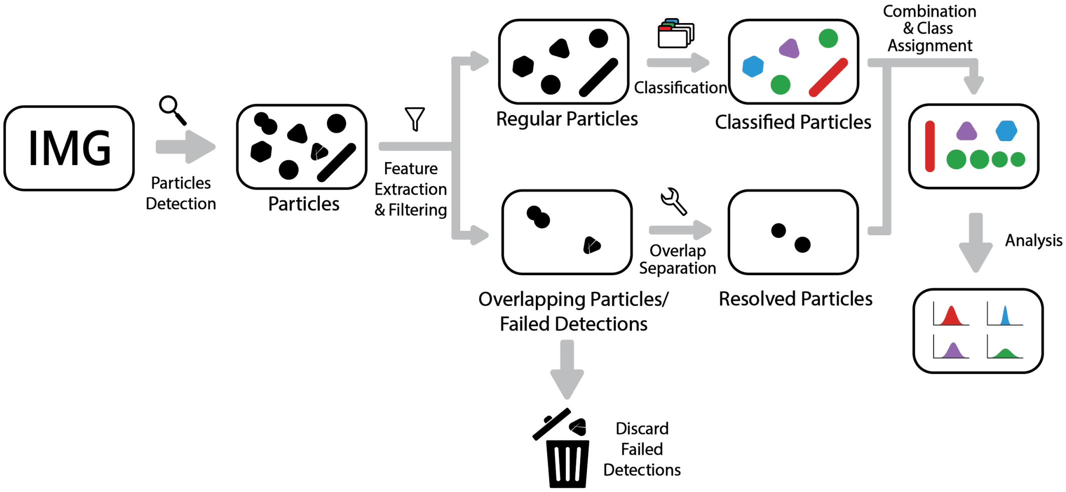


Figure 1.2: *Scheme of the AutoDetect-mNP algorithm.* The algorithm can be divided into four parts: particle detection, feature extraction, filtering and resolution of irregularly shaped particles, and classification of particle shapes.[†]

optimal number of classes, we have also introduced the P_{max} metric to determine K values according to data distributions. P_{max} is defined as the maximum off-diagonal element from the confusion matrix $\boldsymbol{P}(K)$, where

$$P_{ij} = \frac{E[p_j(x_i)]}{E[p_i(x_i)]}, i, j \leq K \tag{1.6}$$

with $p_i(x)$ describes the likelihood of data point x belongs to class i. Therefore, P_{max} is the maximum relative probability that any particle is erroneously classified into a different category. The optimal K that minimizes P_{max} means that the number of categories identified from the dataset ensures any pariticle has a minimal chance of being assigned to the wrong category. Finally, the number of particles in each category was calculated, and features of the classified particles were further analyzed.

The novelty and uniqueness of our algorithm resides in the fact that we are able to cluster and classify extracted particles in a completely unsupervised and automatic fashion such that there is no need to pre-define the shape categories that exist in the dataset. We also do not require training a parametric model using human-labelled data that are both labor intensive and prone to biases. The results of the AutoDetect-mNP algorithm was first evaluated on two data sets of gold nanoparticles (AuNPs) with different shapes comprised of short and long rods and triangular prisms. When compared with two previous

[†] Figure reproduced with permission from: Wang X*, Li J*, Ha HD, Dahl JC, Ondry JC, Moreno-Hernandez I, Head-Gordon T, Alivisatos AP. AutoDetect-mNP: an unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. *Jacs Au.* **2021**, 1(3):316-327. (* denotes equal first authors)

methods[97, 102], our algorithm has showed to be the most accurate in determining the counts of NPs with different morphologies from a collection of 20 TEM images, and our method is the only one to differentiate between long and short rods. Furthermore, we also found our algorithm runs significantly faster than previous methods, which enables our method to be applied to larger datasets and images with higher resolution to avoid information loss. We have also demonstrated successful applications of AutoDetect-mNP to more challenging NP systems beyond AuNPs, which showed that our algorithm can serve as a general and unbiased metric for reporting shape and shape attribute distributions of various mNP systems using TEM or similar image acquisition strategies. This development facilitates future research on automated platforms for high-throughput mNP synthesis and time resolved characterization of mNP reactivity.

1.4 Better modelling of protein structures using machine learning

Deep learning modelling of molecular structures has been an especially important and popular application of data-driven models in molecular science. This is illustrated by the success of AlphaFold and similar models in predicting protein structures with atomistic accuracy from 1-dimensional amino acid sequences. The strength of these data-driven models present in solving problems that are difficult to tackle with traditional methods.[108, 7, 109, 27] Unfortunately, many of these applications utilize “black box” models that do not have a straightforward explanation on why they work, nor can we obtain physical insight from these models.

For the purpose of protein structure modelling, internal coordinates of a protein provide a compact description of the protein geometry, which is frequently used during geometry optimizations and NMR structure determination.[110] Due to the relatively small variations of bond lengths and bond angles in a protein, they are often treated as fixed values in many protein modelling applications, such as fragment-based protein folding and loop modelling.[111, 112] This treatment indeed reduces the complexity of the problem, but also risks decreased accuracy when building models with less degrees of freedom.

These missing correlations among internal coordinates can be largely recovered using a deep neural network, and by considering these subtle correlations between internal coordinates, more accurate protein structures can be reconstructed from torsion angles alone.

As I will show in Chapter 4, we found there are meaningful correlations between bond lengths and/or bond angle variations with backbone torsion angles and amino acid type. The Int2Cart method predicts bond lengths and bond angles of a protein given the backbone torsion angles and amino acid types.[40] Specifically, the DL model exploits a statistical analysis of a large collection of protein structures in the Protein DataBank, from which we found strong correlations between backbone bond lengths and bond angles with backbone torsion angles ϕ, ψ and ω. Even though the observed distributions of backbone bond lengths and bond angles did not show significant variation between amino acid types, further break down of the bond length and angle distributions as a function of both backbone torsions and residue type exhibit notable variations across *all* 20 amino acids. To capture deeper correlations beyond a single residue, we have trained a recurrent neural network (RNN) model to predict accurate bond lengths and bond angles in the backbone by taking torsion angles and amino acid types as input. The predictions were then used to rebuild the protein structures. We found that we could significantly reduce error in bond length and bond angle predictions when compared with that assuming fixed values for these geometric characters, but more importantly the reconstructed protein structures illustrate less structural discrepancy when compared with the crystal structures, both in terms of reduced backbone root-mean-square-deviation (RMSD), and diminshed loss of secondary structure. We have also showed the superiority of our model compared to a method that uses bond geometries dependent on localized torsion angle and amino acid information on a single residue.[113]

Finally, the usefulness of the Int2Cart algorithm was demonstrated in two applications. We first show that the agreement between predicted bond lengths and bond angles from Int2Cart and those from an AlphaFold2 modelled structure indicate model quality in terms of AlphaFold2's internal confidence estimations, which supports using Int2Cart for structure validation and refinement. Second, the method was tested on an intrinsically disordered protein (IDP) ensemble showing that it is able to decrease structure modelling error as well as reducing steric clashes. This result indicates the method is generalizable to a different type of proteins than folded ones.

1.5 Deep generative models for drug discovery

The discovery of new molecules has traditionally relied on general chemical principles aided by domain expertise. Even though various theoretical models and computational tools exist, they typically aid as opposed to replicate an expert chemist's capability to design new molecules. This was the state of affairs until recently with the advent of deep learning models that also demonstrate versatility in generative tasks such as new molecule generation.[5] Given that molecular structures can be represented as computer-readable data (strings, graphs), combined with the accumulation of molecules in public databases, it is less surprising that a well-trained ML algorithm is able to generate novel molecules that have never been proposed before; this is especially the case for drug discovery.[114, 115, 116, 117, 118, 119, 120, 121, 122, 123]

The true power of these generative models is their potential to be fine-tuned using external feedbacks. Reinforcement learning (RL) is a quite successful technique to improve a generative model when coupled with an external evaluator. In terms of drug discovery, generated molecules can be evaluated through *in silico* docking simulations, which evaluates the potency the molecule binds with a given biomolecule target. Due to the computational cost of docking simulations, many previous algorithms have trained a separate machine learning model to predict the binding potency of a generated molecule to a biological target using experimental datasets specifically for the target.[118, 119, 120, 122, 123] However, these networks may not generalize to unseen ligand structures, hence leading the optimization through RL in the wrong direction. It is therefore worthwhile to explore how to build more physical evaluation metrics into the RL workflow to improve the quality of generated molecules in terms of binding potency to the specified protein target.

In the final chapter of this dissertation, I will present the iMiner approach that combines deep RL with real-time molecular docking for *de novo* molecule generation based on a biological target.[36] By encoding molecular structures as one-dimensional strings, we trained a generative model based on a recurrent neural network (RNN) that creates brand new molecules inside the drug-like chemical space. To improve the chemical validity of the generated molecules, previous methods that use SMILES representations require special treatment in the model.[118] Instead, we have used a special encoding of SELFIES strings to make sure generated molecules are valid through a better chemistry syntax.[47] Docking simulations with AutoDock vina[124] were used to provide more physical estimations of binding affinity between the proposed molecules and a predefined protein pocket. We have also developed an empirical drug-likeness metric to pose constraints on where the generative model should explore. When combined with an advanced RL algorithm to train the model, new molecules generated have noticeable improvement in terms of docking score while maintaining drug-like properties. Furthermore, we have collected generated molecules with better binding potency throughout the training process, instead of using molecules from the final trained model. Further down-selection steps based on a separate docking software Glide[125] to find consensus, and other physical, biological or empirical filters were applied[126, 127], which ensures none of the generated drug molecules will be obvious non-starters during the drug development process.

The whole iMiner workflow was validated using SARS-COV-2 main protease (mPro) as a working example, for which we suggested 54 molecules as potential drug candidates. The efficient exploration of chemical space with better binding potency to a given target is demonstrated by comparing molecules proposed by iMiner with those collected in the COVID moonshot project, a crowd-sourced effort in finding an effective mPro inhibitor.[128] Our molecules present optimized structures and precise non-bonded interactions with the binding pocket of mPro according to docking poses, while being chemically diverse in scaffold and molecular composition. We would expect the method to work well for developing inhibitor molecules for other types of proteins relevant to diseases, or exploring variations on existing molecules that can improve potency. Therefore, it can potentially accelerate hit discovery and hit-to-lead optimizations in drug discovery pipeline.

1.6 References

[1] Stefan Chmiela, Huziel E Sauceda, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. sgdml: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45, 2019.

[2] Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):2903, 2019.

[3] Silvia Amabilino, Lars A Bratholm, Simon J Bennie, Alain C Vaucher, Markus Reiher, and David R Glowacki. Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality. *The Journal of Physical Chemistry A*, 123(20):4486–4499, 2019.

[4] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current opinion in structural biology*, 61:139–145, 2020.

[5] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[6] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[7] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[8] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.

[9] Sebastian Brickel, Akshaya K Das, Oliver T Unke, Haydar T Turan, and Markus Meuwly. Reactive molecular dynamics for the [cl–ch3–br]- reaction in the gas phase and in solution: a comparative study using empirical and neural network force fields. *Electronic Structure*, 1(2):024002, 2019.

[10] Khosrow Shakouri, Jorg Behler, Jorg Meyer, and Geert-Jan Kroes. Accurate neural network description of surface phonons in reactive gas–surface dynamics: N2+ ru (0001). *The journal of physical chemistry letters*, 8(10):2131–2136, 2017.

[11] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[12] Jie Li, Kochise C Bennett, Yuchen Liu, Michael V Martin, and Teresa Head-Gordon. Accurate prediction of chemical shifts for aqueous protein structure on “real world” data. *Chemical science*, 11(12):3180–3191, 2020.

[13] Shuai Liu, Jie Li, Kochise C Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. Multiresolution 3d-densenet for chemical shift prediction in nmr crystallography. *The journal of physical chemistry letters*, 10(16):4558–4565, 2019.

[14] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[15] Mojtaba Haghighatlari, Gaurav Vishwakarma, Mohammad Atif Faiz Afzal, and Johannes Hachmann. A physics-infused deep learning model for the prediction of refractive indices and its use for the large-scale screening of organic compound space. 2019.

[16] Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. Learning to make chemical predictions: the interplay of feature representation, data, and machine learning methods. *Chem*, 6(7):1527–1542, 2020.

[17] Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[18] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[19] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[20] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[21] Vladimir Vovk. Kernel ridge regression. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 105–116. Springer, 2013.

[22] Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

[23] Larry Medsker and Lakhmi C Jain. *Recurrent neural networks: design and applications*. CRC press, 1999.

[24] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[25] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[26] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[27] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022:500902, 2022.

[28] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022.

[29] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[30] Mojtaba Haghighatlari, Jie Li, Xingyi Guan, Oufan Zhang, Akshaya Das, Christopher J Stein, Farnaz Heidar-Zadeh, Meili Liu, Martin Head-Gordon, Luke Bertels, et al. Newtonnet: A newtonian message passing network for deep learning of interatomic potentials and forces. *Digital Discovery*, 1(3):333–343, 2022.

[31] Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pages 9377–9388. PMLR, 2021.

[32] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[33] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.

[34] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[35] Woosung Jeon and Dongsup Kim. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports*, 10(1):22104, 2020.

[36] Jie Li, Oufan Zhang, Fiona L Kearns, Mojtaba Haghighatlari, Conor Parks, Xingyi Guan, Itai Leven, Rommie E Amaro, and Teresa Head-Gordon. Reinforcement learning with real-time docking of 3d structures to cover chemical space: Mining for potent sars-cov-2 main protease inhibitors. *arXiv preprint arXiv:2110.01806*, 2021.

[37] Jie Li, Kochise C Bennett, Yuchen Liu, Michael V Martin, and Teresa Head-Gordon. Accurate prediction of chemical shifts for aqueous protein structure on "real world" data. *Chemical science*, 11(12):3180–3191, 2020.

[38] Jie Li, Jiashu Liang, Zhe Wang, Aleksandra L Ptaszek, Xiao Liu, Brad Ganoe, Martin Head-Gordon, and Teresa Head-Gordon. Highly accurate prediction of nmr chemical shifts from low-level quantum mechanics calculations using machine learning. *arXiv preprint arXiv:2306.08269*, 2023.

[39] Xingzhi Wang, Jie Li, Hyun Dong Ha, Jakob C Dahl, Justin C Ondry, Ivan Moreno-Hernandez, Teresa Head-Gordon, and A Paul Alivisatos. Autodetect-mnp: an unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. *Jacs Au*, 1(3):316–327, 2021.

[40] Jie Li, Oufan Zhang, Seokyoung Lee, Ashley Namini, Zi Hao Liu, João MC Teixeira, Julie D Forman-Kay, and Teresa Head-Gordon. Learning correlations between internal coordinates to improve 3d cartesian coordinates for proteins. *Journal of Chemical Theory and Computation*, 2023.

[41] F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.*, 65(6):386–408, 1958.

[42] Jyrki Taskinen and Jouko Yliruusi. Prediction of physicochemical properties based on neural network modelling. *Adv. Drug Deliv. Rev.*, 55(9):1163–1183, 2003.

[43] Y LeCun, B Boser, J S Denker, D Henderson, R E Howard, W Hubbard, and L D Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.*, 1(4):541–551, dec 1989.

[44] G Huang, Z Liu, G Pleiss, L Van Der Maaten, and K Weinberger. Convolutional Networks with Dense Connectivity. *IEEE Trans. Pattern Anal. Mach. Intell.*, page 1, 2019.

[45] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[46] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7(1):1–34, 2015.

[47] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[48] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[49] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[50] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[51] Renjie Liao, Zhizhen Zhao, Raquel Urtasun, and Richard S Zemel. Lanczosnet: Multi-scale deep graph convolutional networks. *arXiv preprint arXiv:1901.01484*, 2019.

[52] Hiroyuki Shindo and Yuji Matsumoto. Gated graph recursive neural networks for molecular property prediction. *arXiv preprint arXiv:1909.00259*, 2019.

[53] Chao Shang, Qinqing Liu, Ko-Shin Chen, Jiangwen Sun, Jin Lu, Jinfeng Yi, and Jinbo Bi. Edge attention-based multi-relational graph convolutional networks. *arXiv preprint arXiv: 1802.04944*, 2018.

[54] Xiaofeng Wang, Zhen Li, Mingjian Jiang, Shuang Wang, Shugang Zhang, and Zhiqiang Wei. Molecule property prediction based on spatial graph embedding. *Journal of chemical information and modeling*, 59(9):3817–3828, 2019.

[55] Gary Bécigneul, Octavian-Eugen Ganea, Benson Chen, Regina Barzilay, and Tommi S Jaakkola. Optimal transport graph neural networks. 2020.

[56] Youjun Xu, Jianfeng Pei, and Luhua Lai. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of chemical information and modeling*, 57(11):2672–2685, 2017.

[57] Michael Withnall, Edvard Lindelöf, Ola Engkvist, and Hongming Chen. Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction. *Journal of cheminformatics*, 12(1):1–18, 2020.

[58] Hao Yuan and Shuiwang Ji. Structpool: Structured graph pooling via conditional random fields. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

[59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[60] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA Neural networks Mach. Learn.*, 4(2):26–31, 2012.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv*, 2014.

[62] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci.*, 116(32):15849–15854, 2019.

[63] C Y Zhao, H X Zhang, X Y Zhang, M C Liu, Z D Hu, and B T Fan. Application of support vector machine (SVM) for prediction toxic activity of different data sets. *Toxicology*, 217(2-3):105–119, 2006.

[64] Matthias Rupp, Raghunathan Ramakrishnan, and O Anatole Von Lilienfeld. Machine learning for quantum mechanical properties of atoms in molecules. *J. Phys. Chem. Lett.*, 6(16):3309–3313, 2015.

[65] Will Gerrard, Lars A Bratholm, Martin J Packer, Adrian J Mulholland, David R Glowacki, and Craig P Butts. IMPRESSION–prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.*, 2020.

[66] Federico M Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nat. Commun.*, 9(1):4501, 2018.

[67] Albert P Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87(18):184115, 2013.

[68] Jörg Behler and Michele Parrinello. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.*, 98(14):146401, apr 2007.

[69] Yang You, James Demmel, Cho-Jui Hsieh, and Richard Vuduc. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems. *Proc. 2018 Int. Conf. Supercomput.*, pages 307–317, 2018.

[70] Leo Breiman. *Classification and regression trees.* Routledge, 2017.

[71] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[72] Neil E Jacobsen. *NMR data interpretation explained: understanding 1D and 2D NMR spectra of organic compounds and natural products.* John Wiley & Sons, 2016.

[73] Peter J Hore. *Nuclear magnetic resonance.* Oxford University Press, USA, 2015.

[74] Andrew E Derome. *Modern NMR techniques for chemistry research.* Elsevier, 2013.

[75] Kurt Wüthrich. Protein structure determination in solution by nmr spectroscopy. *J. Biol. Chem.*, 265(36):22059–22062, 1990.

[76] Dominique Marion. An introduction to biological nmr spectroscopy. *Mol. Struct. Proteom.*, 12(11):3006–3025, 2013.

[77] Chris J Pickard and Francesco Mauri. All-electron magnetic response with pseudopotentials: Nmr chemical shifts. *Physical Review B*, 63(24):245101, 2001.

[78] Graham A Webb. *Modern magnetic resonance: Part 1: Applications in chemistry, biological and marine sciences, Part 2: Applications in medical and pharmaceutical sciences, Part 3: Applications in materials science and food science.* Springer Science & Business Media, 2007.

[79] Pablo A Unzueta, Chandler S Greenwell, and Gregory JO Beran. Predicting density functional theory-quality nuclear magnetic resonance chemical shifts via δ-machine learning. *J. Chem. Theory Comput.*, 17(2):826–840, 2021.

[80] Jürgen Gauss. Analytic second derivatives for the full coupled-cluster singles, doubles, and triples model: Nuclear magnetic shielding constants for bh, hf, co, n 2, n 2 o, and o 3. *J. Chem. Phys.*, 116(12):4773–4776, 2002.

[81] Yang Shen and Ad Bax. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, 48(1):13–22, 2010.

[82] Beomsoo Han, Yifeng Liu, Simon W Ginzinger, and David S Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, 50(1):43, 2011.

[83] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[84] Andrew M Teale, Ola B Lutnæs, Trygve Helgaker, David J Tozer, and Jürgen Gauss. Benchmarking density-functional theory calculations of nmr shielding constants and spin–rotation constants using accurate coupled-cluster calculations. *J. Chem. Phys.*, 138(2):024111, 2013.

[85] Caspar Jonas Schattenberg and Martin Kaupp. Extended benchmark set of main-group nuclear shielding constants and nmr chemical shifts and its use to evaluate modern dft methods. *J. Chem. Theory Comput.*, 17(12):7602–7621, 2021.

[86] Chris J Pickard and Francesco Mauri. All-electron magnetic response with pseudopotentials: Nmr chemical shifts. *Physical Review B*, 63(24):245101, 2001.

[87] Trygve Helgaker, Michal Jaszunski, and Kenneth Ruud. Ab initio methods for the calculation of nmr shielding and indirect spin-spin coupling constants. *Chem. Rev.*, 99:293–352, 1999.

[88] James R Cheeseman, Gary W Trucks, Todd A Keith, and Michael J Frisch. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J. Chem. Phys.*, 104(14):5497–5509, 1996.

[89] Denis Flaig, Marina Maurer, Matti Hanni, Katharina Braunger, Leonhard Kick, Matthias Thubauville, and Christian Ochsenfeld. Benchmarking hydrogen and carbon nmr chemical shifts at hf, dft, and mp2 levels. *J. Chem. Theory Comput.*, 10(2):572–578, 2014.

[90] Yanfei Guan, SV Shree Sowndarya, Liliana C Gallegos, Peter C St John, and Robert S Paton. Real-time prediction of 1 h and 13 c chemical shifts with dft accuracy using a 3d graph neural network. *Chem. Sci.*, 12(36):12012–12026, 2021.

[91] Sean K. Mulligan, Jeffrey A. Speir, Ivan Razinkov, Anchi Cheng, John Crum, Tilak Jain, Erika Duggan, Er Liu, John P. Nolan, Bridget Carragher, and Clinton S. Potter. Multiplexed TEM Specimen Preparation and Analysis of Plasmonic Nanoparticles. *Microscopy and Microanalysis*, 21(4):1017–1025, 2015.

[92] Yong Zi Tan, Anchi Cheng, Clinton S. Potter, and Bridget Carragher. Automated data collection in single particle electron microscopy. *Microscopy (Oxford, England)*, 65(1):43–56, 2016.

[93] Martin Schorb, Isabella Haberbosch, Wim J.H. Hagen, Yannick Schwab, and David N. Mastronarde. Software tools for automated transmission electron microscopy. *Nature Methods*, 16(6):471–477, 2019.

[94] Stephen D. House, Yuxiang Chen, Rongchao Jin, and Judith C. Yang. High-throughput, semi-automated quantitative STEM mass measurement of supported metal nanoparticles using a conventional TEM/STEM. *Ultramicroscopy*, 182:145–155, 2017.

[95] John M Sosa, Daniel E Huber, Brian Welk, and Hamish L Fraser. Development and application of MIPAR™: a novel software package for two- and three-dimensional microstructural characterization. *Integrating Materials and Manufacturing Innovation*, 3(1):123–140, 2014.

[96] Lionel Cervera Gontard, Dogan Ozkaya, and Rafal E. Dunin-Borkowski. A simple algorithm for measuring particle size distributions on an uneven background from TEM images. *Ultramicroscopy*, 111(2):101–106, 2011.

[97] Christine R. Laramy, Keith A. Brown, Matthew N. O'Brien, and Chad A. Mirkin. High-Throughput, Algorithmic Determination of Nanoparticle Structure from Electron Microscopy Images. *ACS Nano*, 9(12):12488–12495, 2015.

[98] Luca Boselli, Hender Lopez, Wei Zhang, Qi Cai, Valeria A. Giannone, Jingji Li, Alirio Moura, João M. de Araujo, Jennifer Cookman, Valentina Castagnola, Yan Yan, and Kenneth A. Dawson. Classification and biological identity of complex nano shapes. *Communications Materials*, 1(1):1–12, 2020.

[99] Catherine K Groschner, Christina Choi, and M. C. Scott. Methodologies for Successful Segmentation of HRTEM Images via Neural Network. *arXiv*, 2020.

[100] Yanjun Qian, Jianhua Z. Huang, Xiaodong Li, and Yu Ding. Robust nanoparticles detection from noisy background by fusing complementary image information. *IEEE Transactions on Image Processing*, 25(12):5713–5726, 2016.

[101] Chiwoo Park, Jianhua Z. Huang, David Huitink, Subrata Kundu, Bani K. Mallick, Hong Liang, and Yu Ding. A multistage, semi-automated procedure for analyzing the morphology of nanoparticles. *IIE Transactions (Institute of Industrial Engineers)*, 44(7):507–522, 2012.

[102] C. Park, J. Z. Huang, J. X. Ji, and Y. Ding. Segmentation, Inference and Classification of Partially Overlapping Nanoparticles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):669–681, 2013.

[103] Chiwoo Park and Yu Ding. Automating material image analysis for material discovery. *MRS Communications*, 9(2):545–555, 2019.

[104] Eric A. Grulke, Stephen B. Rice, Jin Cheng Xiong, Kazuhiro Yamamoto, Tae Hyun Yoon, Kevin Thomson, Meghdad Saffaripour, Greg J. Smallwood, Joshua W. Lambert, Arnold J. Stromberg, Ryan Macy, Nicholas J. Briot, and Dali Qian. Size and shape distributions of carbon black aggregates by transmission electron microscopy. *Carbon*, 130:822–833, 2018.

[105] Eric A. Grulke, Xiaochun Wu, Yinglu Ji, Egbert Buhr, Kazuhiro Yamamoto, Nam Woong Song, Aleksandr B. Stefaniak, Diane Schwegler-Berry, Woodrow W. Burchett,

Joshua Lambert, and Arnold J. Stromberg. Differentiating gold nanorod samples using particle size and shape distributions from transmission electron microscope images. *Metrologia*, 55(2):254–267, 2018.

[106] Ming Kuei Hu. Visual Pattern Recognition by Moment Invariants. *IRE Transactions on Information Theory*, 8(2):179–187, 1962.

[107] J. P. MacSleyne, J P Simmons, and M. De Graef. On the use of 2-D moment invariants for the automated classification of particle shapes. *Acta Materialia*, 56(3):427–437, 2008.

[108] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Ž´, Alexander W R Nelson, Alex Bridgland, and Others. Improved protein structure prediction using potentials from deep learning. *Nature*, pages 1–5, 2020.

[109] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[110] Jon Baker, Don Kinghorn, and Peter Pulay. Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules. *J. Chem. Phys.*, 110(11):4986–4991, 1999.

[111] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier, 2004.

[112] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[113] Donald S Berkholz, Maxim V Shapovalov, Roland L Dunbrack Jr, and P Andrew Karplus. Conformation dependence of backbone geometry in proteins. *Structure*, 17(10):1316–1325, 2009.

[114] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954. PMLR, 2017.

[115] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[116] Akshay Subramanian, Utkarsh Saha, Tejasvini Sharma, Naveen K Tailor, and Soumitra Satapathi. Inverse design of potential singlet fission molecules using a transfer learning based approach. *arXiv preprint arXiv:2003.07666*, 2020.

[117] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[118] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[119] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning*, pages 3668–3679. PMLR, 2020.

[120] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[121] Alex Zhavoronkov, Vladimir Aladinskiy, Alexander Zhebrak, Bogdan Zagribelnyy, Victor Terentiev, Dmitry S Bezrukov, Daniil Polykovskiy, Rim Shayakhmetov, Andrey Filimonov, Philipp Orekhov, et al. Potential 2019-ncov 3c-like protease inhibitors designed using generative deep learning approaches. 2020.

[122] Navneet Bung, Sowmya R Krishnan, Gopalakrishnan Bulusu, and Arijit Roy. De novo design of new chemical entities for sars-cov-2 using artificial intelligence. *Future medicinal chemistry*, 13(06):575–585, 2021.

[123] Jannis Born, Matteo Manica, Joris Cadow, Greta Markert, Nil Adell Mill, Modestas Filipavicius, Nikita Janakarajan, Antonio Cardinale, Teodoro Laino, and María Rodríguez Martínez. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on sars-cov-2. *Machine Learning: Science and Technology*, 2(2):025024, 2021.

[124] Oleg Trott and Arthur J Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, jan 2010.

[125] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, David E Shaw, Perry Francis, and Peter S Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, mar 2004.

[126] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in

drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[127] Jayme L Dahlin, J Willem M Nissink, Jessica M Strasser, Subhashree Francis, LeeAnn Higgins, Hui Zhou, Zhiguo Zhang, and Michael A Walters. Pains in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging hts. *Journal of medicinal chemistry*, 58(5):2091–2113, 2015.

[128] Hagit Achdout, Anthony Aimon, Elad Bar-David, Haim Barr, Amir Ben-Shmuel, James Bennett, Melissa L Bobby, Juliane Brun, Sarma BVNBS, Mark Calmiano, et al. Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv*, 2020.

CHAPTER 2

Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data[†]

2.1 Introduction

Chemical shifts are a readily obtainable NMR observable that can be measured with high accuracy for proteins, and sensitively probe the local electronic environment that can yield quantitative information about protein secondary structure[1, 2, 3], estimation of backbone torsion angles,[4, 5] or measuring the exposure of the amino acid residues to solvent.[6] But in order to take full advantage of these high quality NMR measurements, there is a necessary reliance on a computational model that can relate the experimentally measured NMR shifts to structure with high accuracy. Existing methods for chemical shift prediction rely on extensive experimental databases together with useful heuristics to rapidly, but non-rigorously, simulate protein chemical shifts. As of yet, quantum mechanical methods which would in principle provide more rigor to chemical shift prediction are still in progress.[7]

The heuristic chemical shift back-calculators are formulated as approximate analytical models such as shAIC,[8] PPM,[9] and PPM_One,[10] empirical alignment-based methods such as SHIFTY[11] and SPARTA,[12, 13] 3D representations for machine learning of chemical shifts in solid state NMR for small molecules,[14, 15] and feature-based methods including SHIFTCALC,[16] SHIFTX,[17] PROSHIFT,[18] Camshift, [19] and SPARTA+;[20] in the case of SHIFTX2 [21] both alignment and features are utilized. Some of the most successful alignment-based methods rely on the fact that proteins with similar sequences will also share similar structures which lead to their exhibiting similar chemical shifts. This idea was first exploited by the program SHIFTY[11] which "transferred" the chemical shifts

[†] Reproduced with permission from: Li J, Bennett KC, Liu Y, Martin MV, Head-Gordon T. Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data. *Chemical science.* **2020**, 11(12), 3180-3191.

from known sequences in the database to the query sequence based on a global sequence alignment. Higher accuracy was achieved in the formulation of SHIFTY+ by replacing the global sequence alignment with a local sequence alignment, and is included in the most recent chemical shift prediction program SHIFTX2.[21] Alignment-based methods in general yield predictions with higher accuracy when a good sequence homologue is found in the database, and the constant increase in the number of sequences and associated chemical shifts deposited into the Biomolecular Magnetic Resonance Bank (BMRB),[22] suggests that a similar sequence to the query sequence will continue to increase steadily.

On the other hand, methods that are based on sequence alignments will by definition fail if sequence similarity between the query sequence with any sequence in the database is too low, as well as the possibility that similar sequences can adopt very different structural folds.[23] For query sequences with low sequence identity, the analytical or feature extraction methods predict secondary chemical shifts (*i.e.* from a random coil reference[24]) by providing data input formulated as hypersurfaces of structural data attributes such as backbone ϕ, ψ angles and hydrogen bonding derived from X-ray structures or calculated from quantum mechanics, or physical data observables such as ring currents,[25, 26] electric field effects,[27] or Lipari–Szabo order parameters,[28] that can be generated from easily parsed computational models. These feature extracted data are then used to establish empirical hypersurfaces such as used in SHIFTS[7, 29] and Camshift, [19] or to train a machine learning algorithm in the cases of PROSHIFT,[18] SPARTA+,[20] and the SHIFTX+ component of SHIFTX2.[21]

In the evaluation of chemical shifts calculators like SPARTA+ and SHIFTX2, extensive effort has been put into making the test dataset as clean as possible. However, for chemical shifts of real-world proteins, large deviations from predicted values are typically considered as "outliers" and removed from the test dataset post-prediction.[20] The definition of outliers can be arbitrary and results in higher test performance than operating on a real-world data set that may not necessarily be plagued by experimental errors. In this work we examine the current performance of feature extracted methods represented by SPARTA+, as well as the combination of sequence alignment and feature extracted method as implemented in SHIFTX2 on a randomly selected test dataset with minimal data filtering. First we assess the performance in terms of root mean square error (RMSE) with respect to experimental chemical shifts for SPARTA+ and SHIFTX2 when evaluated on a fully independent set of test proteins with high-resolution (<2.4 Å) X-ray structures. We use chemical shift data deposited in the BMRB, in which protein chemical shifts have been re-referenced with respect to high-resolution X-ray structures using RefDB developed by Wishart and co-workers.[30] We also assess SPARTA+ and SHIFTX2 performance when eliminating putative outliers as determined by test data set filtering using PANAV,[13] or removing test proteins with >30% sequence identity to the training set, which are dataset preparation steps that provide a more fair comparison to the two standard chemical shift calculators.

Furthermore, we show that higher accuracy for chemical shifts can be achieved with an enhanced hybrid algorithm, UCBShift, that makes predictions using machine learning on a more extended set of extracted features and transferring experimental chemical shifts from a database by utilizing both sequence and structural alignments. Although we find that we can realize better RMSEs if we also filter the different aspects of the test data, the resulting UCBShift chemical shift prediction method on all of the data including outliers yields RMSEs that will be at the level of 0.31, 0.19, 0.84, 0.81, 1.00 and 1.81 ppm for H, Hα, C', Cα, Cβ and N respectively when evaluated on any independently generated test sequence.

The improved chemical shift performance of UCBShift can be utilized in several predictive contexts such as detection of erroneous chemical shift assignments and errors in reference shifts, to refine single folded structures[31] or refinement of ensembles such as we have done in our Experimental Inferential Structure Determination Method (EISD)[32, 33] for folded and intrinsically disordered proteins (IDPs). To illustrate the usefulness of the UCBShift method, we consider the discrimination among alternative folds or selection of native structures among structural "decoys". We determine that UCBShift is able to identify the native structure of two different proteins that comprise two different decoy data sets with certainty by examining the correlation between predicted and experimental chemical shifts, with significant improvement over SPARTA+ and SHIFTX2 prediction methods when sequence or structural homology is available.

2.2 Methods

Preparation of training and new testing datasets

A high-quality database of protein structures and associated accurately referenced chemical shifts are crucial for composing a machine learning approach that can make reliable predictions of the chemical shifts, and for faithfully comparing the performance of existing alternative approaches such as SPARTA+ and SHIFTX2. Several publicly available data sources, including the SPARTA+ training set and the training and testing set for SHIFTX+, were combined into a single training dataset that captures the structure and chemical shift relationship. Since all of these data were used in the development of the original SPARTA+ and SHIFTX2 methods, it ensures that corrections for chemical shift reference values were included in our dataset as well.

Unlike previous incarnations of these data sets, which stripped out all presence of crystal waters and ligands, we generated a data set that retained the small molecules in the crystal structures. Our hypotheses is that for crystal waters especially, they often are highly conserved and functional, and are likely highly populated even in solution NMR experiments.[34, 35] Any reported hydrogen atoms in the Research Collaboratory for Structural Bioinformatics protein databank (RCSB or PDB) structures[36] were removed and a systematic approach for adding a complete set of hydrogens used the program REDUCE[37] to keep consistency in the structural data used across all approaches. To ensure more robust training, for each atom type, residues with chemical shifts deviating from the average by 5 standard deviations and residues that DSSP[38] failed to generate secondary structure predictions were removed, which accounted for the removal of 40, 5, 18, 147 and 1 training examples for H, Hα, Cα, Cβ and N shifts, respectively. When stereochemically inequivalent Hα were present, their shifts were averaged. In the creation of data for each of the individual atom types, any residues that do not have recorded chemical shifts in the database were eliminated.

Before excluding redundant chains from the database, there were altogether 852 proteins in the training dataset. Duplicate chains were identified and excluded from our dataset: two chains are regarded as duplicates if the sequences and their structures are exactly the same, or eliminating the shorter sequence if it is a sub-sequence of a longer sequence (which is kept). However, 32 chains in the SPARTA+ dataset were retained because although they had identical sequences, they were found to have different structures and thus different chemical shifts. After excluding the duplicate chains by this prescription, the number of protein chains in the training dataset decreased to 647. The filtering of the training dataset based on RCI-S^2 [39] in principle excludes flexible residues whose chemical shifts are harder to predict. We did not exclude training data based on RCI-S^2 as was done in some other chemical shift predictors, because a complete training set that covers different prediction difficulties is crucial for obtaining reliable performance for real-world applications. Table 2.1 reports the total number of training data examples for each of the different atom types. In addition, the RefDB database,[30] which is a database for re-referenced protein chemical shifts assignments extracted and corrected from BMRB, was also compiled for the alignment-based chemical shifts prediction.

Table 2.1: *Total number of training and testing examples for chemical shift prediction for each atom type.* The training set is comprised of the combination of the SPARTA+ training set and the training and testing set for SHIFTX+, and removing all redundant chains. We have developed a new test set comprised of 200 high-resolution proteins with chemical shifts available from RefDB; the test data eliminates duplicate chains, and residues with no deposited chemical shift values. The LH-Test set refers to the subset of the total set of test proteins with only low sequence homology to other proteins such that sequence or structural homology cannot be exploited. We also created two curated test sets which additionally exclude paramagmetic proteins, some Hα chemical shifts that have calculated ring current effect exceeding 1.5 ppm, and "outliers" detected by the PANAV program [13]. Further information is provided in Methods and Appendix.

	# of PDBs	**H**	**Hα**	**C$'$**	**Cα**	**Cβ**	**N**
Train	647	72894	56149	58228	79611	70621	74896
Test	200	19120	11727	8231	13140	10139	15374
Test (curated)	200	18494	11240	7861	12533	9883	14610
LH-Test	100	8634	4979	3332	5685	4278	6576
LH-Test (curated)	100	8606	4950	3331	5251	4201	6480

Since the training dataset in Table 2.1 covers all of the data from SPARTA+ and

SHIFTX2, a separate test dataset needs to be prepared for a fair comparison of all of the chemical shift programs. Therefore, 200 proteins with high-resolution (<2.4 Å) X-ray structures and with chemical shifts available in the RefDB were selected at random to form a separate test set that do not share the exact same sequence as training structures. These structures were downloaded from RCSB and again hydrogens were added with the REDUCE algorithm.[37] Erroneous chemical shifts assignments were removed from this test dataset, which include 9 (H, Cβ, and N) chemical shifts that were significantly offset from the random coil average; 8 Cβ chemical shifts from cysteines that show strong disagreement with their expected chemical shifts under their oxidation state in the crystal structures, and all C' chemical shifts from 3 proteins that are anticorrelated with predictions from SPARTA+, SHIFTX2 and UCBShift (see Appendix for details). It is essential to remove these chemical shifts from the test set because of evidence for the existence of experimental or recording errors in these data; but no further processing was done on the test set so that it is a good representation of a "real-world" application.

A more carefully "curated" test dataset based on these 200 proteins was also prepared, which additionally exclude paramagmetic proteins, some Hα chemical shifts that have calculated ring current effect exceeding 1.5 ppm, "outliers" detected by the PANAV program,[13] and chemical shifts corrected by PANAV that are different from their original values by more than 0.3 ppm for hydrogens, 1.0 ppm for carbons and 1.5 ppm for nitrogen. These additional test filters are similar to the procedures used by SPARTA+ and SHIFTX2 in preparing their test datasets.[20, 21] A complete list of the 200 testing proteins are given in the Appendix (Table 2.B.1). As is inevitable, some of these 200 proteins share high sequence identity with some of the training data, so we also generate test datasets after filtering out proteins with >30% sequence identity to yield a low-homology test set (LH-Test) with 100 test proteins (Table 2.1).

Machine learning for chemical shift prediction

The new UCBShift chemical shift prediction program is composed of two sub-modules: the transfer prediction module (UCBShift-Y) that utilizes sequence and structural alignments to "transfer" the experimental chemical shift value to the query example, and a machine learning module (UCBShift-X) that learns the mapping between the feature extracted data to the experimental chemical shift in the training data. The overall structure of the algorithm is depicted in Fig. 2.1.

UCBShift-Y module. UCBShift-Y is similar in spirit to the SHIFTY+ component of SHIFTX2, in that the experimental chemical shift for a given atom type in a given residue can be transferred to the query protein when the sequence of the protein in the database is highly similar or even identical to the sequence of the query protein. However, instead of relying on the sequence alignment alone, we have developed an algorithm that relies on both sequence alignment and structural alignment, which allows for proteins that are highly related in structure but remotely related in sequence to be utilized. The use of structural alignments

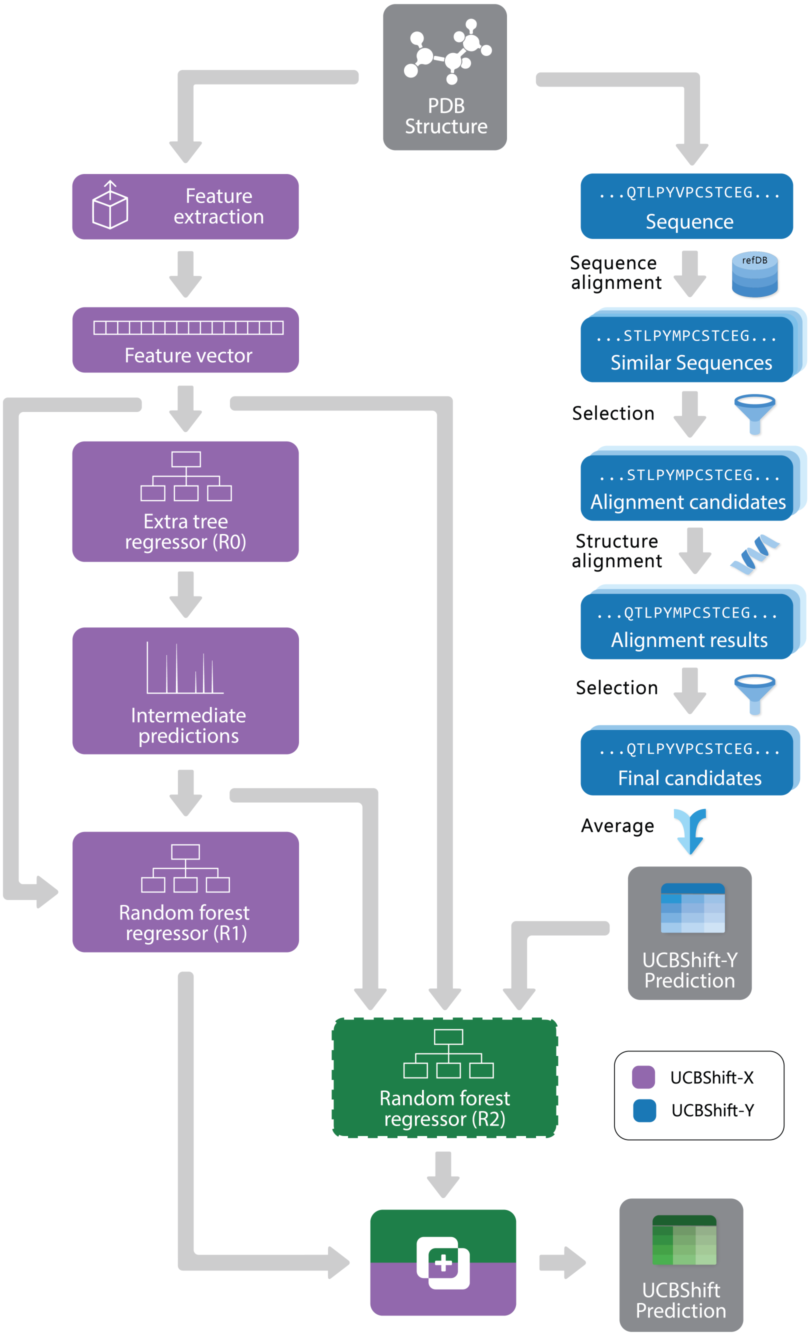


Figure 2.1: *The overall design of the UCBShift chemical shift prediction algorithm.* It combines both a transfer prediction module that relies on both sequence and structural alignments, and a machine learning module that trains a tree regression model on augmented feature extracted data.

also prevents proteins that have high sequence similarity but low structural homology to mislead an algorithm to erroneous chemical shifts transfers.

For the UCBShift-Y module, a query sequence is first aligned with all sequences in the RefDB database using the local BLAST algorithm,[40] and the PDB files for all sequences generating significant matches are further aligned with the query PDB structure using the mTM-align algorithm.[41] The alignments were further filtered to only keep those alignments that have TM score greater than 0.8 and an RMSD with the query protein structure that is smaller than 1.75 Å. For each of the aligned PDB sequences, its best alignment with the RefDB sequence is determined using the Needleman–Wunsch alignment.[42] If the residues are exactly the same, the shifts from RefDB are directly transferred to the target; otherwise, the secondary chemical shifts from RefDB are transferred to account for the different chemical shift reference states for different amino acids. To be more specific, the target shift for atom A and residue I is calculated from the matching residue J:

$$\delta_{\mathrm{I,A}} = \delta_{\mathrm{rc,I,A}} + (\delta_{\mathrm{J,A}} - \delta_{\mathrm{rc,J,A}}) \tag{2.1}$$

where $\delta_{\mathrm{rc,I,A}}$ and $\delta_{\mathrm{rc,J,A}}$ are the random coil shifts for atom type A in residue I and J, respectively, and $\delta_{\mathrm{J,A}}$ is the chemical shift for the matching residue in the database.

When multiple significant structural alignments exist for a given residue, the secondary shifts from these references are averaged with an exponential weighting w_{I},

$$w_{\mathrm{I}} = e^{5(S_{\mathrm{NA}} \times S_{\mathrm{TM}}) + B_{\mathrm{IJ}}} \times \mathbf{1}(B_{\mathrm{IJ}} \geq 0) \tag{2.2}$$

given by the normalized sequence alignment score, $S_{\mathrm{NA}} = S_{\mathrm{blast}}/\max(S_{\mathrm{blast}})$, where S_{blast} is the blast score of the matching sequence, and $\max(S_{\mathrm{blast}})$ is the maximum blast score from all blast hits; the structure alignment TM score S_{TM} is the pairwise TM score between the query structure and the aligned structure, and B_{IJ} is the substitution likelihood between the residue in the query sequence and the residue in the matching sequence using the BLOSUM62 substitution matrix.[43] Weights with negative substitution scores are set to zero.

UCBShift-X module. The UCBShift-X module requires the formulation of feature extracted data of a given atom type in a query residue, and the ability to map the feature extracted data to the chemical shift value during the training. Similar to the SPARTA+ program or for the SHIFTX+ component of SHIFTX2, we have developed residue-specific features for the query residue and the previous and next residues to the query residue, but we have included more features and polynomial transformations of the features to improve prediction. The feature extracted data generated from the PDB structures of individual residues include:

- 20 numbers representing the score for substituting the residue to any other amino acid, and taken from the BLOSUM62 substitution matrix[43]
- Sine and cosine values of the ϕ and ψ dihedral angles at the residue. Taking the sine and cosine values of the dihedral angles prevents the discontinuity when the dihedral

angle goes from +180° to -180°. For the undefined dihedral angles, for example the ϕ angle of a residue at the N-terminus and the ψ angle of a residue at the C-terminus, both the sine and cosine values were set to zero.

- A binary number indicating whether χ_1 or χ_2 dihedral angles for the side chain exists (existence indicator), and the sine and cosine values of these angles when they are defined for the same reasons described for backbone dihedral angles.

- Existence indicators and geometric descriptors for the hydrogen bond between the amide hydrogen and carboxyl oxygen, and between the Cα hydrogen and a carboxyl group (so called α-hydrogen bonds). For each position in the query residue that hydrogen bonds can form, a group of five numbers describe the properties of the hydrogen bond: a boolean number indicating its existence, the distance between the closest hydrogen bond donor–acceptor pair, the cosine values for the angles at the donor hydrogen atom and at the acceptor atom, and the energy of the hydrogen bond calculated with the DSSP model.[38] For the query residue, all hydrogen descriptors for amide hydrogen, carboxyl oxygen and α hydrogen are included, but only the carboxyl oxygen features are included for the previous residue, and the amide hydrogen features for the next residue. These add up to 25 hydrogen bond descriptors for any given residue.

- S^2 order parameters calculated by the contact model[28]

- Absolute and relative accessible surface area produced by the DSSP program.

- Hydrophobicity of the residue by the Wimley–White whole residue hydrophobicity scales.[44]

- Ring current effect calculated by the Haigh–Mallion model.[25, 26] For each training model for a specific atom type, the ring current for that atom type is included, while the ring currents for other atom types are excluded from the feature set.

- The one-hot representation of the secondary structure of the residue produced by DSSP program (composed of eight categories)

- Average B factor of the residue extracted from the PDB file.

- Half-sphere exposure of the residue[45]

- Polynomial transformations of some of the residue-specific features, such as the hydrogen bond distances (d_{HB}), by including d_{HB}^2, d_{HB}^{-1}, d_{HB}^{-2}, d_{HB}^{-3}, and the squares of the cosine values of the dihedral angles are also included as additional features. These polynomial quantities have been found to be correlated with secondary chemical shifts, and have occurred in several empirical formulas for calculating chemical shifts.[3, 46]

Unlike SPARTA+ and SHIFTX+, we have developed a pipeline with an extra tree regressor[47] followed by random forest regressor[48] as the machine learning based predictor shown in Fig. 2.1. Both the extra tree regressor and random forest regressor are ensembles of tree regressors that split the data using a subset of the features, and make ensemble-based predictions via a majority vote. However, extra tree regressors split the nodes in each tree randomly by selecting an optimal cut-point from uniformly distributed cut-points in the range of the feature, while the random forest regressors calculate the locally optimal cut in a feature by comparing the information entropy difference before and after the split. The random forest regressor learns based on the predictions from the first tree regressor and all the other input features, which can be regarded as a variant of the boosting algorithm,[49] since it learns from the mistakes the first predictor makes. The pipeline was first optimized using the TPOT tool[50] with 3-fold cross validation on the training set, and all the parameters were fine-tuned using a temporal validation dataset with 50 structures randomly selected from the training set. Because tree-based ensemble models are robust to the inclusion of irrelevant features,[51] feature selection was not performed. A more detailed analysis of the feature importance will be given in the Results.

Algorithmically, two separate random forest (RF) regressors are trained. The first RF regressor (R_1) only accepts features extracted from the structure and the prediction from the extra tree regressor, and the second RF regressor (R_2) additionally takes the secondary shift output from UCBShift-Y, together with additional scores and coverage indicating the quality of the alignments, and is trained using only a subset of the training data for which UCBShift-Y is able to make a prediction. Based on the availability of UCBShift-Y predictions, the final prediction of the whole algorithm is generated either by R_1 (when no UCBShift-Y predictions are available) or R_2 (when UCBShift-Y is able to make predictions). Finally, the random coil reference values are added back to the prediction to complete the total chemical shift prediction, *i.e.* the predictions are calculated with

$$\delta_{pred} = \begin{cases} f_{R_1}(X, f_{R_0}(X)) + \delta_{\mathrm{RC}} & \text{when UCBShift-Y generates no prediction} \\ f_{R_2}(X, f_{R_0}(X), S) + \delta_{\mathrm{RC}} & \text{when UCBShift-Y generates predictions} \end{cases} \tag{2.3}$$

where f_{R_0} represents the first-level extra tree regressor, f_{R_1} and f_{R_2} are the two second-level random forest regressors, X are all the features extracted from the structure, S are the predictions from UCBShift-Y and the identity scores, and δ_{RC} is the random coil chemical shift for the given residue.

2.3 Results

The performance of SPARTA+, SHIFTX2, and UCBShift are evaluated across the newly created test dataset of 200 proteins (test) and the subset of 100 low sequence homology with respect to the training set (LH-Test), each of which is uncurated or curated as described in Methods (Table 2.2). The mean average errors (MAEs) and correlation coefficients (R^2) are available in Table 2.C.1. In general, the performance of SPARTA+ is even across both the curated test and curated LH-Test datasets. The average RMSE error for SPARTA+ (and for all chemical shift predictors) on the uncurated Test and uncurated LH-Test datasets increases further, in which we provide the minimum error and the maximum error for each protein for SPARTA+ in graphical form in Fig. 2.C.1 and 2.C.2.

SHIFTX2 is seen to outperform SPARTA+ for chemical shift RMSE for all atom types on the curated dataset when there is high sequence homology for which it was designed, and it performs comparably to SPARTA+ on the curated data for target sequences with low sequence similarity to the training data. However, we find that the actual performance on curated data set is less accurate than the reported performance of the SHIFTX2 method.[21] One possible explanation is that a sequence similarity analysis revealed that out of the original 61 testing proteins of SHIFTX2, 4 proteins had 100% sequence alignment with a protein in the training dataset, sometimes under different identification numbers (Table 2.C.2). This problem of training data leakage into the testing data of the original SHIFTX2 method could be a non-trivial source of the better performance of SHIFTX2 reported in the literature. The protein-specific average RMSE error and the scatter plots for the SHIFTX2 predicted and experimental shifts are also given in Fig. 2.C.1 and 2.C.2 on the uncurated test dataset.

By comparison we find that filtering of the test set for outliers that disagree with the predictions, the elimination of paramagnetic proteins, and removing test shifts for hydrogen due to potentially inaccurate and large ring currents effects has more limited effect on prediction performance. To illustrate, the distributions of absolute errors from SPARTA+ for paramagnetic proteins and diamagnetic proteins in the Test dataset are shown in Fig. 2.C.3 The error distributions are not that different for H, Hα, Cβ, and N, and while paramagnetic proteins show higher prediction errors than diamagnetic proteins for the C' and Cα data types, they are not egregious errors.

We find that UCBShift outperforms SPARTA+ and the SHIFTX2 algorithm for chemical shift prediction RMSE when tested on the uncurated test data set, the more carefully curated test data, and regardless of the level of sequence homology. The protein-specific average RMSE error and the scatter plots for the UCBShift predicted and experimental shifts are also given in Fig. 2.C.1 and 2.C.2 on the uncurated Test dataset. Therefore UCBShift is more accurate for real-world applications, where the types of proteins may be more diverse than the test sets for SPARTA+ and SHIFTX2. In order to understand the improved performance of UCBShift in particular, we analyze the components of the algorithm including the UCBShift-X and UCBShift-Y modules, as well as the importance of the extracted features, utilizing the full test set of 200 proteins in more detail below.

Analysis of UCBShift-Y module

The major difference of our transfer prediction module (UCBShift-Y) in comparison with SHIFTY or SHIFTY+ is the inclusion of a structural alignment to select reference sequences for transfer of the chemical shift value. There is a trade-off between the coverage UCBShift-

Table 2.2: *Test set RMSE between predicted and experimental chemical shifts of SPARTA+, SHIFTX2, and UCBShift of relevant atom types found in proteins.* We compare the performance of SPARTA+[a], SHIFTX2[b], and UCBShift across an independently generated uncurated test dataset of 200 proteins that do not share the same sequence as the training set (Test) and a subset of 100 proteins with <30% sequence identity to the training set (LH-Test). We also compare the 3 methods against curated test data that removes "outliers" according to SHIFTX2 and SPARTA+ standards. Uncertainties are calculated from 50 random draws of 75% of the test data. All in units of ppm

Dataset	Test			LH-Test		
Atom type	SPARTA+	SHIFTX2	UCBShift	SPARTA+	SHIFTX2	UCBShift
H	0.51 ± 0.003	0.44 ± 0.003	$\mathbf{0.31 \pm 0.003}$	0.49 ± 0.004	0.49 ± 0.003	$\mathbf{0.45 \pm 0.004}$
Hα	0.27 ± 0.002	0.23 ± 0.003	$\mathbf{0.19 \pm 0.002}$	0.27 ± 0.003	$\mathbf{0.26 \pm 0.003}$	$\mathbf{0.26 \pm 0.003}$
C$'$	1.25 ± 0.01	1.16 ± 0.01	$\mathbf{0.84 \pm 0.01}$	1.16 ± 0.01	1.20 ± 0.01	$\mathbf{1.14 \pm 0.01}$
Cα	1.16 ± 0.01	1.05 ± 0.01	$\mathbf{0.81 \pm 0.01}$	1.13 ± 0.02	1.15 ± 0.01	$\mathbf{1.09 \pm 0.01}$
$C\beta$	1.35 ± 0.02	1.27 ± 0.03	$\mathbf{1.00 \pm 0.03}$	1.36 ± 0.05	1.37 ± 0.06	$\mathbf{1.34 \pm 0.05}$
N	2.72 ± 0.02	2.40 ± 0.02	$\mathbf{1.81 \pm 0.02}$	2.73 ± 0.02	2.73 ± 0.03	$\mathbf{2.61 \pm 0.02}$
Dataset	Test (curated)			LH-Test (curated)		
Atom type	SPARTA+	SHIFTX2	UCBShift	SPARTA+	SHIFTX2	UCBShift
H	0.49 ± 0.002	0.42 ± 0.002	$\mathbf{0.30 \pm 0.002}$	0.48 ± 0.003	0.47 ± 0.003	$\mathbf{0.43 \pm 0.003}$
Hα	0.26 ± 0.002	0.22 ± 0.002	$\mathbf{0.18 \pm 0.002}$	0.26 ± 0.003	0.25 ± 0.002	$\mathbf{0.24 \pm 0.003}$
C$'$	1.15 ± 0.009	1.06 ± 0.009	$\mathbf{0.77 \pm 0.008}$	1.16 ± 0.01	1.19 ± 0.01	$\mathbf{1.13 \pm 0.01}$
Cα	1.09 ± 0.008	0.98 ± 0.009	$\mathbf{0.76 \pm 0.009}$	1.08 ± 0.01	1.10 ± 0.01	$\mathbf{1.04 \pm 0.01}$
Cβ	1.17 ± 0.009	1.09 ± 0.02	$\mathbf{0.82 \pm 0.01}$	1.15 ± 0.01	1.15 ± 0.01	$\mathbf{1.12 \pm 0.02}$
N	2.59 ± 0.02	2.25 ± 0.02	$\mathbf{1.71 \pm 0.01}$	2.67 ± 0.02	2.66 ± 0.02	$\mathbf{2.55 \pm 0.02}$

[a] Reported SPARTA+ values from ref. 20: 0.49 ppm for H, 0.25 ppm for Hα, 1.09 ppm for C' , 0.94 ppm for Cα, 1.14 ppm for Cβ, and 2.45 ppm for N.

[b] Reported SHIFTX2 values from ref. 52: 0.17 ppm for H, 0.12 ppm for Hα, 0.53 ppm for C' , 0.44 ppm for Cα, 0.52 ppm for Cβ, and 1.12 ppm for N.

Y can achieve and the average accuracy of the prediction, requiring tuning the thresholds for accepting an imperfectly aligned protein as reference. Empirically we chose a relatively permissive threshold for sequence alignment to enable sequences that do not have much similarity with the query protein proceed to the next step in case it generates a good structure alignment. The thresholds for the TM score and RMSD in structure alignment were optimized during training to ensure the reference structures are close enough to the query structure. A TM score threshold of 0.8 was selected because NMR chemical shifts are sensitive to local structures, and only a well-aligned structure provides reliable chemical shift references.

We hypothesized that a structure-based alignment followed by sequence alignment would be more reliable since it would (1) allow for transferring shifts from structurally homologous proteins with low sequence identity, while also (2) ensuring that the transferred chemical shift values are not from a protein that has high sequence similarity but low structural homology with the query protein. This is confirmed in Fig. 2.2 which plots the difference of the RMSE on amide hydrogen chemical shift prediction between our UCBShift-Y and SHIFTY+ as a function of sequence identity. Plots for other atom types are given in Fig.2.D.1. Here the sequence identity is defined as the ratio of the number of matched residues to either the length of the query sequence or the length of the matched sequence, whichever is longer. Furthermore, the UCBShift results are reported with a specifically designed "test mode" which will not utilize sequences with more than 99% identity with the query sequence for making the prediction; this practice ensures the testing performance is a more realistic reflection of the actual performance when operating on input data which is not included in the search database. It is evident that on average predictions on query sequences with low sequence similarity but high structural homology are greatly improved with UCBShift-Y.

A particularly interesting example is the prediction for adenylate kinase (PDB ID: 4AKE)[53] and its mutant (PDB ID: 1E4V),[54] both of which are identical but for a single substitution of a valine for a glycine residue at position 10 (Fig. 2.3). Even with such high sequence identity, these two proteins adopt quite different tertiary structures with a backbone RMSD of 7.08 Å as can be seen from the overlay of their two structures in Fig. 2.3a. Hence while the experimental chemical shifts for these two proteins have a root-mean-square difference (RMSD) of 0.38 ppm for amide hydrogen shifts overall, the maximum H chemical shift difference is much larger at 1.34 ppm and is reflected in the surprisingly lower correlation (R-value = 0.86) between the amide hydrogen shifts for two proteins given the high sequence similarity (Fig. 2.3b). Therefore when using SHIFTY or SHIFTY+ for the 1E4V query sequence, the best sequence match will be 4AKE, thus increasing the chemical shift prediction error due to the huge structural deviation between the two proteins. Instead when using our UCBShift-Y module it selects two alternative proteins, 1AKE and 2CDN, which share an average sequence similarity of only 67% with the query protein. The correlation between the predicted 1E4V amide hydrogen shifts with UCBShift-Y which chooses references based on structural alignment and the experimental values are given in Fig. 2.3c, raising the R-value to 0.94 and lowering the RMSE to 0.25 ppm.

Fig. 2.D.2 summarizes the results of UCBShift-Y vs. SHIFTY+ for chemical shift

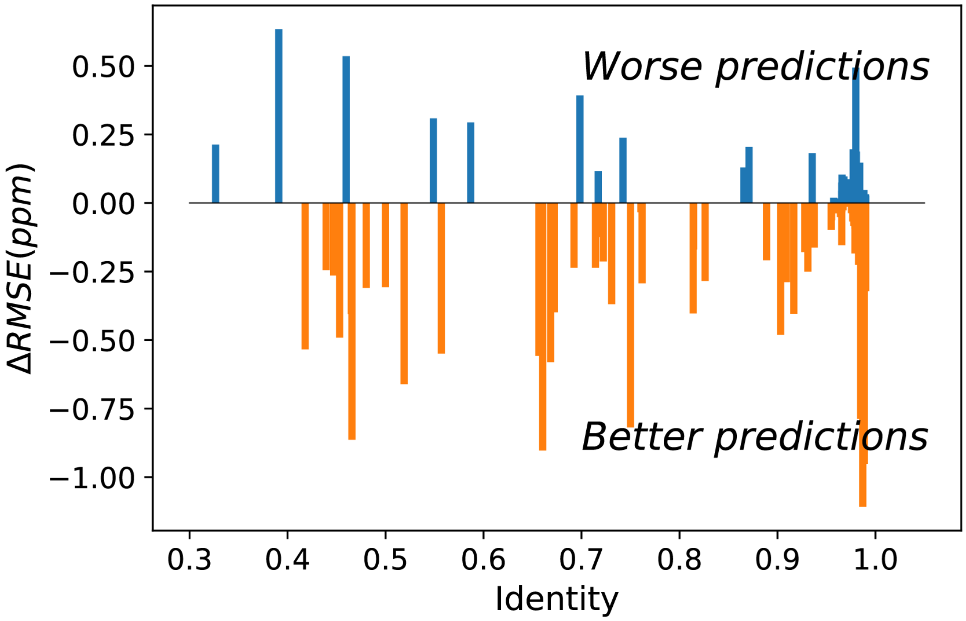


Figure 2.2: *Difference between UCBShift-Y and SHIFTY+ for protein specific RMSEs for amide hydrogens as a function of sequence identity.* The presence of more negative values indicates UCBShift-Y is making better predictions than SHIFTY+ across the range of sequence identity. The better RMSE even at low sequence identity arises from finding a structural homolog.

prediction for all atom types, validating that the structural alignments successfully found better reference proteins for the query protein which improved the overall prediction quality. In comparison with SHIFTY+, all atom types other than carboxyl carbon are improved in accuracy; although predictions for the carboxyl carbons are at the same level of accuracy as SHIFTY+, the failure to improve this atom type with UCBShift-Y is likely due to the lower number of chemical shifts available for transfer prediction for this atom type. Finally we note that our UCBShift-Y can be used as a standalone chemical shift predictor when sequence and structural alignments exist and have available experimental chemical shifts.

Analysis of the UCBShift-X module

The connection between features extracted from PDB files and the secondary chemical shifts was explored using several machine learning methods, including neural networks with a single hidden layer, deep fully-connected neural networks, residual neural networks, convolutional neural networks, recurrent neural networks, as well as tree-based ensemble models. The more complex and deeper neural networks performed well on the training dataset and vali-

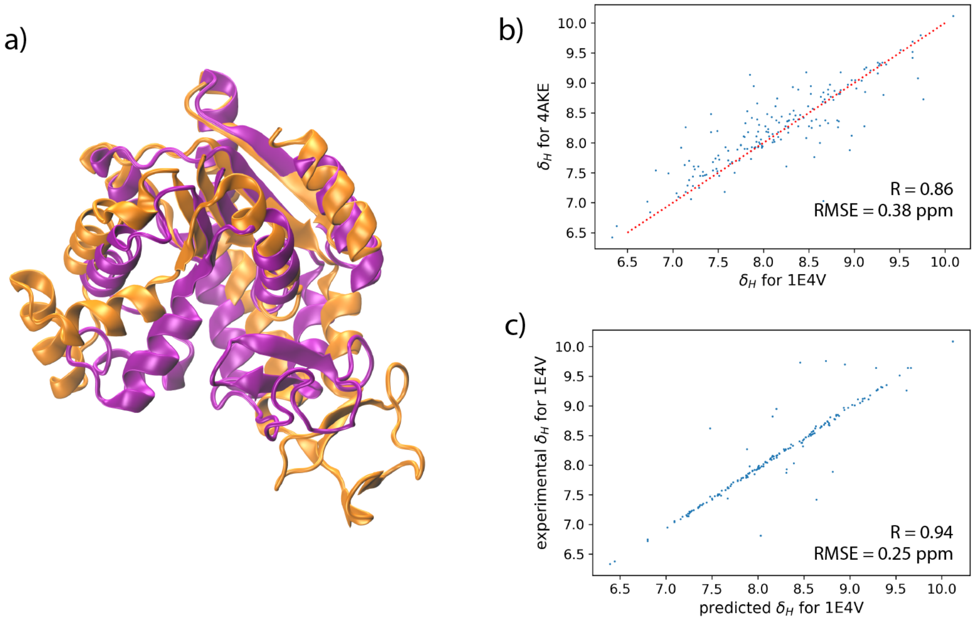


Figure 2.3: *Analysis of the transfer prediction module for UCBShift which uses sequence and structural alignment.* (a) Structural alignment of adenylate kinase (4AKE, orange) and the mutant G10V of adenylate kinase (1E4V, purple). (b) Correlation between experimental chemical shifts of the amide hydrogen for 4AKE and 1E4V. (c) Correlation between predicted amide hydrogen chemical shifts using UCBShift with experimental values. In this case structural alignments instead of sequence alignments were used for selecting references for the transfer prediction.

dation dataset, however their performance on the test data was found to be no better than SPARTA+ or SHIFTX+, likely indicating that more feature extracted data is needed and/or due to problems with data representation, to fully exploit the potential of these methods. Thus the tree-based ensemble models stood out as a more competitive machine learning predictor for chemical shifts with limited data. Even so, the learning curves for the random forest models show that the cross-validation error steadily decreases as the number of training examples increases (Fig. 2.E.1), suggesting even better predictions can be achieved if more training data were available.

The RMSE of the pipeline with extra tree regressor and random forest regressor but without inputs from UCBShift-Y (R_1) between the predicted chemical shifts and the observed shifts is summarized in Table 2.3 and named UCBShift-X. It is found to be statistically better for all the atom types when compared with SPARTA+, or the SHIFTX+ component of SHIFTX2, which also use no sequence and/or structural alignments. The overall performance of the UCBShift-X machine learning module is promising, and it also can be used as a reliable

Table 2.3: RMSE for the individual elements of transfer prediction (UCBShift-Y) and machine learning module (UCBShift-X) on the test dataset. The standalone UCBShift-Y prediction when sequence and structural alignments exist and have available experimental chemical shifts. The chemical shift prediction of the machine learning module (UCBShift-X) which is trained independent of any transfer prediction. The prediction of the R_2 module with input from UCBShift-Y module, and the combined R_1 and R_2 modules that defines the UCBShift calculator

UCBShift components	**H**	**Hα**	**C$'$**	**Cα**	**C$\boldsymbol{\beta}$**	**N**
UCBShift-X (R_1)	0.44	0.25	1.17	1.08	1.28	2.49
UCBShift-Y	0.21	0.17	0.64	0.57	0.67	1.25
ML with UCBShift-Y input (R_2)	0.19	0.15	0.66	0.57	0.70	1.23
UCBShift (utilizing both R_1 and R_2)	0.31	0.19	0.84	0.81	1.00	1.81

standalone predictor for chemical shifts, especially when no faithful alignment is found using UCBShift-Y.

If we consider using the R_2 module (which is trained using only a subset of the training data for which UCBShift-Y is able to make a prediction), the errors of some atom types further decrease (Table2.3). Interestingly, the averaged RMSE from R_2 for H, Hα and N is even smaller than the average RMSE of UCBShift-Y, indicating that the second ML module is doing better than just combining the results from UCBShift-Y and from the first level machine learning module R_0 for these atom types. But given the uncertainties in sequence and structural alignments or the lack of chemical shift data for UCBShift-Y, both the R_1 and R_2 machine learning modules are utilized to yield the final UCBShift algorithm and results for chemical shifts as given in Table 2.3 for all the six atom types.

Analysis of the data representation

A further test is done to analyze the contributions of different features extracted from the structural PDB files to the R_0, R_1, and R_2 pipelines that define the UCBShift algorithm (Fig. 2.1). Relative feature importance is calculated as the total decrease in node impurity weighted by the probability of reaching a node decided with that feature, and averaged over all the trees in the ensemble.[55] The results are analyzed on the predictions for amide hydrogen as a working example, and are given in Table 2.4. For the R_0 module we find that the most predictive features are the backbone dihedral angles, the secondary structure, BLOSUM numbers, hydrogen bond features, and the ring current effect which are included in SPARTA+ and SHIFTX2. However, the polynominal transformations of the structural data and half surface exposure are unique in our feature set, and they have very high importance among all the features for the R_0 component. Not surprisingly, the R_0 input is nearly half of the important features for R_1, but the backbone dihedral angles and the polynomial transformations account for an additional 25% of the important extracted features.

Table 2.4: Importance of different input features into the R_0, R_1, and R_2 pipelines of the machine learning modules

Feature categories	$\mathbf{R_0}$	$\mathbf{R_1}$	$\mathbf{R_2}$
Backbone dihedral angles	0.22	0.11	0.04
Transformed features	0.23	0.11	0.04
Secondary structure	0.17	0.005	0.001
BLOSUM numbers	0.11	0.02	0.005
Hydrogen bond	0.11	0.06	0.02
Half surface exposure	0.05	0.06	0.008
Ring current	0.04	0.03	0.005
Sidechain dihedral angles	0.03	0.03	0.005
Atomic surface area	0.02	0.01	0.002
B Factor	0.008	0.01	0.002
S^2 order parameters	0.006	0.01	0.002
Hydrophobicity	0.002	0.001	0.0002
pH values	0.001	0.002	0.001
Prediction from R_0	N/A	0.53	0.19
UCBShift-Y prediction	N/A	N/A	0.67
UCBShift-Y metrics	N/A	N/A	0.01

The UCBShift-Y prediction as well as the prediction from R_0 are the dominant factors for the R_2 model; this result indicates the network is indeed trying to differentiate situations when UCBShift-X predictions are more reliable and when they are not so accurate in comparison to R_0 predictions, as well as based on the other structure-derived features. Therefore, using machine learning to combine the predictions from feature-based prediction and alignment-based prediction is a better strategy than doing a weighted average of the two predictions. Finally features such as hydrophobicity and pH values, and to some extent B-factors and S^2 order parameters, seem to play a minor role in predictive capacity of the ML module.

Application of UCBShift to protein structure discrimination

One practical application of an accurate protein chemical shift calculator is to detect native structures based on the correlation between predicted and experimental chemical shifts.[56]To illustrate UCBShift's applicability to determine the native structure of a protein, we obtained two decoy datasets that have a range of altered and misfolded structures as measured by the α-carbon root mean square deviation (RMSD) with the native state. The average correlation coefficients for each structure with available experimental chemical shifts of H, Hα and N from BMRB 4429 (1CTF) and BMRB 4811 (1HFZ) are plotted over RMSDs to their native structures in Fig. 2.4.

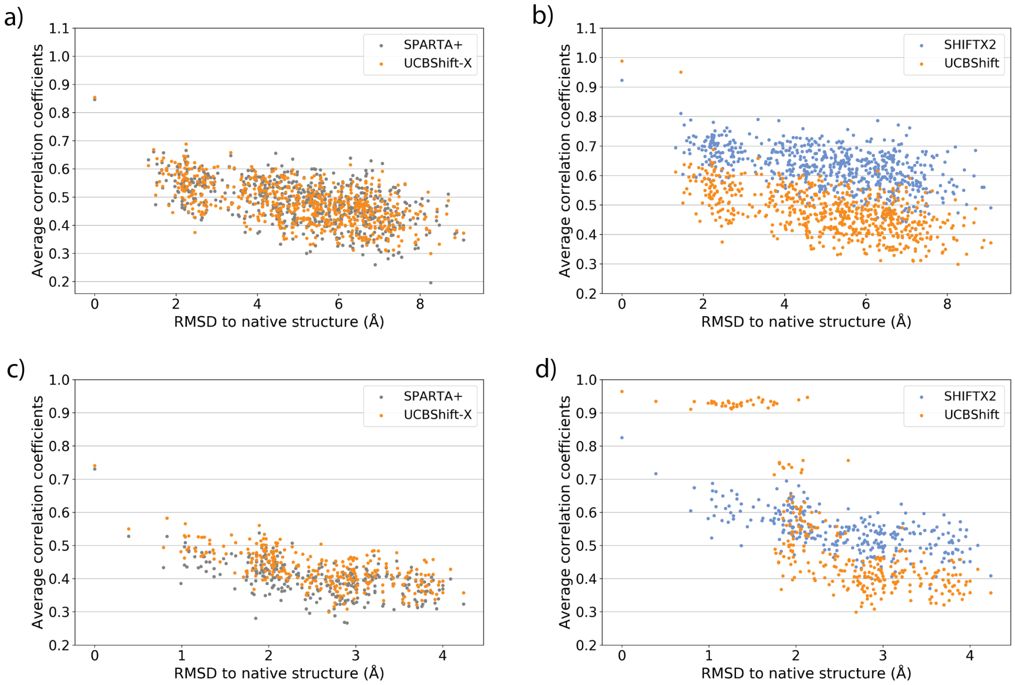


Figure 2.4: Average correlation coefficients between predicted chemical shifts of decoyed structures and observed chemical shifts versus Cα RMSD between decoyed structures and native structure for PDB 1CTF (a and b) and 1HFZ (c and d). Results are visualized as UCBShift-X compared to SPARTA+ (a and c) and UCBShift compared to SHIFTX2 (b and d).

The decoy dataset for PDB structure 1CTF was obtained from the Decoy 'R' Us database[57] which contains the native structure and 630 structures with a range of 1.3–9.1 Å in the α-carbon RMSD, and the decoy dataset for PDB structure 1HFZ generated using 3DRobot[58] which contains the native structure and 298 structures with a range of 0.4–4.2 Å in the α-carbon RMSD. Using UCBShift-X alone has similar discriminative power for the native state as SPARTA+, and predicted chemical shifts for lower RMSD structures also tend to have better correlation with experimental values using UCBShift-X (Fig. 2.4a and c). The complete UCBShift method shows greater discriminative power for the native state than found with either SPARTA+ or SHIFTX2, in which we can differentiate between structures within experimental resolutions (< 2 Å) against unlikely structures more easily, while still retaining the highest correlation for the (experimental) native structure (Fig. 2.4b and d).

2.4 Discussion and conclusion

Prediction of protein chemical shifts from structure has relied on robust and popular algorithms such as SPARTA+ and SHIFTX2 that represent the 3-dimensional structure by a set of extracted features that are presented to a machine learning algorithm, sometimes supplemented with direct transfer of experimental data taken on related proteins of a given query sequence. In this paper, we tested the performance of SPARTA+ and SHIFTX2 on a large test set of proteins not previously encountered in previous training and test sets, and showed that SPARTA+ performs as reported and evenly across high and low sequence homology test data, as expected. SHIFTX2 still outperforms SPARTA+ on test sequences with high sequence homology, but not at the same levels expected from the reported RMSE literature values.[52] This test dataset contains "outliers" which may be harder to predict, and hence is a more faithful representation of actual real-world data. We have also developed and tested a new generation algorithm, UCBShift, for solution chemical shift prediction for all relevant protein atom types, and utilizing more small molecular structure information (water and ligands), physically inspired non-linear transformation of features derived from structure, together with a two-level machine learning pipeline that exploits sequence as well as structural alignments to achieve this current state-of-the art performance. The feature extraction algorithm, the UCBShift prediction program, and all training and testing data can be downloaded from a publicly available github repository https://github.com/THGLab/CSpred.

Although the performance of these algorithms are much better when applied to carefully curated test data, the filtering out of test data risks the inability to distinguish between a poor prediction from a poor experimental chemical shift value. Large outliers would certainly result from the wrong random coil reference for Cβ shifts due to ambiguous cysteine oxidation states, or single whole proteins which exhibit many chemical shift outliers for particular atom types, and should not be considered a failure in algorithmic performance, but a problem of the experimental data. However, further test filtering can start to become arbitrary as we move from deviant to suspicious to acceptable experimental agreement with *the prediction*; one can't have it both ways. Thus in this paper we have provided a realistic range of test performance since scientists use these chemical shift predictors on real-world data that may differ from the original training datasets such that the algorithms do not generalize well-*i.e.* some measure of disagreement with experiment may just simply be prediction error. As such Table 2.2 provides a more realistic range of test reliability for all three methods.

Although we have realized noticeable improvement over other protein chemical shifts predictors, we believe we are reaching the limit of accuracy by using extracted feature from structures or transfer predictions through alignments. Deep learning may be helpful in the next step since it can operate directly on 3D data representations without the potential bias introduced by features extracted by human experts[59, 60] as we and others have shown recently for chemical prediction in the solid state,[14, 15] in which we greatly improved prediction for all atom types and approached chemical accuracy on par with *ab initio* calculations for hydrogen in particular. The ability to move to 3D representations will be important for QM chemical shift predictions for intrinsically disordered proteins, since feature extracted data will be less available and likely less representative for this class of protein, and the results can be ensemble averaged to provide a prediction that can be compared against solution NMR experimental data for structural ensemble refinement as we have shown for IDPs.[32, 33]

2.5 Funding

This work was supported by the National Institutes of Health [5U01GM121667]; and the computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy [DE-AC02-05CH11231]. KCB thanks the California Alliance Postdoctoral Fellowship for early support of this work.

2.6 Acknowledgements

We thank Mojtaba Haghighatlari, Brad Ganoe, Tim Stauch, Lars Urban, Shuai Liu, and Martin Head-Gordon for fruitful discussions.

2.7 References

[1] Vladimir Saudek, Annalisa Pastore, Maria A Castiglione Morelli, Rainer Frank, Heinrich Gausepohl, Toby Gibson, Falk Weih, and Paul Roesch. Solution structure of the dna-binding domain of the yeast transcriptional activator protein gcn4. *Protein Engineering, Design and Selection*, 4(1):3–10, 1990.

[2] Michael P Williamson. Secondary-structure dependent chemical shifts in proteins. *Biopolymers: Original Research on Biomolecules*, 29(10-11):1423–1431, 1990.

[3] David S Wishart, Brian D Sykes, and Frederic M Richards. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *Journal of molecular biology*, 222(2):311–333, 1991.

[4] Silvia Spera and Ad Bax. Empirical correlation between protein backbone conformation and c. alpha. and c. beta. 13c nuclear magnetic resonance chemical shifts. *Journal of the American Chemical Society*, 113(14):5490–5492, 1991.

[5] David S Wishart and Alex M Nip. Protein chemical shift analysis: a practical guide. *Biochemistry and Cell Biology*, 76(2-3):153–163, 1998.

[6] Wim F Vranken and Wolfgang Rieping. Relationship between chemical shift value and accessible surface area for all amino acid atoms. *BMC Structural Biology*, 9(1):1–10, 2009.

[7] David A Case. Chemical shifts in biomolecules. *Current opinion in structural biology*, 23(2):172–176, 2013.

[8] Jakob T Nielsen, Hamid R Eghbalnia, and Niels Chr Nielsen. Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. *Progress in nuclear magnetic resonance spectroscopy*, 60:1–28, 2012.

[9] Da-Wei Li and Rafael Brüschweiler. Ppm: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *Journal of biomolecular NMR*, 54:257–265, 2012.

[10] Dawei Li and Rafael Brüschweiler. Ppm_one: a static protein structure based chemical shift predictor. *Journal of biomolecular NMR*, 62:403–409, 2015.

[11] David S Wishart, M Scott Watson, Robert F Boyko, and Brian D Sykes. Automated 1h and 13c chemical shift prediction using the biomagresbank. *Journal of biomolecular NMR*, 10(4):329, 1997.

[12] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *Journal of biomolecular NMR*, 38:289–302, 2007.

[13] Bowei Wang, Yunjun Wang, and David S Wishart. A probabilistic approach for validating protein nmr chemical shift assignments. *Journal of biomolecular NMR*, 47:85–99, 2010.

[14] Shuai Liu, Jie Li, Kochise C Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. Multiresolution 3d-densenet for chemical shift prediction in nmr crystallography. *The journal of physical chemistry letters*, 10(16):4558–4565, 2019.

[15] Federico M Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nature communications*, 9(1):4501, 2018.

[16] Mitsuo Iwadate, Tetsuo Asakura, and Michael P Williamson. Cα and cβ carbon-13 chemical shifts in proteins from an empirical database. *Journal of biomolecular NMR*, 13:199–211, 1999.

[17] Stephen Neal, Alex M Nip, Haiyan Zhang, and David S Wishart. Rapid and accurate calculation of protein 1 h, 13 c and 15 n chemical shifts. *Journal of biomolecular NMR*, 26(3), 2003.

[18] Jens Meiler and David Baker. Coupled prediction of protein secondary and tertiary structure. *Proceedings of the National Academy of Sciences*, 100(21):12105–12110, 2003.

[19] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131(39):13894–13895, 2009.

[20] Yang Shen and Ad Bax. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *Journal of biomolecular NMR*, 48:13–22, 2010.

[21] Beomsoo Han, Yifeng Liu, Simon W. Ginzinger, and David S. Wishart. SHIFTX2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, mar 2011.

[22] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C. F. Schulte, D. E. Tolmie, R. Kent Wenger, H. Yao, and J. L. Markley. BioMagResBank. *Nucleic Acids Research*, 36(Database):D402–D408, dec 2007.

[23] H. Jane Dyson and Peter E. Wright. Unfolded proteins and protein folding studied by NMR. *Chemical Reviews*, 104(8):3607–3622, jul 2004.

[24] David S. Wishart, Colin G. Bigam, Arne Holm, Robert S. Hodges, and Brian D. Sykes. 1h, 13c and 15n random coil NMR chemical shifts of the common amino acids. i. investigations of nearest-neighbor effects. *Journal of Biomolecular NMR*, 5(1):67–81, jan 1995.

[25] C.W. Haigh and R.B. Mallion. Ring current theories in nuclear magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 13(4):303–344, jan 1979.

[26] DavidA. Case. Calibration of ring-current effects in proteins and nucleic acids. *Journal of Biomolecular NMR*, 6(4), dec 1995.

[27] A. D. Buckingham. CHEMICAL SHIFTS IN THE NUCLEAR MAGNETIC RESONANCE SPECTRA OF MOLECULES CONTAINING POLAR GROUPS. *Canadian Journal of Chemistry*, 38(2):300–307, feb 1960.

[28] Fengli Zhang and Rafael Brüschweiler. Contact model for the prediction of NMR n-h order parameters in globular proteins. *Journal of the American Chemical Society*, 124(43):12654–12655, oct 2002.

[29] David A. Case. Molecular dynamics and NMR spin relaxation in proteins. *Accounts of Chemical Research*, 35(6):325–331, nov 2001.

[30] Haiyan Zhang, Stephen Neal, and David S. Wishart. *Journal of Biomolecular NMR*, 25(3):173–195, 2003.

[31] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M. Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K. Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H. Arrowsmith, Thomas Szyperski, Gaetano T. Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proceedings of the National Academy of Sciences*, 105(12):4685–4690, mar 2008.

[32] David H. Brookes and Teresa Head-Gordon. Experimental inferential structure determination of ensembles for intrinsically disordered proteins. *Journal of the American Chemical Society*, 138(13):4530–4538, mar 2016.

[33] James Lincoff, Mojtaba Haghighatlari, Mickael Krzeminski, João MC Teixeira, Gregory-Neal W Gomes, Claudiu C Gradinaru, Julie D Forman-Kay, and Teresa Head-Gordon. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Communications chemistry*, 3(1):74, 2020.

[34] Masayoshi Nakasako. Water–protein interactions from high–resolution protein crystallography. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1448):1191–1206, aug 2004.

[35] Alfonso De Simone, Guy G. Dodson, Franca Fraternali, and Adriana Zagari. Water molecules as structural determinants among prions of low sequence identity. *FEBS Letters*, 580(10):2488–2494, apr 2006.

[36] H. M. Berman. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, jan 2000.

[37] J.Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation 1 1edited by j. thornton. *Journal of Molecular Biology*, 285(4):1735–1747, jan 1999.

[38] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, dec 1983.

[39] Mark V. Berjanskii and David S. Wishart. A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society*, 127(43):14970–14971, oct 2005.

[40] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, oct 1990.

[41] Runze Dong, Zhenling Peng, Yang Zhang, and Jianyi Yang. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 34(10):1719–1725, dec 2017.

[42] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, mar 1970.

[43] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, nov 1992.

[44] William C. Wimley and Stephen H. White. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Structural &amp Molecular Biology*, 3(10):842–848, oct 1996.

[45] Thomas Hamelryck. An amino acid has two sides: A new 2d measure provides a different view of solvent exposure. *Proteins: Structure, Function, and Bioinformatics*, 59(1):38–48, feb 2005.

[46] Gerhard Wagner, Arthur Pardi, and Kurt Wuethrich. Hydrogen bond length and proton NMR chemical shifts in proteins. *Journal of the American Chemical Society*, 105(18):5948–5949, sep 1983.

[47] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, mar 2006.

[48] Leo Breiman. *Machine Learning*, 45(1):5–32, 2001.

[49] Robert E Schapire. Theoretical views of boosting. In *Computational Learning Theory: 4th European Conference, EuroCOLT'99 Nordkirchen, Germany, March 29–31, 1999 Proceedings*, pages 1–10. Springer, 1999.

[50] Randal S Olson, Ryan J Urbanowicz, Peter C Andrews, Nicole A Lavender, La Creis Kidd, and Jason H Moore. Automating biomedical data science through tree-based pipeline optimization. In *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30–April 1, 2016, Proceedings, Part I 19*, pages 123–137. Springer, 2016.

[51] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[52] Beomsoo Han, Yifeng Liu, Simon W Ginzinger, and David S Wishart. Shiftx2: significantly improved protein chemical shift prediction. *Journal of biomolecular NMR*, 50:43–57, 2011.

[53] CW Müller, GJ Schlauderer, Jochen Reinstein, and Georg E Schulz. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure*, 4(2):147–156, 1996.

[54] Christoph W Müller and Georg E Schulz. Crystal structures of two mutants of adenylate kinase from escherichia coli that modify the gly-loop. *Proteins: Structure, Function, and Bioinformatics*, 15(1):42–49, 1993.

[55] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[56] Anders S. Christensen, Troels E. Linnet, Mikael Borg, Wouter Boomsma, Kresten Lindorff-Larsen, Thomas Hamelryck, and Jan H. Jensen. Protein structure validation and refinement using amide proton chemical shifts derived from quantum mechanics. *PLoS ONE*, 8(12):e84123, dec 2013.

[57] Ram Samudrala and Michael Levitt. Decoys 'r' us: A database of incorrect conformations to improve protein structure prediction. *Protein Science*, 9(7):1399–1401, 2000.

[58] Haiyou Deng, Ya Jia, and Yang Zhang. 3drobot: automated generation of diverse and well-packed protein structure decoys. *Bioinformatics*, 32(3):378–387, oct 2015.

[59] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14), apr 2007.

[60] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The Journal of Physical Chemistry Letters*, 6(12):2326–2331, jun 2015.

Appendix

2.A Exclusion of erroneous chemical shifts assignments in the test dataset

Table 2.A.1: Removed chemical shifts that are significantly offset from random coil average

PDBID	**RESID**	**RESNAME**	**Recorded CS (ppm) [atom]**
1MI4A	153	LEU	0.09[H]
1DSBA	30	CYS	331.45[CB]
2GYKE	62	ASP	89.55[CB]
1GOA	119	TRP	26.57[N]
1OVHA	1	MET	38.5[N]
1D03A	1	ALA	40.47[N]
2IGD	1	MET	39.34[N]
1VB0A	1	LEU	39.84[N]
1QOGA	1	ALA	39.52[N]
1DSBA	51	PRO	152.99[N]

Table 2.A.2: Excluded cysteine Cβ chemical shifts

PDBID	**RESID**	**Oxidation state in crystal structure**	**Recorded Cβ CS (ppm)**
2GOOB	40	reduced	48.59
2PF5A	47	reduced	43.71
1DSBA	33	oxidized	33.6
1VKBA	35	oxidized	31.16
1VKBA	62	oxidized	26.76
2AFGB	117	reduced	39.46
1LJUA	82	oxidized	30.66
1LJUA	89	oxidized	27.12

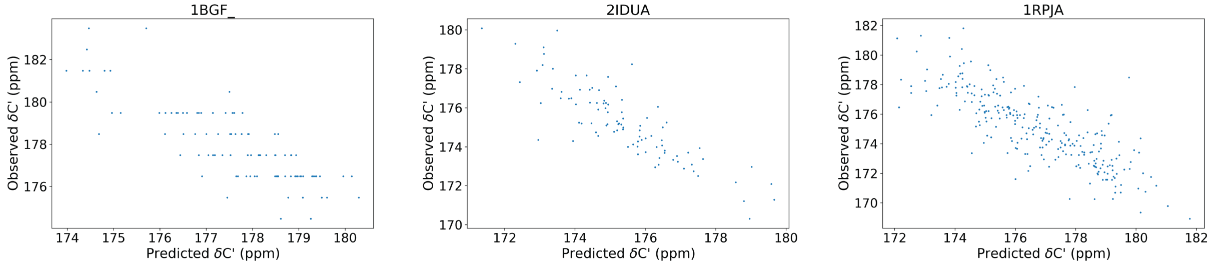


Figure 2.A.1: *Scatter plot for the C' of the three proteins whose C' shifts are removed.* Plotted are the observed C' chemical shifts versus predicted C' chemical shifts with UCBShift.

2.B Compilation of test dataset

Table 2.B.1: *A complete list of the 200 testing proteins.* Provided are the PDB identifier, BMRB identifier, the X-ray resolution (RES.), highest sequence similarity to any example in the training dataset (SIM.), total number of residues and the number of residues and atom types with experimental chemical shifts. Rows with green background are data with low sequence homology (<30%).

						Number of residues				
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C'**	**Cα**	**Cβ**	**N**
1MI4A	4854	1.7	0.998	427	209	198	236	229	199	217
2GP0A	15680	2.05	0.997	288	240	0	149	188	156	151
2H9HA	4836	1.39	0.995	212	201	170	202	212	186	201
1T85A	17415	1.8	0.995	406	305	0	0	299	196	233
2UYZA	4132	1.4	0.994	156	137	142	0	141	107	116
1GOA_	4012	1.9	0.994	155	134	110	0	119	0	118
1SNO_	1878	1.7	0.993	136	0	0	0	0	0	109
5NUC_	5536	2.1	0.993	134	97	93	100	101	93	96
2EYOA	16585	1.7	0.993	135	118	117	125	126	117	118
1NBPA	6621	2.2	0.992	121	114	116	0	116	111	114
1LJUA	4944	1.4	0.992	130	118	105	109	125	116	118
1IP2A	5125	1.8	0.992	130	122	0	0	0	0	118
1AR0A	5888	2.3	0.992	125	114	0	0	112	100	108
7TIMB	16566	1.9	0.992	247	0	0	16	63	69	2
1OMSC	15813	2.3	0.992	114	107	104	112	114	105	107
1BRJA	975	2	0.991	107	101	94	0	0	0	0
2H76A	62	2.25	0.991	108	88	79	0	0	0	0

						Number of residues				
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C'**	**Cα**	**Cβ**	**N**
2HSHA	256, 257	1.35	0.990	105	59	60	0	0	0	0
1RN1B	1658	1.84	0.990	103	0	0	0	0	0	96
2OCTA	15548	1.4	0.990	97	84	0	89	88	83	84
1QOGA	447	1.8	0.990	98	76	79	45	84	80	79
2GYKE	4115, 4116	1.6	0.988	83	77	75	0	80	74	77
1POH_	29	2	0.988	85	82	79	0	0	0	0
1SPQB	15066	2.16	0.988	239	185	0	0	183	146	147
1TPH2	15064	1.8	0.988	245	224	0	223	231	174	204
1J3FA	4568	1.45	0.987	152	144	121	143	146	124	144
109M_	5158	1.83	0.987	154	129	0	131	142	0	129
1F2MA	1875, 495	2	0.987	136	0	108	0	76	1	0
2IN8A	16634	1.7	0.986	139	129	0	0	0	0	122
2FI4I	4877	1.58	0.983	58	46	46	0	0	0	0
1P2OD	45	2	0.983	58	47	44	0	0	0	0
2OW9B	4679	1.74	0.982	166	137	140	145	154	139	137
1BNEA	16169, 16170	2.1	0.982	107	84	55	0	0	0	0
1TXXA	1812, 1813	2.2	0.981	108	0	0	0	0	0	95
2SGDI	1375	1.8	0.980	51	0	0	0	43	15	0
1MOLB	4633	1.7	0.979	94	77	73	0	0	0	0
1IV9A	4222	1.9	0.979	96	84	80	0	0	0	78
1YJFA	5514	1.35	0.979	225	174	96	174	198	136	174
2HWNB	4473	1.6	0.978	45	39	41	0	43	41	39
2HZIB	15488	1.7	0.975	264	244	0	36	41	0	37
1QKRA	15653	1.8	0.973	172	146	0	139	146	126	130
2BC5C	1672	2.25	0.972	106	91	92	0	0	0	91
1XRKB	4785, 4786	1.5	0.968	121	104	104	103	110	104	104
1MYWA	15826	2.2	0.962	228	192	123	184	189	139	175
1DS3I	1374	1.65	0.961	50	39	38	0	0	0	0
1LW6I	4974	1.5	0.953	63	57	55	56	57	54	57
1U06A	7305, 7306	1.49	0.952	55	52	51	0	0	0	0
1SKOB	6181	2	0.946	116	108	104	0	114	106	108
1P7JA	7386	2.1	0.932	53	36	36	30	31	29	30
1AJ6_	5218	2.3	0.932	194	177	143	169	161	106	160
2H61H	4105, 5895	1.9	0.924	90	80	79	78	83	78	80
1TCF_	1553	1.9	0.899	156	57	53	0	0	0	0
1KTZB	5953, 5954, 4779	2.15	0.893	106	90	100	100	101	100	90
2B59A	5267	2.11	0.872	166	157	142	132	153	121	138
1B68A	6524	2	0.859	138	134	0	0	133	120	132
1BD9B	17382	2.05	0.834	185	170	169	181	184	169	170

					Number of residues					
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C'**	**Cα**	**Cβ**	**N**
1N3ZA	4980	1.65	0.815	117	108	108	0	0	0	101
1AG6_	79	1.6	0.788	99	86	83	0	0	0	0
1IKOP	7220	1.92	0.764	141	127	120	127	133	122	127
2BIUX	6141, 7310	1.71	0.745	164	139	123	0	146	127	139
1OFFA	16024	1.8	0.735	95	91	88	0	0	0	91
1MR3F	15626	1.6	0.718	177	145	0	0	135	100	123
1Z7XX	4370	1.95	0.674	126	119	121	0	0	0	111
1GWYA	16362, 16630	1.71	0.674	175	169	154	166	175	154	169
7RXN_	15374	1.5	0.673	52	44	43	0	0	0	44
1YCQA	6248	2.3	0.645	88	75	72	0	78	66	72
1JHFB	6373	1.8	0.644	111	96	0	89	101	83	90
1RDG_	5163	1.4	0.635	52	45	42	0	0	0	0
1PCS_	5475	2.15	0.580	98	84	80	0	0	0	0
2AFGB	15783, 16493, 16494, 16502, 6875	2	0.543	129	123	117	123	129	104	122
1WZVA	16321	2.1	0.535	150	129	50	125	141	136	129
2GCNA	16668	1.85	0.517	177	127	91	138	138	127	127
1L0SB	5573	2.3	0.500	87	76	67	81	79	72	76
1OKHB	4587	1.75	0.478	46	41	42	0	0	0	0
1OBOA	5011	1.2	0.477	169	145	130	0	0	0	136
1RCF_	16593	1.4	0.472	169	98	91	0	2	0	43
1FDQA	5320, 16042, 16046, 16047, 16048, 16049	2.1	0.466	131	128	121	0	130	121	128
1FU0A	4774	1.9	0.448	86	84	77	0	0	0	0
1EK8A	5190	2.3	0.438	185	175	0	181	181	0	167
2GGMB	5503, 6687	2.35	0.436	137	130	121	0	0	0	62
1HFZC	4811, 4332	2.3	0.435	121	111	99	0	0	0	110
1D03A	1673	1.85	0.424	169	160	142	0	0	0	159
2HPWA	16600	1.55	0.417	228	206	111	159	203	168	185
1YW5A	16690	1.6	0.412	177	165	0	170	170	158	165
1GPR_	1663	1.9	0.407	158	146	133	0	0	0	134
2HPR_	932	2	0.402	86	79	74	0	0	0	0
5HPGA	16466	1.66	0.393	84	75	73	0	0	0	75
1WYWB	6304	2.1	0.381	79	77	0	0	0	0	71
1DSBA	16327	2	0.381	188	0	0	146	167	124	151
1KDJ_	7370	1.7	0.373	102	96	0	0	0	0	90
1VB0A	16060	0.92	0.348	61	0	0	53	55	50	51
1ZDNA	17437	1.93	0.348	151	132	0	0	122	104	121

					Number of residues					
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C’**	**Cα**	**Cβ**	**N**
2AAOB	5324	2	0.343	131	118	115	0	114	104	111
1M8AB	15596	1.7	0.343	61	59	58	0	0	0	0
1IOZA	10051	2	0.339	162	156	0	0	0	0	153
2BJDA	6398	1.27	0.337	90	87	81	0	0	0	0
1V6PB	1381	0.87	0.333	62	55	51	0	0	0	0
1RB4B	371	1.9	0.333	32	31	30	0	0	0	0
2PF5A	6392, 6393	1.9	0.327	88	84	78	84	88	78	84
2SEMA	5729	2.2	0.317	58	49	48	50	52	48	48
2IDSA	16740	1	0.295	105	93	0	0	0	0	93
1EQTB	16803	1.6	0.294	67	58	60	57	62	60	58
3DFR_	7200, 7196, 7197, 7198, 7199	1.7	0.286	162	147	0	0	0	0	137
1H0JB	4966	1.9	0.283	60	54	58	0	59	0	0
1BAZC	394, 395	1.9	0.283	46	41	40	0	0	0	0
1DCDB	5249	2	0.278	36	35	26	0	32	0	35
1OKSA	6568, 6569	1.8	0.268	50	48	0	0	0	0	47
1R1TA	4128, 4306	1.7	0.262	98	96	0	0	97	0	93
2GITE	5784, 5169, 3078	1.7	0.260	100	92	97	0	0	0	0
1W7ZC	397	1.67	0.258	31	26	23	0	0	0	0
2IDUA	16741	0.95	0.257	104	95	0	94	99	89	95
1UWXA	1639	2.2	0.254	57	54	50	0	0	0	0
1HC9A	8, 4195, 5006, 5024	1.8	0.243	74	63	68	0	0	0	0
1CM9B	4914, 4852	2.1	0.243	66	60	63	0	0	0	58
1HC9B	15130	1.8	0.243	74	0	70	0	74	0	0
1YU7X	4428	1.5	0.239	64	57	58	0	0	0	56
1ZNIC	1632, 1585, 554, 1344	1.5	0.238	21	20	20	0	0	0	0
2F91B	5274	1.2	0.229	32	29	29	0	0	0	0
2ALGA	16294	2.3	0.228	92	77	69	0	0	0	0
1PZ4A	16662	1.35	0.224	113	106	104	0	109	100	102
2IE2C	17162, 17163	1.7	0.223	212	194	0	0	191	161	174
1CLVI	4490, 4404	2	0.219	32	27	29	0	0	0	0
2GJ2A	7099	2.35	0.212	79	77	74	0	78	68	74
1PHT_	16448	2	0.212	83	0	0	46	69	46	61
2CA5B	15214	2.1	0.212	61	50	0	0	16	10	39
1WKXA	6123	1.7	0.209	43	31	30	0	0	0	0
1LP1B	1120, 4324	2.3	0.207	54	12	13	0	0	0	12

						Number of residues				
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C'**	**Cα**	**Cβ**	**N**
2FFGA	15529	2.31	0.207	77	68	0	0	61	57	61
1LU0B	4246	1.03	0.207	29	27	26	0	0	0	0
1PPEI	199, 314, 2227, 2527	2	0.207	29	3	5	0	0	0	0
1HSLB	16204, 16205	1.89	0.206	238	215	169	221	216	184	204
1PGX_	2575	1.66	0.205	70	58	54	0	0	0	0
1OVHA	915	1.95	0.201	162	152	139	0	0	0	152
1DFNA	16254	1.9	0.200	30	0	0	27	28	24	28
1CQMB	16344	1.65	0.198	98	88	88	0	0	0	88
2IGD_	15283	1.1	0.197	61	0	0	54	54	50	53
1FD3D	4642	1.35	0.195	41	35	35	0	0	0	0
451C_	1333, 10133	1.6	0.195	82	75	73	0	0	0	69
1O5UA	16006	1.83	0.188	88	82	83	82	87	84	82
1RPJA	16984, 16982	1.8	0.188	288	273	0	261	288	263	273
1QG7A	16142	2	0.179	62	58	62	58	62	62	57
1EZGB	5323	1.4	0.179	82	78	74	0	0	0	76
2NWGB	16143	2.07	0.176	64	57	61	57	61	61	55
2GSVB	15350	1.9	0.175	66	63	61	0	65	65	63
1R69_	2539	2	0.174	63	60	58	0	0	0	0
2PSPB	2384	1.95	0.170	105	88	94	0	0	0	0
3WRP_	17010	1.8	0.167	101	85	0	0	57	53	65
1OS3D	1633	1.95	0.167	28	26	25	0	0	0	0
1QE6C	280	2.35	0.167	67	61	62	0	0	0	0
1WTQA	5905, 5908	1.7	0.167	64	60	0	0	0	0	59
1TUKA	4977	1.12	0.164	67	62	60	0	0	0	61
3ERAB	7211	1.7	0.161	62	53	52	0	0	0	51
2DGCA	1396, 1397, 1398, 1399	2.2	0.159	49	48	48	0	0	0	0
1O82D	4112	1.46	0.157	70	69	60	0	0	0	0
1C8CA	4570	1.45	0.156	64	56	51	0	0	0	0
2CG7A	15756	1.2	0.156	90	83	0	0	88	72	83
1BWOB	2065, 4932, 4383	2.1	0.156	90	84	81	0	0	0	0
2H9EC	4396	2.2	0.155	52	50	48	0	2	4	50
1OMYA	7330	2	0.154	64	59	60	0	0	0	0
1MIDA	15143	1.71	0.154	91	83	73	0	0	0	0
1KBAB	1675	2.3	0.152	66	60	58	0	0	0	0
1L1DB	6051	1.85	0.151	145	113	100	119	125	95	99
1YU8X	15245	1.45	0.149	64	39	40	0	0	0	0
1NAQF	15094	1.7	0.143	105	99	41	104	105	98	104

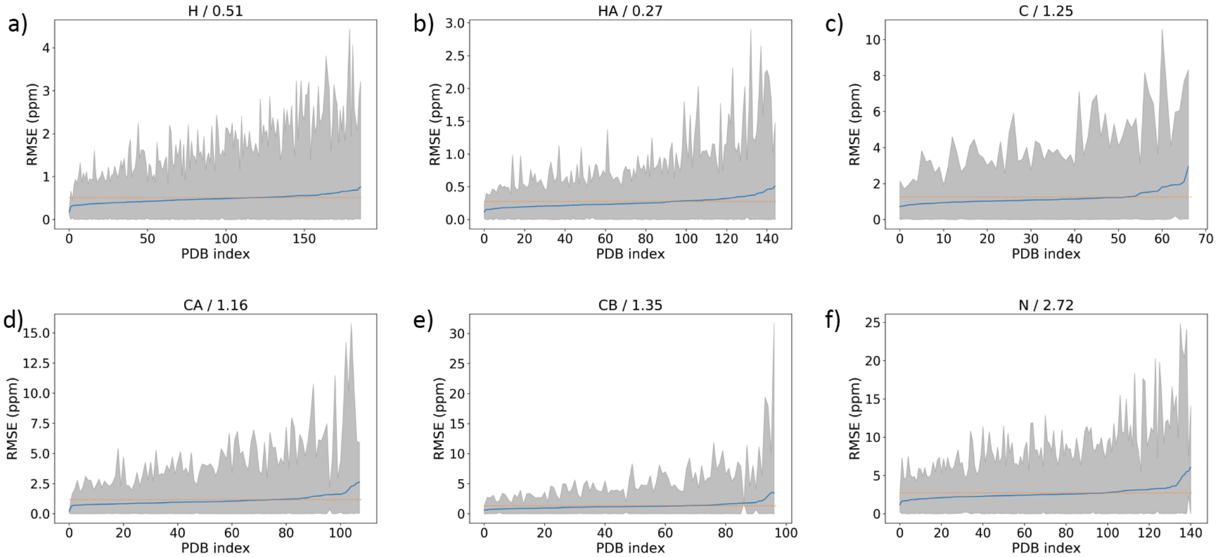
					Number of residues					
PDB	**BMRB**	**RES. (Å)**	**SIM.**	**Total**	**H**	**Hα**	**C'**	**Cα**	**Cβ**	**N**
1GZZB	2498, 4204, 15654	2.3	0.143	60	54	54	57	59	44	50
2A3GA	1442	2.25	0.143	21	19	20	0	0	0	0
1KX9B	5094	1.65	0.143	102	98	95	0	0	0	92
1BNZA	6050	2	0.141	64	51	47	0	0	0	0
1ICFJ	5583	2	0.138	65	58	59	0	0	0	0
1YNRB	10135	2	0.138	79	75	0	0	0	0	72
1GN0A	17136	1.8	0.130	108	103	99	0	108	100	103
1AE3_	2039	2	0.128	86	75	72	0	0	0	0
1HB8B	2049	2	0.116	86	82	80	0	0	0	0
1KTZA	4411	2.15	0.116	82	67	74	0	0	0	67
2GM5D	4269	2.1	0.115	114	79	72	78	81	74	79
1CY5A	4661	1.3	0.113	92	86	86	0	84	12	86
1VKBA	16380	1.9	0.112	147	130	132	0	141	116	118
1QVEA	4918	1.54	0.111	126	114	105	0	0	0	108
2GOOB	15956	2.2	0.107	85	76	79	53	66	57	53
1J1VA	5200	2.1	0.106	91	82	72	85	91	87	86
1THQA	6234	1.9	0.100	147	102	0	0	99	75	91
2BEMC	17160	1.55	0.100	170	152	159	168	169	159	152
1ENFA	16146	1.69	0.099	212	204	0	0	200	165	198
1BGF_	5997	1.45	0.097	124	111	0	96	104	85	111
1Q2UA	17507	1.6	0.095	189	171	0	178	179	148	157
2UV0H	6271	1.8	0.091	163	150	0	161	161	147	150
1AM7A	16664	2.3	0.089	150	143	126	141	146	0	143
1VYKA	17357	1.49	0.087	129	82	71	91	99	96	93
1JPST	16838	1.85	0.087	200	179	0	177	178	66	160
1XIOA	17064	2	0.084	217	0	0	129	132	97	134
2IM8A	16113, 7227	2	0.084	120	95	104	0	94	90	75
2BDYA	16940	1.61	0.083	276	248	0	0	0	0	0
1ASS_	5930	2.3	0.082	152	145	140	133	147	0	140
2Q2TA	16059	2.3	0.072	293	249	0	236	265	235	249
1K82A	5219	2.1	0.071	260	172	0	106	148	95	118
1CKUB	2999	1.2	0.071	85	75	73	0	0	0	0
1PHP_	16464, 16451, 16447	1.65	0.063	394	372	0	359	363	325	366
2ASDA	16869	1.95	0.053	341	189	0	187	191	141	189
2GT8A	17251	2	0.042	298	187	185	168	195	161	169
2ILNI	5617	2	0.000	53	47	53	0	0	0	0

2.C Evaluation of SPARTA+, SHIFTX2 and UCBShift performance on uncurated Test dataset

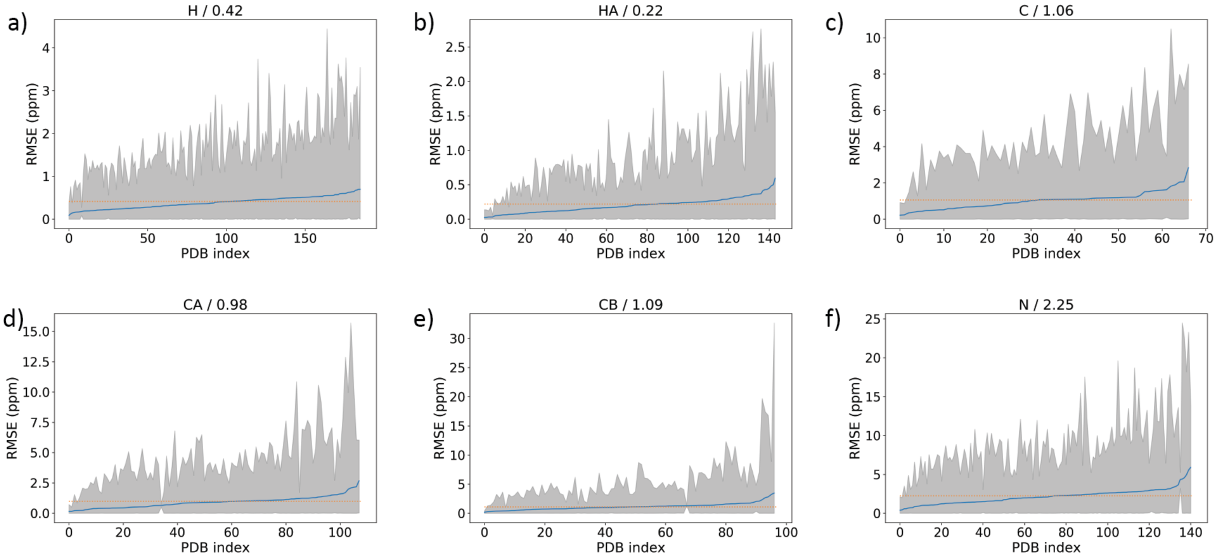
Table 2.C.1: Root mean square error (RMSE), mean squared error (MAE) and Pearson's correlation coefficients (R^2) for UCBShift in the curated and uncurated test dataset and its low homology subset.

Dataset		**Test**			**LH-Test**	
Atom Type	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
H	0.31 ± 0.003	0.18 ± 0.001	0.90	0.45 ± 0.004	0.32 ± 0.001	0.77
Hα	0.19 ± 0.002	0.11 ± 0.001	0.94	0.26 ± 0.003	0.18 ± 0.002	0.87
C	0.84 ± 0.01	0.48 ± 0.004	0.93	1.14 ± 0.01	0.81 ± 0.007	0.86
Cα	0.81 ± 0.01	0.43 ± 0.004	0.99	1.09 ± 0.01	0.73 ± 0.005	0.97
Cβ	1.00 ± 0.03	0.47 ± 0.005	1.00	1.34 ± 0.05	0.83 ± 0.009	0.99
N	1.81 ± 0.02	1.06 ± 0.006	0.95	2.61 ± 0.02	1.89 ± 0.01	0.88
Dataset		**Test (Curated)**			**LH-Test (Curated)**	
Atom Type	**RMSE**	**MAE**	**R^2**	**RMSE**	**MAE**	**R^2**
H	0.30 ± 0.002	0.17 ± 0.001	0.90	0.43 ± 0.003	0.32 ± 0.002	0.78
Hα	0.18 ± 0.002	0.10 ± 0.001	0.94	0.24 ± 0.003	0.17 ± 0.002	0.88
C	0.77 ± 0.008	0.44 ± 0.004	0.94	1.13 ± 0.01	0.81 ± 0.006	0.86
Cα	0.76 ± 0.009	0.41 ± 0.003	0.99	1.04 ± 0.01	0.71 ± 0.006	0.98
Cβ	0.82 ± 0.01	0.45 ± 0.004	1.00	1.12 ± 0.02	0.78 ± 0.007	1.00
N	1.71 ± 0.01	1.03 ± 0.008	0.95	2.55 ± 0.02	1.86 ± 0.01	0.88

SPARTA+



SHIFTX2



UCBShift

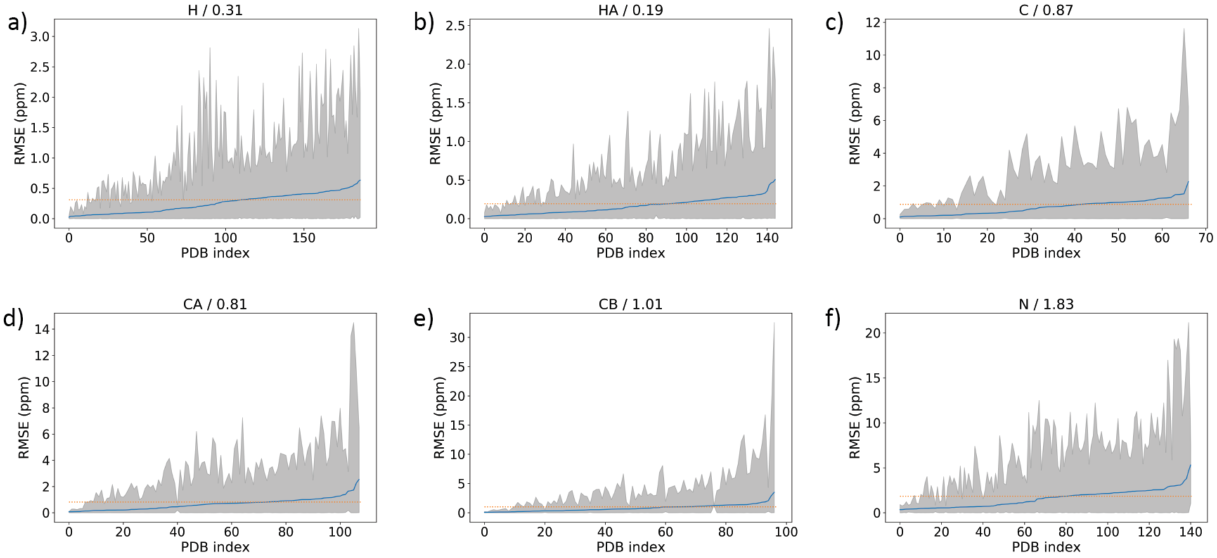
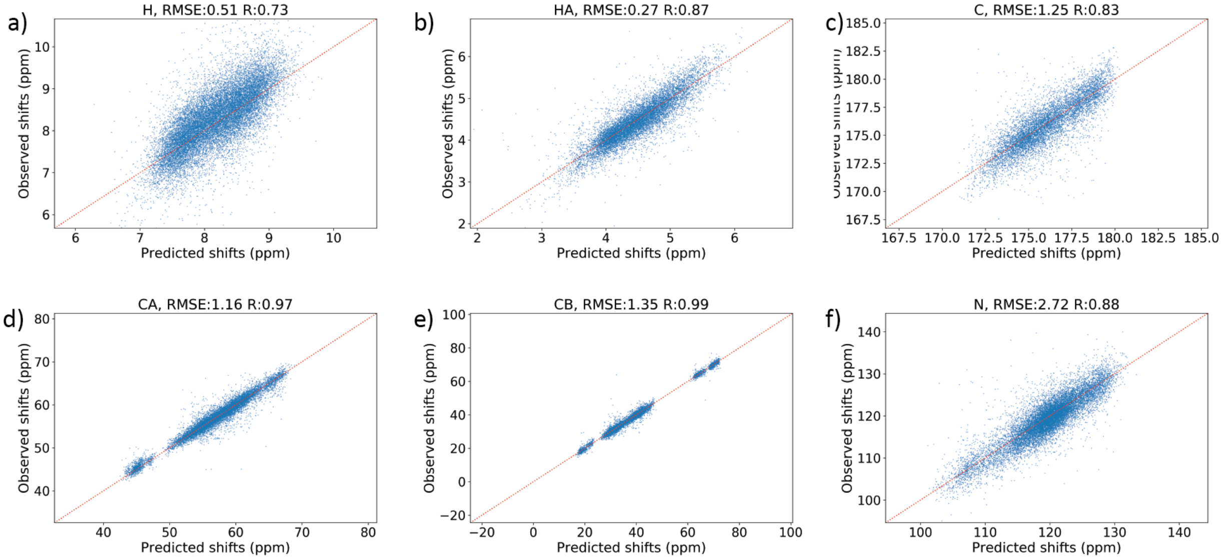
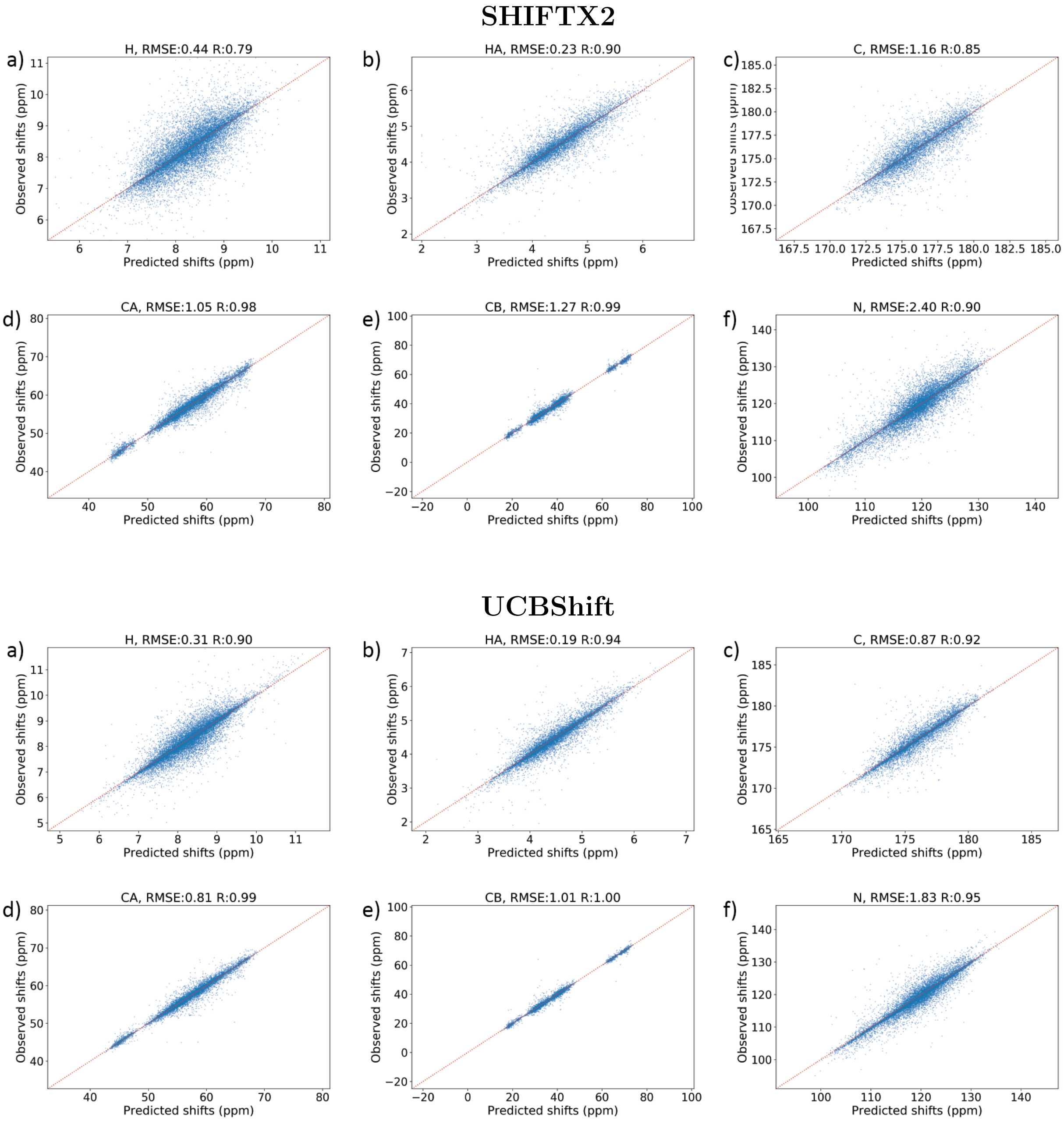


Figure 2.C.1: *Error analysis for SPARTA+, SHIFTX2 and UCBShift.* In each graph, the dotted line indicates the average RMSE over all the predicted residues. PDBs are sorted according to the prediction RMSE over that specific structure, and the grey region represents the minimum and maximum RMSE of a single residue in a protein for atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N atom types.

SPARTA+





SHIFTX2

UCBShift

Figure 2.C.2: *Scatter plot for SPARTA+, SHIFTX2 and UCBShift predictions.* The predicted chemical shifts for six atom types plotted with the experimental shifts types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N atom types.

Table 2.C.2: Examples of SHIFTX2 that have information leakage from training to test set.

Identifier in training	**Identifier in testing**	**Sequence similarity**	**Structure RMSD (Å)**
R014_3LZTA	A055_1YKYX	100%	0.77
R114_1VDQA	A055_1YKYX	100%	0.41
R020_1RUVA	A001_1KF3A	100%	0.20
R006_2CPLA	A054_1CWCA	98%[a]	0.21
R129_1ZJLA	A022_1ZJLA	100%	0
R072_3RN3A	A001_1KF3A	100%	0.11

[a] Training (R006) is a sub-sequence of testing (A054) with 1 aa missing on N-terminus and 2 aa missing on C-terminus

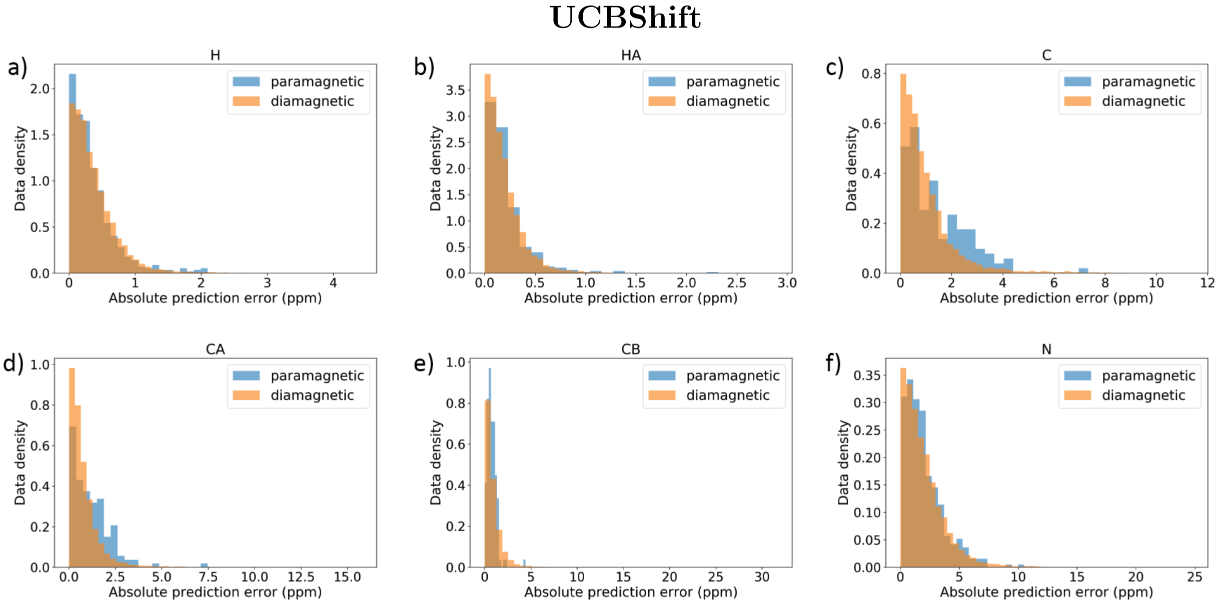


Figure 2.C.3: *Comparison of absolute error distribution for paramagnetic proteins and diamagnetic proteins based on SPARTA+ predictions.* Atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N.

2.D Performance analysis of UCBShift-Y in comparison with SHIFTY+

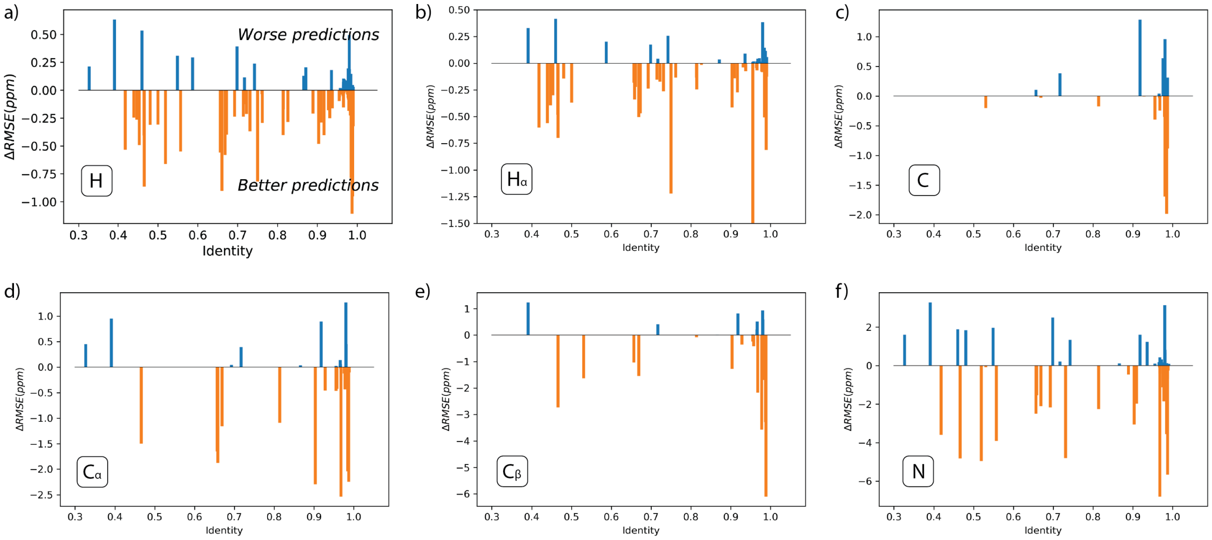


Figure 2.D.1: *Difference between UCBShift-Y and SHIFTY+ for protein specific RMSEs for different atom types as a function of sequence identity.* Atom types (a) H, (b) Hα, (c) C', (d) Cα, (e) Cβ, and (f) N.

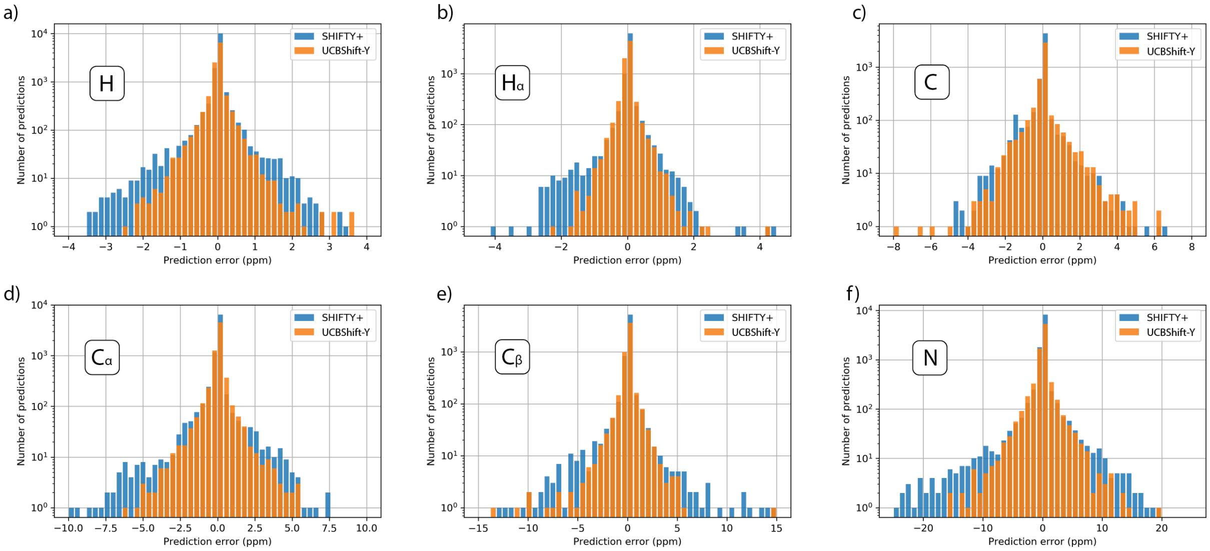


Figure 2.D.2: *Error distributions for UCBShift-Y and SHIFTY+ across all atom types.* This compares only the transfer prediction module of UCBShift (UCBShift-Y) which uses sequence and structural alignment.

2.E Training curves for random forest models

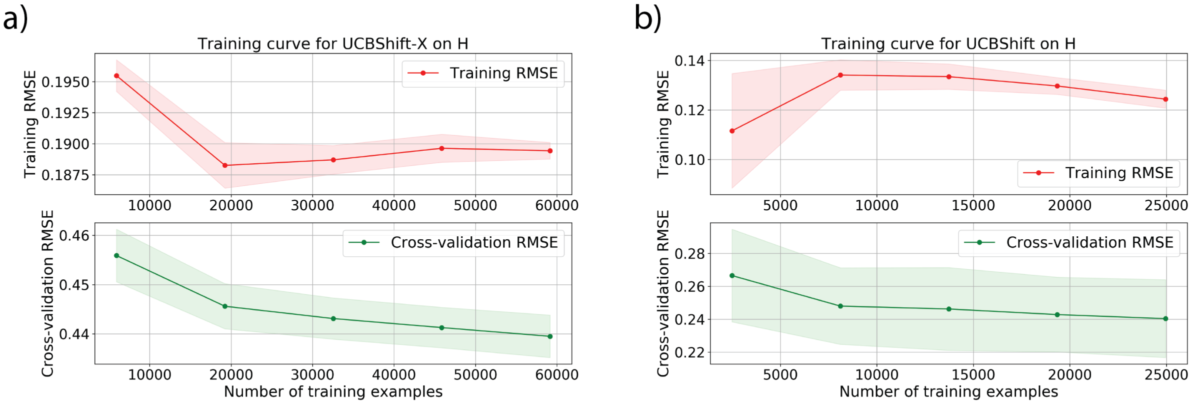


Figure 2.E.1: Training curves for R1 and R2 in UCBShift-X module for hydrogen.

CHAPTER 3

Highly Accurate Prediction of NMR Chemical Shifts from Low-Level Quantum Mechanics Calculations Using Machine Learning[†]

3.1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is a highly accurate experimental technique to probe chemical bonding and subtle environmental differences of atoms in various molecular systems, ranging from small molecules[1, 2, 3], natural products[1, 4, 5], biopolymers[6, 7], to materials.[8, 9, 10] The NMR chemical shift (CS), which describes the shielding effect offset of a nucleus of interest relative to a defined standard molecule, is one of the most informative data obtained from an NMR measurement, especially for molecular structure[11], identifying the crystal morphology from a selection of candidates[10], distinguishing among synthetic outcomes for natural products,[5] and building and refining atomic level models for proteins.[12]. Therefore, accurate CS back-calculators which connect structure to shift perturbations are an indispensable tool in trying to help scientists understand and make good use of NMR chemical shifts measurements.

Chemical shifts arise from the electron shielding of a nucleus under an external magnetic field. The shift values can be calculated from first principles[13, 14, 15] using the second order magnetic shielding tensor $\hat{\sigma}$, that describes the response of the induced magnetic field in all directions, but usually only the isotropic component $\sigma_{iso} = \frac{1}{3}\mathrm{Tr}(\hat{\sigma})$ is mapped to an experimental observable.[16] Calculation of chemical shifts can be done with exceptional

[†] Reproduced with permission from: Li J*, Liang J*, Wang Z, Ptaszek AL, Liu X, Ganoe B, Head-Gordon M, Head-Gordon T. Highly Accurate Prediction of NMR Chemical Shifts from Low-Level Quantum Mechanics Calculations Using Machine Learning. *arXiv preprint* arXiv:2306.08269. 2023 Jun 14. (* denotes equal first authors)

accuracy using coupled-cluster theory with single and double excitation and perturbative-approximated triple excitations [CCSD(T)] together with a complete basis set (CBS) or one that is sufficiently large for convergence.[17, 18, 19] However, with present-day algorithms and computing resources, such calculations are essentially impractical for any complex systems that contain more than ten heavy atoms (non-hydrogen atoms), due to their computational scaling. Efforts continue to reduce the cost by approaches such as composite methods[20, 21], and nucleus-optimized electronic structure models.[22]

Alternatively, data-driven approaches have also been quite successful in predicting experimental or calculated chemical shifts at greatly reduced cost. For aqueous proteins, chemical shifts can be predicted from carefully curated features extracted from 3-dimensional geometries of the peptides using machine learning (ML) methods including neural networks and random forests, such as implemented in SPARTA+[23], SHIFTX2[24] and UCBShift[25]. For organic small molecules in crystalline form, kernel ridge regression (KRR) [10]and 3D convolutional networks (CNN)[26] have been employed to predict chemical shieldings calculated using gauge-including projector-augmented waves (GIPAW) density functional theory (DFT) methods from merely the molecular structure inputs. Recent work by Guan et al. has trained a 3D graph neural network to predict H and C chemical shifts for neutral organic molecules found in NMRShiftDB[27] using quantum mechanics (QM) optimized geometries and DFT calculated chemical shifts, and then transfer learning to predict experimental chemical shifts from force-field optimized geometries.[28] These ML models that directly predict chemical shifts from input geometries are orders of magnitude faster than QM calculations, and can usually achieve comparable accuracy to the quantum mechanical method they have been trained on. However, this has typically relied upon DFT that can calculate chemical shieldings at a much more acceptable cost, but also can often suffer from insufficient accuracy[29, 30, 31]. In addition, machine learning methods are not expected to generalize to a different molecular system, unlike QM methods that are still much more generalizable and rigorous in terms of predicting chemical shieldings for a specific input geometry.

The question arises whether a machine learning method can be used to “amend” a low-level QM prediction to high accuracy, hence achieving generalizability and speed at the same time. An intuitive way is to use machine learning to predict the difference between a high-level and low-level calculation, using molecular geometries as input. Such Δ-machine learning idea are exemplified in the work of Unzueta, et al. that predicts a correction to a cheap DFT calculation using small basis set and arrives at the target accuracy of the same DFT method with a large basis set.[32] Very recently, Büning and Grimme have shown that a similar approach can correct DFT predictions of chemical shieldings to CCSD(T) quality, signifying an important step in predicting CS at the highest level of theory achievable from theoretical calculations.[33]

But what is true about many such ML approaches is that they can be poor in predicting out-of-distribution cases, i.e. outside the specifics of the training data.[34] Ideally, good feature engineering can provide an augmented chemical representation beyond just molecular configuration[34, 25], information that is preferably derived from a cheap calculation but which is nonetheless invaluable information for not only obtaining high-level accuracy, but transferability. This very idea has been proven for predicting correlation energies at MP2 and CCSD level using molecular orbital features at the mean-field Hartree-Fock (HF) level[35].

In this work we present a novel feature representation obtained from a low-level DFT chemical shielding calculation of the diamagnetic (DIA) and paramagnetic (PARA) shielding tensor elements, and combine it with geometric-dependent features that are used as input into a neural network model to predict chemical shieldings equivalent to CCSD(T)/CBS accuracy.[21] In addition we introduce a novel active learning (AL) training procedure that selects out-of-distribution training data with increasing number of heavy atoms from a full set of off-equilibrium geometries obtained from the ANI-1 dataset.[36] Finally, to analyze the transferability of iShiftML to other systems, we find that error estimations in terms of the standard deviation among a committee of ML models is well correlated with the actual error without knowing the target values, signaling when the model is or is not trustworthy for applications outside the original training set.

The resulting iShiftML model trained with data up to 7 heavy atoms has exceptional predictive performance when evaluated on the 8 heavy atom test data, achieving prediction errors of 0.11 ppm for H, 1.54 ppm for C, 3.90 ppm for N and 6.33 for O between predicted chemical shieldings and the target CCSD(T) composite method values. The iShiftML model when compared against experimental gas phase CS measurements for molecules that are not included in the training set reduces the error of the low-level DFT calculation by at least 50%. Furthermore, we have used our method to predict experimental CSs for natural products that are vastly larger and more chemically complex than any molecule from our training dataset, illustrated with strychine and vannusal, in which we show that diastereomers of the vannusal B molecule can be easily differentiated by inspecting the errors between predicted CS and experimental measurements. We expect the iShiftML method to be used extensively to achieve highly accurate predictions of chemical shifts of various molecular systems, and also facilitate theoretical research of NMR chemical shieldings at the CCSD(T)/CBS level.

3.2 Methods and Models

Feature Selection for Machine Learning of Chemical Shifts

The magnetic shielding tensor $\hat{\sigma}$ is defined as the total second derivative of the energy E with respect to nuclear spin $\mathbf{M}^A$ at nucleus A and the external magnetic field $\mathbf{B}^{ext}$, with components defined as

$$\sigma_{ab} = \left.\frac{\mathrm{d}^2 E(\mathbf{M}^A, \mathbf{B})}{\mathrm{d}M_a^A \mathrm{d}B_b}\right|_{\mathbf{B}=0, \mathbf{M}^A=0} \tag{3.1}$$

Here "d" means total derivative, and a, b correspond to Cartesian indices. For a variationally optimized wavefunction with parameters, $\boldsymbol{\theta}$ (even the exact wavefunction), the total derivative has two partial derivative contributions:

$$\sigma_{ab} = \left.\left\{ \frac{\partial^2 E(\mathbf{M}^A, \mathbf{B})}{\partial M_a^A \partial B_b} + \frac{\partial^2 E(\mathbf{M}^A, \mathbf{B})}{\partial M_a^A \partial \boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}}{\partial B_b} \right\}\right|_{\mathbf{B}=0, \mathbf{M}^A=0, \boldsymbol{\theta}=\boldsymbol{\theta}_{\mathrm{opt}}} \tag{3.2}$$

Given that the chemical shielding tensor and each of its components, σ_{ab}, can also be decomposed into diamagnetic and paramagnetic components within the DFT gauge-including atomic orbitals (GIAO) approach[37, 38],

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} = \begin{pmatrix} \mathrm{DIA}_{xx} & \mathrm{DIA}_{xy} & \mathrm{DIA}_{xz} \\ \mathrm{DIA}_{yx} & \mathrm{DIA}_{yy} & \mathrm{DIA}_{yz} \\ \mathrm{DIA}_{zx} & \mathrm{DIA}_{zy} & \mathrm{DIA}_{zz} \end{pmatrix} + \begin{pmatrix} \mathrm{PARA}_{xx} & \mathrm{PARA}_{xy} & \mathrm{PARA}_{xz} \\ \mathrm{PARA}_{yx} & \mathrm{PARA}_{yy} & \mathrm{PARA}_{yz} \\ \mathrm{PARA}_{zx} & \mathrm{PARA}_{zy} & \mathrm{PARA}_{zz} \end{pmatrix} \quad (3.3)$$

it comes naturally that the isotropic chemical shieldings at the same level of theory can be calculated as

$$\sigma_{iso} = \frac{1}{3}(\sigma_{xx}+\sigma_{yy}+\sigma zz) = \frac{1}{3}(\mathrm{DIA}_{xx}+\mathrm{DIA}_{yy}+\mathrm{DIA}_{zz}+\mathrm{PARA}_{xx}+\mathrm{PARA}_{yy}+\mathrm{PARA}_{zz}) \quad (3.4)$$

in which the off-diagonal elements have a contribution of zero to the final isotropic chemical shielding formula. However, the full tensor of Eq. 3.3 still encodes useful information about the local atomic environments for each nucleus and might be helpful with predicting chemical shieldings at a higher level of accuracy. Hence we formulate the chemical shift tensor components DIA and PARA as a feature set for the machine learning approach described further below.

In addition, we use Atomic Environment Vectors (AEVs) as geometric descriptors that are used to describe the atomic environments at each nucleus, following previous studies[36, 32]. AEVs are reformulations of the atomic symmetry functions used by Behler and Parinello in their neural networks for predicting molecular energies[39], which contain orientation-independent angular and radius terms that are determined by local geometries of nearby atoms categorized by atom type within a cutoff. The 384-dimensional AEV for an atom constitutes a radial part (the first 64 elements) and an angular part (the remaining 320 elements). The radial elements for atom i are calculated as

$$G_{A,n}^{(rad)} = \sum_{j \in \mathcal{N}[i]} e^{-\eta(R_{ij}-R_n)^2} f_C(R_{ij}) \quad (3.5)$$

where A denotes a specific atom type of H, C, N, O for the second atom, and n is a distance index that defines the different reference distances R_n from the center atom. The summation is done over all neighbor atoms j with type A near the central atom i within a cutoff, and R_{ij} is the distance between atoms i and j. The reference distances are defined as $R_n = 0.9 + a_0/2 * n$ where $a_0 = 0.529177$Å is the Bohr radius and n ranges from 0 to 15. $\eta = 16$ was used to adjust the width of each Gaussian so that it matches with the separation between two consecutive reference distances. Finally, $f_C(R_{ij})$ is a cutoff function that smoothly modulates the Gaussian term around the cutoff radius, with the following formula and cutoff radius $R_C = 5.2$Å:

$$f_C(R_{ij}) = \begin{cases} (1 + cos(\pi \frac{R_{ij}}{R_C}))/2 & R_{ij} \leq R_C \\ 0 & \text{otherwise} \end{cases} \quad (3.6)$$

We have used 16 distance indices for each atom type and hence 64 radial AEV values.

Similarly, the angular components of an AEV vector are defined as

$$G^{(ang)}_{A,B,m,n} = 2^{1-\xi} \sum_{j,k \in \mathcal{N}[i], j \neq k} (1 + cos(\theta_{ijk} - \theta_m))^{\xi} f_{(R,n)}(R_{ij}, R_{ik}) \tag{3.7}$$

$$f_{R,n}(R_{ij}, R_{ik}) = e^{-\eta(R_{ij}+R_{ik})/2-R_n)^2} f_C(R_{ij}) f_C(R_{ik}) \tag{3.8}$$

with A, B defining the two different atom types for nearby atoms, and thus $4+3+2+1 = 10$ different atom type combinations are possible. m and n are the angle and distance indices that define the reference angles and positions by $\theta_m = \frac{2m+1}{16}\pi$ with m from 0∼7, $R_n = (0.90, 1.55, 2.20, 2.85)$Å, and θ_{ijk} denotes the angle centered at atom i. The same mathematical format of the distance cutoff function was used, but with a radial cutoff value of $R_C = 3.5$Å. The normalization constant $\xi = 32$. The 10 atom type combinations, 8 reference angles and 4 reference distances altogether defines 320 different angular components of the AEV vector. The calculation of AEVs were performed with the precompiled C++ code from Ref. 32.

NMR shielding calculations and stability analysis

Recently Liang et al. presented a systematic investigation on using locally dense basis sets (LDBS) and composite QM methods for chemical shieldings calculations, which have been categorized into low-level, middle-level and high-level effectiveness based on a balance of accuracy and computational cost.[21] We selected the ωB97X-V functional [40] in conjunction with the pcSseg-1 basis set [41] as our low-level method. The ωB97X-V functional offers robust and transferable performance for various properties prediction,[42, 43, 44, 45, 46, 47, 48, 49] particularly the dipole moment,[44] a simple but effective measure of electron density in polar molecules. We opted for this functional over the low-level methods recommended in Ref. 21, which provide more accurate shielding predictions, because those methods could potentially benefit from error cancellation. Thus, we believe it is more advantageous to use ωB97X-V as input for predicting high-level results. The advantage of using ωB97X-V for the low-level input was also validated by its better in-distribution and out-of-distribution predition error when comparing models trained with different low-level methods as input, which are described in Appendix Table 3.C.1. The ORCA 5.0.3 software [50] was utilized for these calculations, and local exchange-correlation integrals were computed over DefGrid3, a default ORCA grid, for all atoms. GIAOs[38] were used in all shielding calculations, including subsequent high-level computations.

We directly adopted the high-level method suggested in Ref. 21, namely CCSD(T)/pcSseg-1 with a basis set correction between pcSseg-1 and pcsSeg-3 calculated from the resolution of identity Møller-Plesset second-order perturbation theory (RIMP2), abbreviated with CCSD(T)(1)∪RIMP2(3). This high-level method can achieve impressively low root mean square errors (RMSEs) (0.048 ppm for H, 0.47 ppm for C, 3.58 ppm for N, and 4.68 ppm for O) in comparison to the theoretical best estimates, CCSD(T) with a complete basis set

(CBS). The CFOUR program package, version 2.1, was utilized for CCSD(T) computations [51, 52, 53], while ORCA was used for RIMP2 calculations. In RIMP2 calculations, the def2-JK [54] auxiliary basis set was employed for the Coulomb and exchange part, whereas the cw5C [55] auxiliary basis set was used for auxiliary correlation fitting to expedite the computation.

As our training set encompasses many conformations far from equilibrium and quantum mechanical (QM) calculations are likely to fail, we employed the stability analysis[56] at HF/pcSseg-1 level to validate our calculations. We exclude all conformations that might exhibit instabilities, including Restricted HF (RHF) → RHF, RHF→ Unrestricted HF, and RHF → Complex RHF.

Dataset preparation

The ANI-1 dataset [36], which contains over 20 million off-equilibrium geometries of small organic molecules up to 8 heavy atoms obtained through normal mode sampling, together with the equilibrium structures of these 57,462 molecules, were used to define the most inclusive dataset (DS-ANI-1) used in this work. However, it is very challenging to perform chemical shielding calculations for all the data in DS-ANI-1, even at a low-level DFT level of theory, and is not accessible for the CCSD(T) calculations that are orders of magnitude more time-consuming than DFT calculations.

To reduce the size of the dataset while keeping the diversity of the conformations of the molecules, a “farthest sampling” algorithm was developed that down-samples off-equilibrium geometries for each molecule in the ANI-1 dataset. The root-mean-square-deviations (RMSDs) for molecules after the optimal alignment using the Quarternions method [57] was used to evaluate conformation dis-similarities between geometries of the same molecule. A conformation collection pool was defined with the first conformation of a molecule being the first element. In each iteration, the aligned RMSDs for all geometries in ANI-1 dataset but not in the collected pool were calculated towards all conformations in the collected pool, and the geometry with the highest RMSD was added to the collected pool.

The total number of collected conformations depends on the number of heavy atoms in the molecule. For molecules up to 4 heavy atoms, 200 most dissimilar conformations were collected into the pool. For molecules with 5, 6 and 7 heavy atoms, the number of non-equilibrium conformations collected for each molecule were 100, 50 and 5 respectively. The equilibrium geometries for molecules with 5-7 heavy atoms were always included in the dataset. A stability analysis was performed to further exclude systems for which the NMR shielding calculations are likely to fail or be erroneous. This collection of a sub-sampled dataset (DS-SS) is our primary data for model training and development of the active learning workflow of the iShiftML model, which contains 12,677 geometries for molecules up to 4 heavy atoms, 13,313 geometries for molecules with 5 heavy atoms, 31,462 geometries for molecules with 6 heavy atoms and 37,105 geometries with 7 heavy atoms.

Using the geometries of all these data, we calculated the DIA and PARA matrix elements under the low-level composite DFT method ωB97X-V/pcSseg-1 DFT.[21] For the dataset with 5-7 heavy atoms, 1500 geometries were selected from active learning to perform the high-level composite method[21]. The active learning dataset covering all data using the high-level target values are subsequently labeled DS-AL-N, where N ranges from 4-7, which represents the maximum number of heavy atoms included in the dataset. Finally, 41 randomly selected molecules with 8 heavy atoms were collected from DS-ANI-1. The equilibrium geometries and a random non-equilibrium geometry for each of the 41 molecules were used to define our test dataset. Any data point for carbon chemical shielding with significant deviation between calculated low-level and high-level chemical shieldings was excluded. Our full training and testing dataset are provided in Appendix.

iShiftML Ensemble Model and Training Details

We have employed an ensemble machine learning approach by randomly splitting the training and validation data into 5 even portions, and 5 separate ML models were trained, each model using a different portion as validation data and the rest as training data. In addition, the network parameters for these five models were also initialized with different random numbers. After all models have been trained, they are combined into an ensemble model. When making predictions, each model in the ensemble predicts a value, and the prediction is given by the average from each of the 5 individual models in the ensemble.

Because outliers resulting from failed predictions may contaminate the average, any outliers should be identified and excluded from the calculation. To estimate outliers, we used the local outlier factor (LOF) algorithm implemented in the scikit-learn package to detect outliers.[58] The algorithm relies on a local neighbor density estimation to identify outliers as data points that have a significantly lower density of neigbors than the rest of the data points. Finally, the average and standard deviation among the non-outlier predictions were calculated.

All five ML models are trained by minimizing the mean squared error between the predicted isotropic chemical shieldings and the calculated high-level targets, under the following loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{n}(f_\theta(X_n) - Y_n)^2 \tag{3.9}$$

where f_θ represents the networks parameterized by θ, X_n are the input features, and Y_n are the target values. Weight decay of 3×10^{-5} and dropout with probability 0.1 [59] were used after each linear layer to reduce overfitting to the training data. Starting from a learning rate of 1×10^{-3}, a stepwise learning rate decay schedule was used that monitors evaluation performance on the validation dataset, and reduces learning rate by 30% if the validation error did not decrease after 20 epoch since last error reduction on the validation dataset, unless the learning rate is already smaller than 1×10^{-6}. The neural network was implemented in pytorch[60] and optimized using the Adam optimizer[61] with a batch size of 128 and was trained for 750 epochs.

3.3 Results

We begin with the results concerning the iShiftML model itself, the benefits of ensemble training, and a new active learning protocol in order to emphasize the ability to generalize, predict error confidence, and to construct affordable datasets for chemical shift prediction. A schematic of the iShiftML model architecture is depicted in Figure 3.1. For a given input geometry, the atomic environment vectors together with the paramagnetic and diamagnetic elements of the shielding tensor are calculated with the lower-level ωB97X-V/pcSseg-1 composite method, and are used as neural network inputs that are trained to predict chemical shieldings of the high-level composite method CCSD(T)(1)$\cup$RIMP2(3) for the four atom types for which we predict chemical shieldings: hydrogen, carbon, nitrogen and oxygen.

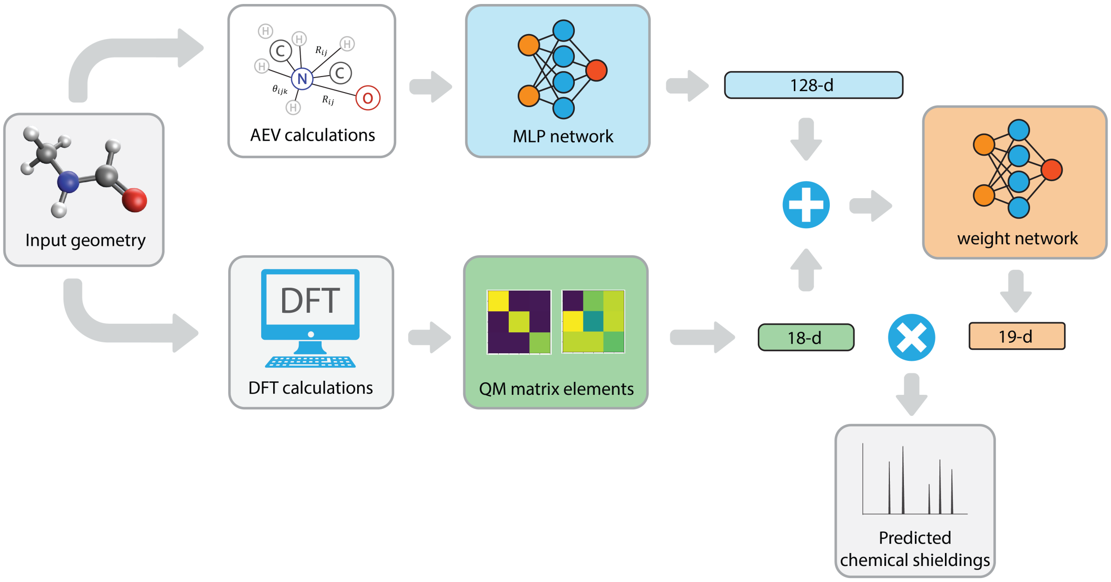


Figure 3.1: *The iShiftML ensemble learning model that uses low-level QM calculations of the shielding tensor and AEVs to predict high-level chemical shieldings.* (a) Given a molecular geometry, the AEV around each nucleus is prepared, and are sent into a multi-layer perception (MLP) network with two layers, each of which contains 128 neurons, in which the ReLU activation function[62] is used for the first layer to encode the AEVs into an 128-dimension internal representation. On a second branch, we perform low-level composite QM calculations to obtain the 18 DIA and PARA chemical shielding values that are concatenated with the AEVs from the first branch to provide input for the second MLP weight network. The weight MLP is composed of a first layer containing 64 neurons and uses ReLU activation, followed by a second layer of 19 neurons and a bias term without an activation function.

Figure 3.2 shows the distribution of the learnable weights of the 18 DIA and PARA values from the network for the test dataset of the hydrogen atom after the training converges. Even without explicit enforcement, the diagonal elements from the DIA and PARA matrices have weights that are close to $\frac{1}{3}$ and off-diagonal elements distributed around 0, which is consistent with Eq. 3.4; the bias term of -0.17 indicates the low-level chemical shieldings have a systematic offset from the more accurate high-level targets. This result proves that the model captures the physical connection between the isotropic chemical shieldings and the intermediate QM matrix elements, and should be generalizable to new predictions even outside of the training dataset, as long as the low-level QM matrix elements are reasonably accurate.

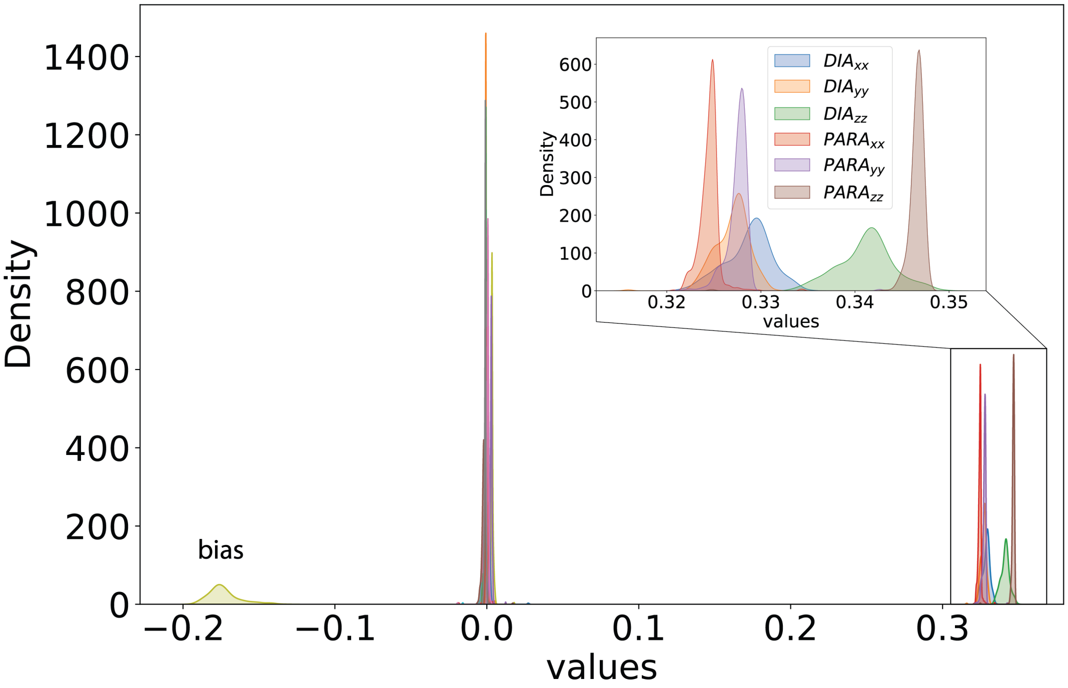


Figure 3.2: *Distributions of weight network outputs for hydrogen model evaluated on test data.* Distributions of the weights for diagonal elements in the DIA and PARA matrices are centered close to 1/3, off-diagonal elements are centered around 0, and the bias term is distributed around -0.17.

We have also employed an ensemble prediction technique to improve on the accuracy compared to any individual training of the iShiftML model (Figure 3.3a). Table 3.1 shows the performance comparisons for individual models and the ensemble average for a model trained with DS-AL-4 for oxygen. We see that while an individual model may make large errors, such as in models 3 and 5, the ensemble average model can mitigate these erroneous predictions, and still reach a consensus prediction that has a lower RMSE and standard deviation than any individual model.

Table 3.1: Root mean square errors (RMSE) and standard deviations from individual models and from the ensemble model for oxygen prediction when trained using DS-AL-4. Data with high standard deviations (std>30) has been excluded to make the trend more concise. All units in ppm. See Methods for further detail.

	RMSE	standard deviation
Model 1	8.30	5.23
Model 2	8.65	6.18
Model 3	16.76	15.01
Model 4	8.86	5.73
Model 5	23.34	21.72
Ensemble model	7.60	4.82

But just as importantly the ensemble model can provide standard deviations that can be used to estimate actual prediction errors even without knowing the actual ground truth for the chemical shift value. Figure 3.3b shows an undertrained model using DS-AL-4 evaluated on 8 heavy atom test data, and compares the predicted and target chemical shielding values with data points colored by the standard deviations from the ensemble. We find that when the standard deviations are small, the predictions are accurate, and correspondingly all data points with large standard deviations correlate with high predicted errors. Figure 3.3c further illustrates the correlation between the prediction standard deviation and the absolute error from the ensemble prediction. Because we find that standard deviations are good approximators for prediction error, the iShiftML model can be applied to and make good prediction for any organic chemical system by only selecting predictions with low standard deviations.

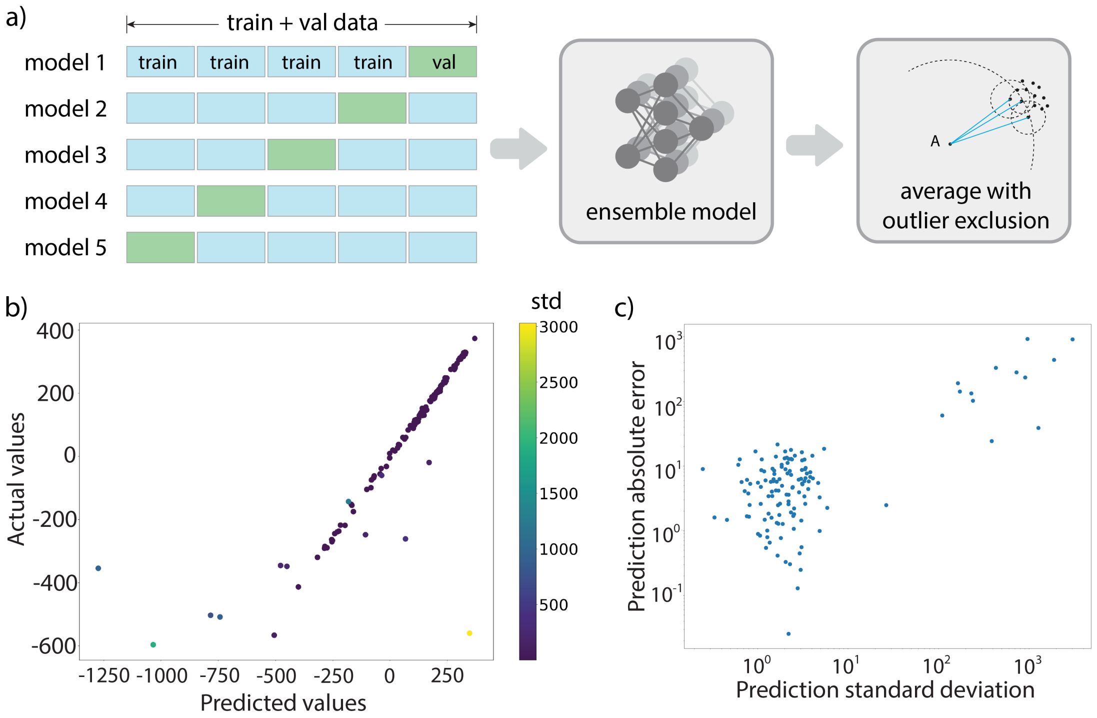


Figure 3.3: *Ensemble prediction and correlation with actual prediction error.*(a) An ensemble learning approach using 5-fold cross validation to train individual models in the ensemble. The final prediction is the average prediction from the models after excluding outliers recognized by the Local Outlier Factor algorithm [58].(b) An undertrained model for oxygen tested on the 8-heavy-atom test dataset, showing correlation between predicted and actual values. Data points are colored according to their standard deviation (STD), with warm colors representing high STDs and cool colors representing low STDs. (c) Prediction errors compared to reference values are found to be well correlated with standard deviations of the predictions in the ensemble on a log-log plot. See Methods for further detail.

Finally, the ability to identify out-of-distribution data not effectively covered by existing training data through ensemble learning has inspired a novel active learning technique to select only the most important training data to calculate time-consuming high-level chemical shieldings while still improving model performance. In particular given that the high-level QM calculation scales as $O(N^7)$ with system size, it is best to generate as many training data with smaller number of heavy atoms in order to reduce the number of calculations needed for molecules with more heavy atoms (Figure 3.4a).

In this case we start by training a model with all subsampled data with up to 4 heavy atoms (DS-AL-4) to allow sufficient initial coverage of the chemical space, and which provides a good starting point for the AL workflow. After training converges with DS-AL-4, the model was used to predict chemical shieldings on the 5 heavy atom data using the low-level QM features. Large standard deviations from the ensemble prediction were utilized to select 1500 structures to generate the next batch of high-level target chemical shieldings which are then added to the training dataset to define the next DS-AL-5 dataset. This process continues until we have included high-level calculations for molecules up to 7 heavy atoms in the training dataset.

After each AL iteration, the model performance was evaluated on the test dataset composed of randomly selected molecules with 8 heavy atoms to show the effectiveness of the AL approach. Test errors in terms of RMSE for the four atom types are visualized in Figure 3.4b-e, which show the trend of error decrease as larger molecules are added to the training dataset. As a reference, a linear regression (LR) model that uses QM features in DS-AL-7 was also trained, which acts as a baseline equivalent to a model that has fixed coefficients on the DIA and PARA terms instead of atomic environment dependent weights.

Figure 3.4b-e shows that even the model trained with DS-AL-4 surpasses the LR reference performance in all atom types other than nitrogen. With more training data included, which we emphasize is that every new training dataset only has ∼10% more data (1500 more molecules) than the dataset with one less heavy atoms, the model continues to systematically improve on the 8 heavy atom test dataset. After the model has been trained with DS-AL-7, the RMSE between predicted and actual high-level QM chemical shieldings are only 0.11 ppm for hydrogen, 1.54 ppm for carbon, 3.90 ppm for nitrogen and 6.33 ppm for oxygen, very close to the error between the target high-level method and the theoretical best estimates. An increase in the proportion of data that can be successfully predicted for this 8 heavy atom test dataset was also observed as more heavy atom data was included in the training dataset (Appendix Table 3.C.3). In the final model trained with DS-AL-7, all test data was predicted with small standard deviations and no data point was excluded from the calculation.

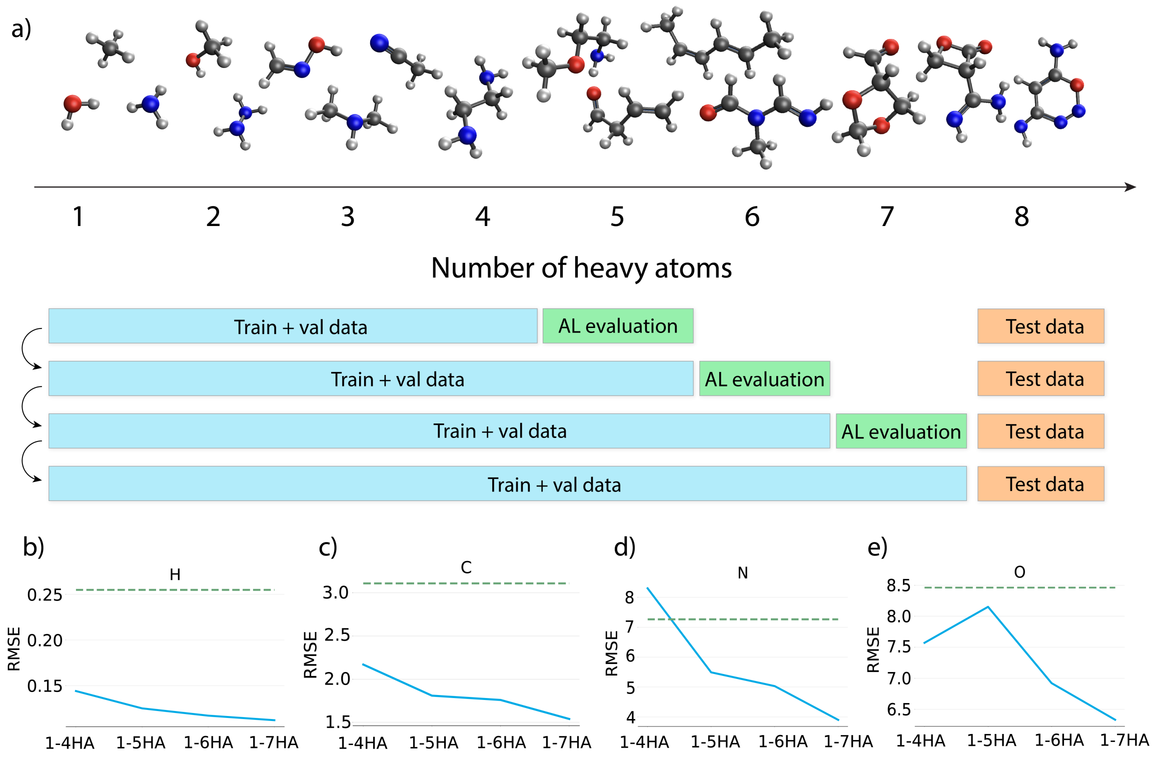


Figure 3.4: *Procedure and results of the active learning workflow.* a) The active learning (AL) workflow. Starting from a model trained with data up to 4 heavy atoms (HA), data with 5HA are evaluated using the trained model and 1500 structures with largest predicted standard deviations from 5HA dataset was included to define the training dataset for the next iteration, until the dataset contains molecules up to 7HA. The 8HA dataset was always used for test. b-e) RMSE on the 8HA test dataset for models trained with AL under training dataset containing molecules with different sizes (blue curve), and also a baseline model that is trained using linear regression (green dotted line). Figures are for hydrogens (b), carbons (c), nitrogens (d) and oxygens (e). (b-e) are also provided in tabular form in Appendix Table 3.C.2. Note that the RMSEs are calculated with uncertain predictions excluded, which removes any prediction with ensemble standard deviation larger than 30. The proportion of data that has been excluded is also listed in Appendix Table 3.C.3.

Application to predicting gas phase chemical shifts

The iShiftML model can predict NMR chemical shieldings at a high-level CCSD(T) composite method accuracy using only a tiny fraction of calculation time of a low-level DFT calculation, which enables us to explore new possibilities of experimental CS prediction as well. We first show that experimental gas phase CS for molecules not included in the training dataset can be accurately predicted and error is significantly reduced compared to the low-level DFT that provides the QM matrix elements. Gas phase chemical shifts were used to minimize the effect of environmental complexities, including any influence of solvent and perturbations to chemical shifts due to other molecules nearby. We also consider a more challenging application of the iShiftML method to highlight the transferability of the model for predicting CS for natural products that are much larger and more complex than any molecule in our training dataset. Specifically, calculated CS for 8 diaesteromers of the vannusal B molecules were compared to the experimental measurements to demonstrate that the matching structure can be confidently selected relying on our iShiftML method, which would greatly assist synthetic chemists.

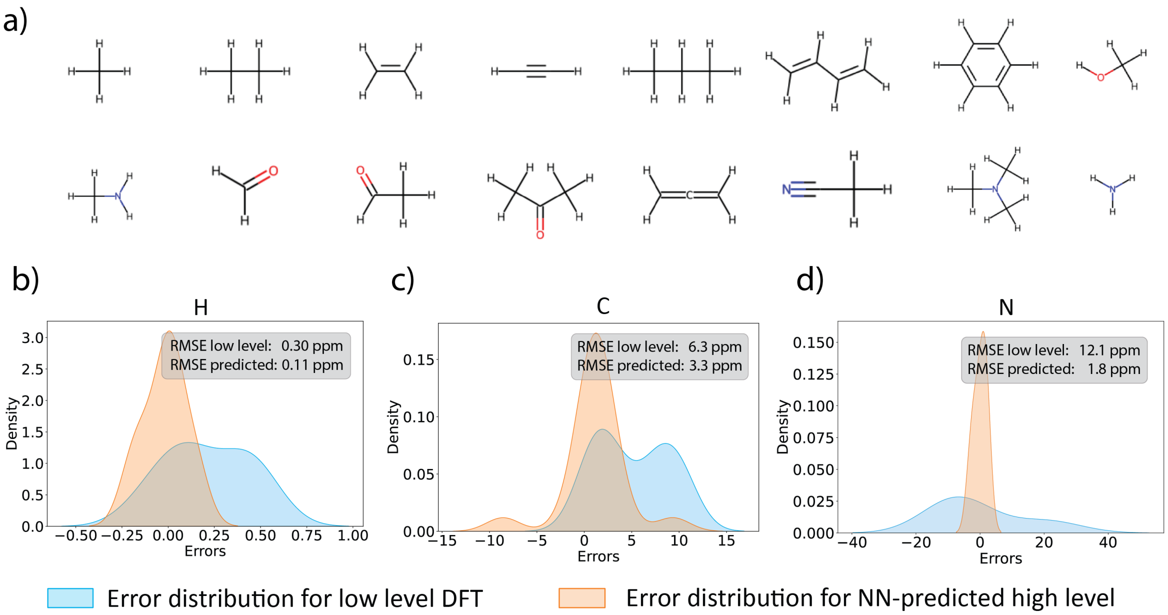


Figure 3.5: *Predicting experimental gas phase chemical shifts for small organic molecules.* (a) the small molecules under investigation. 3D geometries of these molecules are taken from NS372[63] and NIST database.[64] (b-d) Distributions of errors between predicted and experimental gas phase NMR chemical shifts for low level DFT calculations (ωB97X-V/pcSseg-1, blue distributions) and iShiftML predictions for the high level CCSD(T) composite method, orange distributions values for hydrogens (b), carbons (c) and nitrogens (d). Also see Appendix Figure 3.B.1 and Appendix Table 3.C.4.

Figure 3.5a shows a set of 16 molecules that were collected from the literature for their experimental gas phase CS values[65, 66, 67], and the geometries of the molecules were taken from NS372[63] and NIST database.[64] Because some of these molecules were already in the

DS-AL-7 data set, the iShiftML models were retrained after excluding all molecules in Figure 3.5a that are to be tested.

Chemical shifts were calculated with two techniques. For H and C, the reference chemical shieldings for the respective nuclei in the standard substance tetramethylsilicane (TMS) were calculated at the low-level ωB97X-V/pcSseg-1 and high-level CCSD(T)(1)$\cup$RIMP2(3), and chemical shifts were calculated using $\delta = \sigma_{ref} - \sigma_{nuc}$, where σ_{ref} is the isotropic chemical shielding for TMS, and σ_{nuc} is the isotropic chemical shielding for the target nucleus. Reference chemical shieldings are 31.766 ppm and 189.588 ppm using the low-level theory for hydrogen and carbon, respectively, while the references were 31.522 ppm and 193.972 ppm using the high-level theory for hydrogen and carbon, respectively. Due to lack of standard substance for nitrogen, a linear model was fit between the predicted chemical shieldings and experimental chemical shifts using a fixed slope of -1 such that only the intercept was fitted. The resulting intercept is -137.9 ppm and -128.3 ppm for the low level and high level theory, respectively. Oxygen nuclei were not assessed due to lack of experimental gas phase chemical shifts for this test set.

When compared directly to experimental measurements, we find that iShiftML can predict CS for hydrogen nuclei with RMSE of 0.11 ppm, 3.3 ppm for carbon and 1.80 ppm for nitrogen. By comparison, the low-level DFT calculations gives an RMSE of 0.30 ppm for hydrogen, 6.3 ppm for carbon and 12.1 ppm for nitrogen indicating that with an inexpensive method we have significantly reduced error by 2-6 fold. Figures 3.5b-d show the error distributions for the low-level calculated chemical shifts and high-level predicted chemical shifts, both compared with experimental CS for different nuclei. We see that the low-level CS has a systematic offset for the nuclei under investigation, resulting in error distributions shifted towards positive values for hydrogen and carbon, and negative values for nitrogen. This systematic trend was corrected in the predicted high-level CS, whose errors are centered around zero with a much sharper distribution, in line with its overall superior performance compared to the low-level DFT calculations. One carbon CS (acetylene) that had a high prediction error was also found to have high standard deviation around 13 ppm, which again shows that standard deviations give good estimates of prediction error.

Application to natural product chemical shifts prediction

Finally we consider a more challenging application of iShiftML to highlight the transferability of the model. Synthetic chemists often rely on NMR CS as an essential tool to validate the structural correctness of synthesized molecules, especially for natural products.[15] In turn, automated methods such as DP4[68] and DP4+[69] and corresponding ML advances such as DP4-AI[70] for computing NMR spectra reliably enough to confirm the chemical composition and stereochemistry of natural products are a critically important counterpart to the experimental data.[71, 72, 69, 73] Here we demonstrate that iShiftML can also improve the accuracy of predicted CS for a given molecular structure when compared with experimental measurements. We have used strychnine[74, 75, 76, 77, 72] as a starting example since it is a relatively rigid molecule (Figure 3.6a) so that conformational averaging will not play a major role in predicting its chemical shifts accurately. Figure 3.6b and c and shows the absolute errors between experimental and calculated CSs using both the low level DFT and high level predictions from the iShiftML model for hydrogens and carbons, and the correlation plots are provided in Appendix Figure 3.B.2. All iShiftML predictions were made with small standard deviations and hence no outliers were found.

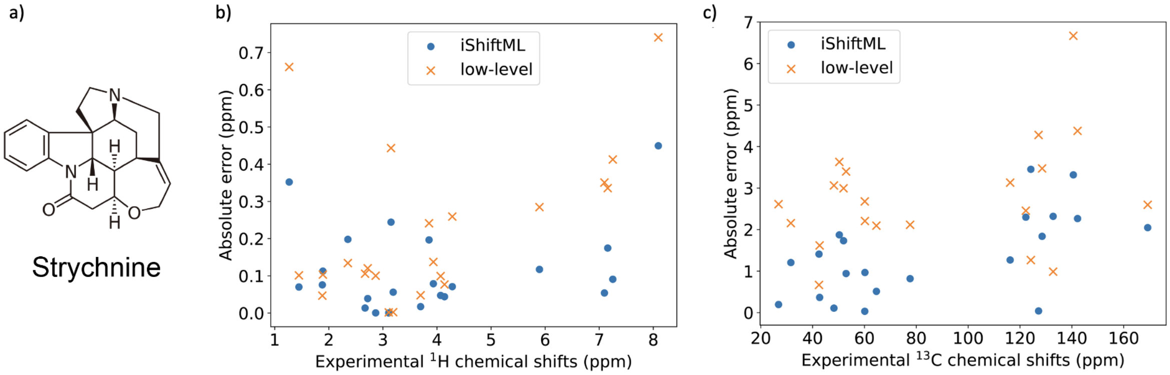


Figure 3.6: *Results on predicting and comparing CS for the strychnine natural product.* a) Molecular structure of strychine. b) Absolute prediction error for the low-level DFT method and iShiftML across the experimental CS range for hydrogens. c) Absolute prediction error for the low-level DFT method and iShiftML across the experimental CS range for carbons. To avoid any inaccuracy in the reference values, all calculated CS were re-referenced to have identical mean values to match the same reference used in the experimental CS.

The RMSEs between experiment and calculated CS for low level DFT, high level iShiftML predictions, along with four other DFT methods reported in Ref. 76 after re-referencing are also provided in Table 3.2. We find that iShiftML has significantly improved over the low level ωB97X-V/pcsSeg-1 DFT calculation that provides input for our model, and is as good or better than other DFT methods that use a much larger basis set. Hence even though strychnine is significantly larger and its fused ring system is not covered by our training dataset, we still realize significant improvements over much more expensive methods, with errors that remain commensurate with the errors of the 8HA test dataset for high level CCSD(T) calculations. This demonstrates the reliability and generalizability of our model.

Table 3.2: RMSEs between predicted and measured CS in strychnine using different methods. The 3-dimensional geometry of the strychnine molecule and the experimental measurements of CS are taken from Ref. 76. Predicted CS are re-referenced to have same mean values as experimental measurements to avoid any referencing errors. However, the slopes are fixed at unity.

Method	**H**[a]	**C**[b]
B3LYP/cc-pVTZ[c]	**0.162**	2.095
PBE1PBE/cc-pVTZ[c]	0.202	2.032
BP/TZP[c]	0.177	3.145
BP/TZ2P[c]	0.177	2.895
ωB97X-V/pcsSeg-1 (low-level)	0.296	3.068
iShiftML	**0.160**	**1.701**

[a] Experimental CS data from Ref. 74

[b] Experimental CS data from Ref. 75

[c] Refitted with unity slope using original data from Ref. 76

Finally we consider a more challenging natural product synthesis application to identify the correct molecular structure of vannusal B (5-2), whose structural assignment had been uncertain due to the errors in back-calculations and comparison to experiment of a set of highly similar diastereomers of the natural product itself (Figure 3.7a).[78, 79] Here we have use iShiftML to investigate the match between experimental and calculated CS for carbon atoms, and compare our results with the M06/pcS-2 DFT method reported in Ref. 80. However, we did not rescale predicted CS values as was done in Ref. 80, so that our reported errors reflect true prediction errors on various atoms in the molecule. Additionally, sp^2 hybridized carbons (C1, C2, C11, C12, C21, C31) were retained in our analysis, unlike the original study, as the iShiftML model should provide accurate predictions (or indicate if it is an outlier) without any prior system knowledge.

Figure 3.7b provides the RMSEs between predicted and experimental CSs for vannusal B (5-2) and same for the structures of the other diastereomers (2-1, 2-2, 3-1, 3-2, 4-1, 4-2, and 5-1). We find that iShiftML consistently predicts lower RMSE across all molecules compared with the low-level DFT method or M06/pcS-2 from Ref. 80 (i.e. the bottom of the blue bars for iShiftML are well below the bottom of the orange and green bars). Furthermore, in Figure 3.7b the bottom of each bar provides the RMSE between the experimental chemical shifts that match the true structure of each diastereomer, while the top position of the RMSE bar shows the error made if the experimental CS for natural product structure 5-2 involved an (erroneous) assignment to the diastereomer structure of interest. On average, iShiftML has a larger RMSE margin (longer bars) between the correct structure assignment of the given diastereomer and the erroneous matching (to 5-2) based on the two sets of experimental CS. Therefore iShiftML can identify the correct structure from other candidates with higher confidence, as well as recognize the true vannusal B molecule with ease.

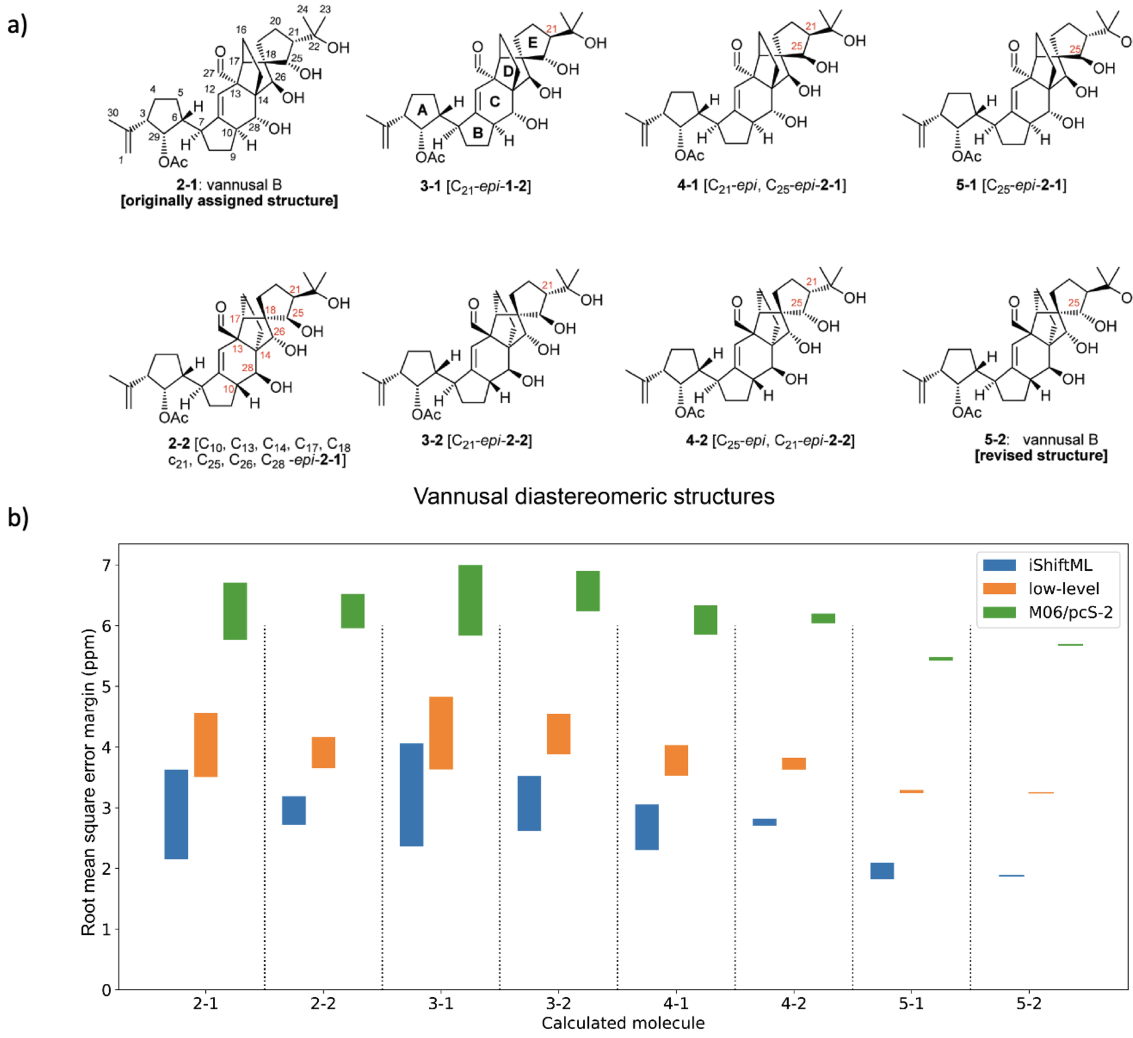


Figure 3.7: *Results on predicting and comparing CS for 8 diastereomers of vannusal B.* a) Molecular structures of the 8 diastereomers of vannusal B. b) Prediction RMSE margins when comparing various vannusal B isomers with their corresponding true experimental CS (bottom position in each bar) and comparing to the vannusal B CS in its native form, 5-2 (top position of each bar) using iShiftML, low-level DFT and M06/pcS-2 (the latter from Ref. 80). Large bars with a low bottom therefore indicate good discrimination between predicted CS for the true structure versus false identification with CS of the native structure. To avoid any inaccuracy in the reference values, all calculated CS were re-referenced to have identical mean values to match the same reference used in the experimental CS.

3.4 Conclusion

Methods for *ab initio* calculation of chemical shieldings lie on a spectrum, with one end being DFT calculations that are cheap but less accurate, and the other end being CCSD(T)/CBS methods that are highly accurate but prohibitively expensive for large systems. We have now created a tool to bridge the two ends using machine learning, so that with input features coming from a relatively fast DFT calculation, the predictions can approach the highest level of accuracy achieveable through quantum mechanics calculations, without incurring extra cost. By utilizing a feature set that relies on chemical shielding DIA and PARA tensor components, together with features that describe molecular geometry, we demonstrated that iShiftML can achieve not only excellent accuracy compared to the high level target chemical shieldings, but greater transferability to test molecules larger than any molecule contained in our training dataset, approaching the intrinsic errors for the high level targets when compared to CCSD(T)/CBS calculations.

While iShiftML is readily helpful for those who study the chemical shieldings of small organic molecules using coupled cluster methods, its broader applicability is exemplified with predicting experimental chemical shifts with higher accuracy for arbitrary systems. Our trained model without any fine-tuning can predict gas phase experimental chemical shifts for small organic molecules with excellent accuracy and reduce error by more than 50% compared to the direct calculation using the same level of QM theory as our input features. When applying this method to synthesized natural products, we illustrated it could achieve better agreement between predicted and measured chemical shifts when the structures match, and provide better differentiation capability between matched and mismatched diastereomer structures given the CS experimental data. We believe there are many more application possibilities of our method, including predicting chemical shifts for proteins, correcting assignment errors in databases, and aiding drug discovery in determining structure-activity relationships.

There are also some limitations of the current method. It is trained with equilibrium and non-equilibrium geometries of closed-shell small organic molecules that contain only H, C, N and O atoms. Also, only single molecule data were included in our training dataset. Therefore it is not expected to work for open-shell molecules, molecules containing other elements, or for molecular systems in which intermolecular interactions play a major role in the chemical shifts. However, we are planning to improve the method in the future to make it even more transferable and widely applicable. For example, adding support for more atom types will be our first step to allow this method to work for a broader range of organic compounds. Nevertheless, we believe in its current form iShiftML can already benefit those in need of a fast and reliable chemical shift predictor.

3.5 Acknowledgements

This work was supported by funding from the National Institute of General Medical Sciences under grant number 5U01GM121667. This research used computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. A. L. Ptaszek was funded by the Christian Doppler Laboratory for High-Content Structural Biology and Biotechnology, Austria and received further support from the DosChem doctoral school program faculty of Chemistry, University of Vienna.

3.6 References

[1] Neil E Jacobsen. *NMR data interpretation explained: understanding 1D and 2D NMR spectra of organic compounds and natural products.* John Wiley & Sons, 2016.

[2] Peter J Hore. *Nuclear magnetic resonance.* Oxford University Press, USA, 2015.

[3] Andrew E Derome. *Modern NMR techniques for chemistry research.* Elsevier, 2013.

[4] Alessandro Bagno, Federico Rastrelli, and Giacomo Saielli. Toward the complete prediction of the 1h and 13c nmr spectra of complex organic molecules by dft methods: application to natural substances. *Chem. Eur. J.*, 12(21):5514–5525, 2006.

[5] Giacomo Saielli, KC Nicolaou, Adrian Ortiz, Hongjun Zhang, and Alessandro Bagno. Addressing the stereochemistry of complex organic molecules by density functional theory-nmr: Vannusal b in retrospective. *J. Am. Chem. Soc.*, 133(15):6072–6077, 2011.

[6] Kurt Wüthrich. Protein structure determination in solution by nmr spectroscopy. *J. Biol. Chem.*, 265(36):22059–22062, 1990.

[7] Dominique Marion. An introduction to biological nmr spectroscopy. *Mol. Struct. Proteom.*, 12(11):3006–3025, 2013.

[8] Steven P Brown. Applications of high-resolution 1h solid-state nmr. *Solid State NMR*, 41:1–27, 2012.

[9] Kenneth JD MacKenzie and Mark E Smith. *Multinuclear solid-state nuclear magnetic resonance of inorganic materials.* Elsevier, 2002.

[10] Federico M Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Emsley. Chemical shifts in molecular solids by machine learning. *Nature Comm.*, 9(1):4501, 2018.

[11] Giampaolo Barone, Luigi Gomez-Paloma, Dario Duca, Arturo Silvestri, Raffaele Riccio, and Giuseppe Bifulco. Structure validation of natural products by quantum-mechanical giao calculations of 13c nmr chemical shifts. *Chem. Eur. J.*, 8(14):3233–3239, 2002.

[12] Lars A Bratholm and Jan H Jensen. Protein structure refinement using a quantum mechanics-based chemical shielding predictor. *Chem. Sci.*, 8(3):2061–2072, 2017.

[13] Trygve Helgaker, Michal Jaszunski, and Kenneth Ruud. Ab initio methods for the calculation of nmr shielding and indirect spin-spin coupling constants. *Chem. Rev.*, 99:293–352, 1999.

[14] Jürgen Gauss and John F Stanton. Electron-correlated approaches for the calculation of nmr chemical shifts. *Adv. Chem. Phys.*, 123:355–422, 2002.

[15] Michael W Lodewyk, Matthew R Siebert, and Dean J Tantillo. Computational prediction of 1h and 13c chemical shifts: a useful tool for natural product, mechanistic, and synthetic organic chemistry. *Chem. Rev.*, 112(3):1839–1862, 2012.

[16] Graham A Webb. *Modern magnetic resonance: Part 1: Applications in chemistry, biological and marine sciences, Part 2: Applications in medical and pharmaceutical sciences, Part 3: Applications in materials science and food science.* Springer Science & Business Media, 2007.

[17] Jürgen Gauss. Analytic second derivatives for the full coupled-cluster singles, doubles, and triples model: Nuclear magnetic shielding constants for bh, hf, co, n 2, n 2 o, and o 3. *J. Chem. Phys.*, 116(12):4773–4776, 2002.

[18] Andrew M Teale, Ola B Lutnæs, Trygve Helgaker, David J Tozer, and Jürgen Gauss. Benchmarking density-functional theory calculations of nmr shielding constants and spin–rotation constants using accurate coupled-cluster calculations. *J. Chem. Phys.*, 138(2):024111, 2013.

[19] Caspar Jonas Schattenberg and Martin Kaupp. Extended benchmark set of main-group nuclear shielding constants and nmr chemical shifts and its use to evaluate modern dft methods. *J. Chem. Theory Comput.*, 17(12):7602–7621, 2021.

[20] David M Reid and Michael A Collins. Approximating ccsd (t) nuclear magnetic shielding calculations using composite methods. *J. Chem. Theory Comput.*, 11(11):5177–5181, 2015.

[21] Jiashu Liang, Zhe Wang, Jie Li, Jonathan Wong, Xiao Liu, Brad Ganoe, Teresa Head-Gordon, and Martin Head-Gordon. Efficient calculation of nmr shielding constants using composite method approximations and locally dense basis sets. *J. Chem. Theory Comput.*, 19:514–523, 2023.

[22] Jonathan Wong, Brad Ganoe, Xiao Liu, Tim Neudecker, Joonho Lee, Jiashu Liang, Zhe Wang, Jie Li, Adam Rettig, Teresa Head-Gordon, et al. An in-silico nmr laboratory for nuclear magnetic shieldings computed via finite fields: Exploring nucleus-specific renormalizations of mp2 and mp3. *J. Chem. Phys.*, 158(16):164116, 2023.

[23] Yang Shen and Ad Bax. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR*, 48:13–22, 2010.

[24] Beomsoo Han, Yifeng Liu, Simon W Ginzinger, and David S Wishart. Shiftx2: significantly improved protein chemical shift prediction. *J. Biomol. NMR*, 50:43–57, 2011.

[25] Jie Li, Kochise C Bennett, Yuchen Liu, Michael V Martin, and Teresa Head-Gordon. Accurate prediction of chemical shifts for aqueous protein structure on "real world" data. *Chem. Sci.*, 11(12):3180–3191, 2020.

[26] Shuai Liu, Jie Li, Kochise C Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. Multiresolution 3d-densenet for chemical shift prediction in nmr crystallography. *J. Phys. Chem. Lett.*, 10(16):4558–4565, 2019.

[27] Stefan Kuhn and Nils 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. *Mag. Res. Chem.*, 53(8):582–589, 2015.

[28] Yanfei Guan, SV Shree Sowndarya, Liliana C Gallegos, Peter C St John, and Robert S Paton. Real-time prediction of 1 h and 13 c chemical shifts with dft accuracy using a 3d graph neural network. *Chem. Sci.*, 12(36):12012–12026, 2021.

[29] Trygve Helgaker, Michal Jaszunski, and Kenneth Ruud. Ab initio methods for the calculation of nmr shielding and indirect spin-spin coupling constants. *Chem. Rev.*, 99:293–352, 1999.

[30] James R Cheeseman, Gary W Trucks, Todd A Keith, and Michael J Frisch. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J. Chem. Phys.*, 104(14):5497–5509, 1996.

[31] Denis Flaig, Marina Maurer, Matti Hanni, Katharina Braunger, Leonhard Kick, Matthias Thubauville, and Christian Ochsenfeld. Benchmarking hydrogen and carbon nmr chemical shifts at hf, dft, and mp2 levels. *J. Chem. Theory Comput.*, 10(2):572–578, 2014.

[32] Pablo A Unzueta, Chandler S Greenwell, and Gregory JO Beran. Predicting density functional theory-quality nuclear magnetic resonance chemical shifts via δ-machine learning. *J. Chem. Theory Comput.*, 17(2):826–840, 2021.

[33] Julius B. Kleine Büning and Stefan Grimme. Computation of ccsd(t)-quality nmr chemical shifts via δ-machine learning from dft. *J. Chem. Theory Comput.*, 0(0):null, 0. PMID: 37262324.

[34] Mojtaba Haghighatlari, Jie Li, Farnaz Heidar-Zadeh, Yuchen Liu, Xingyi Guan, and Teresa Head-Gordon. Learning to make chemical predictions: The interplay of feature representation, data, and machine learning methods. *Chem*, 6(7):1527–1542, 2020.

[35] Matthew Welborn, Lixue Cheng, and Thomas F Miller III. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.*, 14(9):4772–4779, 2018.

[36] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data*, 4(1):1–8, 2017.

[37] Krzysztof Wolinski, James F. Hinton, and Peter Pulay. Efficient implementation of the gauge-independent atomic orbital method for nmr chemical shift calculations. *J. Am. Chem. Soc.*, 112(23):8251–8260, 1990.

[38] Georg Schreckenbach and Tom Ziegler. Calculation of nmr shielding tensors using gauge-including atomic orbitals and modern density functional theory. *J. Phys. Chem.*, 99(2):606–611, 1995.

[39] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.

[40] Narbe Mardirossian and Martin Head-Gordon. ωb97x-v: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.*, 16(21):9904–9924, 2014.

[41] Frank Jensen. Segmented contracted basis sets optimized for nuclear magnetic shielding. *J. Chem. Theory Comput.*, 11(1):132–138, 2015.

[42] Narbe Mardirossian and Martin Head-Gordon. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.*, 115(19):2315–2372, June 2017.

[43] Lars Goerigk, Andreas Hansen, Christoph Bauer, Stephan Ehrlich, Asim Najibi, and Stefan Grimme. A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.*, 19(48):32184–32215, 2017.

[44] Diptarka Hait and Martin Head-Gordon. How accurate is density functional theory at predicting dipole moments? an assessment using a new database of 200 benchmark values. *J. Chem. Theory Comput.*, 14(4):1969–1981, March 2018.

[45] Sebastian Dohm, Andreas Hansen, Marc Steinmetz, Stefan Grimme, and Marek P Checinski. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *J. Chem. Theory Comput.*, 14(5):2596–2608, 2018.

[46] Srimukh Prasad Veccham and Martin Head-Gordon. Density functionals for hydrogen storage: Defining the h2bind275 test set with ab initio benchmarks and assessment of 55 functionals. *J. Chem. Theory Comput.*, 16(8):4963–4982, 2020.

[47] Minho Kim, Tim Gould, Ekaterina I Izgorodina, Dario Rocca, and Sébastien Lebègue. Establishing the accuracy of density functional approaches for the description of noncovalent interactions in ionic liquids. *Phys. Chem. Chem. Phys.*, 23(45):25558–25564, 2021.

[48] Diptarka Hait, Yu Hsuan Liang, and Martin Head-Gordon. Too big, too small, or just right? a benchmark assessment of density functional theory for predicting the spatial extent of the electron density of small chemical systems. *J. Chem. Phys.*, 154(7):074109, 2021.

[49] Jiashu Liang, Xintian Feng, Diptarka Hait, and Martin Head-Gordon. Revisiting the performance of time-dependent density functional theory for electronic excitations: Assessment of 43 popular and recently developed functionals from rungs one to four. *J. Chem. Theory Comput.*, 18(6):3460–3473, 2022.

[50] Frank Neese, Frank Wennmohs, Ute Becker, and Christoph Riplinger. The orca quantum chemistry program package. *J. Chem. Phys*, 152(22):224108, 2020.

[51] Devin A Matthews, Lan Cheng, Michael E Harding, Filippo Lipparini, Stella Stopkowicz, Thomas-C Jagau, Péter G Szalay, Jürgen Gauss, and John F Stanton. Coupled-cluster techniques for computational chemistry: The cfour program package. *J. Chem. Phys*, 152(21):214108, 2020.

[52] Cfour, a quantum chemical program package written by stanton, j. f.; gauss, j.; cheng, l.; harding, m. e.; matthews, d. a.; szalay, p. g. with contributions from auer, a. a., bartlett, r. j., benedikt, u., berger, c., bernholdt, d. e., bomble, y. j., christiansen, o., engel, f., faber, r., heckert, m., heun, o., hilgenberg, m., huber, c., jagau, t.-c., jonsson, d., jusólius, j., kirsch, t., klein, k., lauderdale, w. j., lipparini, f., metzroth, t., mück, l. a., o'neill, d. p., price, d. r., prochnow, e., puzzarini, c., ruud, k., schiffmann, f., schwalbach, w., simmons, c., stopkowicz, s., tajti, a., vázquez, j., wang, f., watts, j. d. and the integral packages molecule (almlöf j. and taylor p.r.), props (taylor p.r.), abacus (helgaker t., jensen h.j. aa., jørgensen p., and olsen j.), and ecp routines by mitin a. v. and wüllen c. van. http://www.cfour.de (accessed Sep 1, 2022).

[53] Michael E Harding, Thorsten Metzroth, Jürgen Gauss, and Alexander A Auer. Parallel calculation of ccsd and ccsd (t) analytic first and second derivatives. *J. Chem. Theory Comput.*, 4(1):64–74, 2008.

[54] Florian Weigend. Hartree–fock exchange fitting basis sets for h to rn. *J. Comput. Chem.*, 29(2):167–175, 2008.

[55] Christof Hättig. Optimization of auxiliary basis sets for ri-mp2 and ri-cc2 calculations: Core–valence and quintuple-ζ basis sets for h to ar and qzvpp basis sets for li to kr. *Phys. Chem. Chem. Phys.*, 7(1):59–66, 2005.

[56] Rolf Seeger and John A Pople. Self-consistent molecular orbital methods. xviii. constraints and stability in hartree–fock theory. *J. Chem. Phys.*, 66(7):3045–3050, 1977.

[57] Marko Melander, Kari Laasonen, and Hannes Jonsson. Removing external degrees of freedom from transition-state search methods using quaternions. *J. Chem. Theory Comput.*, 11(3):1055–1062, 2015.

[58] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[59] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. ML Res.*, 15(1):1929–1958, 2014.

[60] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[62] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[63] Caspar Jonas Schattenberg and Martin Kaupp. Extended benchmark set of main-group nuclear shielding constants and nmr chemical shifts and its use to evaluate modern dft methods. *J. Chem. Theory Comput.*, 17(12):7602–7621, 2021.

[64] Russell D Johnson et al. Nist computational chemistry comparison and benchmark database. *http://srdata. nist. gov/cccbdb*, 2006.

[65] Thomas Zuschneid, Holger Fischer, Thomas Handel, Klaus Albert, and Günter Häfelinger. Experimental gas phase 1h nmr spectra and basis set dependence of ab initio giaomo calculations of 1h and 13c nmr absolute shieldings and chemical shifts of small hydrocarbons. *Z. Naturforsch. B*, 59(10):1153–1176, 2004.

[66] James R Cheeseman, Gary W Trucks, Todd A Keith, and Michael J Frisch. A comparison of models for calculating nuclear magnetic resonance shielding tensors. *J. Chem. Phys.*, 104(14):5497–5509, 1996.

[67] Adriana Gregušová, S Ajith Perera, and Rodney J Bartlett. Accuracy of computed 15n nuclear magnetic resonance chemical shifts. *J. Chem. Theory Comput.*, 6(4):1228–1239, 2010.

[68] Steven G Smith and Jonathan M Goodman. Assigning stereochemistry to single diastereoisomers by giao nmr calculation: The dp4 probability. *J. Am. Chem. Soc.*, 132(37):12946–12959, 2010.

[69] Maribel O Marcarino, Soledad Cicetti, María M Zanardi, and Ariel M Sarotti. A critical review on the use of dp4+ in the structural elucidation of natural products: the good, the bad and the ugly. a practical guide. *Nat. Prod. Rep.*, 39(1):58–76, 2022.

[70] Alexander Howarth, Kristaps Ermanis, and Jonathan M. Goodman. Dp4-ai automated nmr data analysis: straight from spectrometer to structure. *Chem. Sci.*, 11(17):4351–4359, 2020.

[71] Alessandro Bagno and Giacomo Saielli. Addressing the stereochemistry of complex organic molecules by density functional theory-nmr. *WIREs Comput. Mol. Sci.*, 5(2):228–240, 2015.

[72] Valentin A Semenov and Leonid B Krivdin. Dft computational schemes for 1h and 13c nmr chemical shifts of natural products, exemplified by strychnine. *Magn. Reson. Chem.*, 58(1):56–64, 2020.

[73] Callum I. MacGregor, Bing Yuan Han, Jonathan M. Goodman, and Ian Paterson. Toward the stereochemical assignment and synthesis of hemicalide: Dp4f giao-nmr analysis and synthesis of a reassigned c16–c28 subunit. *Chem. Comm.*, 52(25):4632–4635, 2016.

[74] James C Carter, George W Luther III, and Thomas C Long. Proton magnetic resonance spectra and assignments of strychnine and selectively deuterated strychnine. *J. Magn. Res.*, 15(1):122–131, 1974.

[75] Gary E Martin, Chad E Hadden, Ronald C Crouch, and VV Krishnamurthy. Accord-hmbc: advantages and disadvantages of static versus accordion excitation. *Magn. Reson. Chem.*, 37(8):517–528, 1999.

[76] Alessandro Bagno, Federico Rastrelli, and Giacomo Saielli. Toward the complete prediction of the 1h and 13c nmr spectra of complex organic molecules by dft methods: application to natural substances. *Chem. Eur. J.*, 12(21):5514–5525, 2006.

[77] Jeffrey I Seeman and Dean J Tantillo. From decades to minutes: steps toward the structure of strychnine 1910–1948 and the application of today's technology. *Ang. Chem. Int. Ed.*, 59(27):10702–10721, 2020.

[78] KC Nicolaou, Adrian Ortiz, Hongjun Zhang, Philippe Dagneau, Andreas Lanver, Michael P Jennings, Stellios Arseniyadis, Raffaella Faraoni, and Dimitrios E Lizos. Total synthesis and structural revision of vannusals a and b: Synthesis of the originally assigned structure of vannusal b. *J. Am. Chem. Soc.*, 132(20):7138–7152, 2010.

[79] KC Nicolaou, Adrian Ortiz, Hongjun Zhang, and Graziano Guella. Total synthesis and structural revision of vannusals a and b: synthesis of the true structures of vannusals a and b. *J. Am. Chem. Soc.*, 132(20):7153–7176, 2010.

[80] Giacomo Saielli, KC Nicolaou, Adrian Ortiz, Hongjun Zhang, and Alessandro Bagno. Addressing the stereochemistry of complex organic molecules by density functional theory-nmr: Vannusal b in retrospective. *J. Am. Chem. Soc.*, 133(15):6072–6077, 2011.

[81] Stefan Grimme. Semiempirical gga-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.*, 27(15):1787–1799, 2006.

[82] Narbe Mardirossian and Martin Head-Gordon. Mapping the genome of meta-generalized gradient approximation density functionals: The search for b97m-v. *J. Chem. Phys*, 142(7):074111, 2015.

[83] Thomas W Keal and David J Tozer. A semiempirical generalized gradient approximation exchange-correlation functional. *J. Chem. Phys*, 121(12):5654–5660, 2004.

[84] Jianwei Sun, Adrienn Ruzsinszky, and John P Perdew. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.*, 115(3):036402, 2015.

Appendix

3.A Dataset and code links

GitHub repository link:
https://github.com/THGLab/iShiftML

DS-SS (subsampled dataset from ANI-1 with unstable molecules excluded):
https://github.com/THGLab/iShiftML/blob/master/dataset/DS-SS.txt

DS-AL (active learning dataset):
https://github.com/THGLab/iShiftML/blob/master/dataset/DS-AL.txt

Removed chemical shielding:
8_atom/mol_34274/99.xyz/atom_6 (calculated low level chemical shielding: -2.066, calculated high level chemical shielding: 197.792)

3.B Supporting Figures

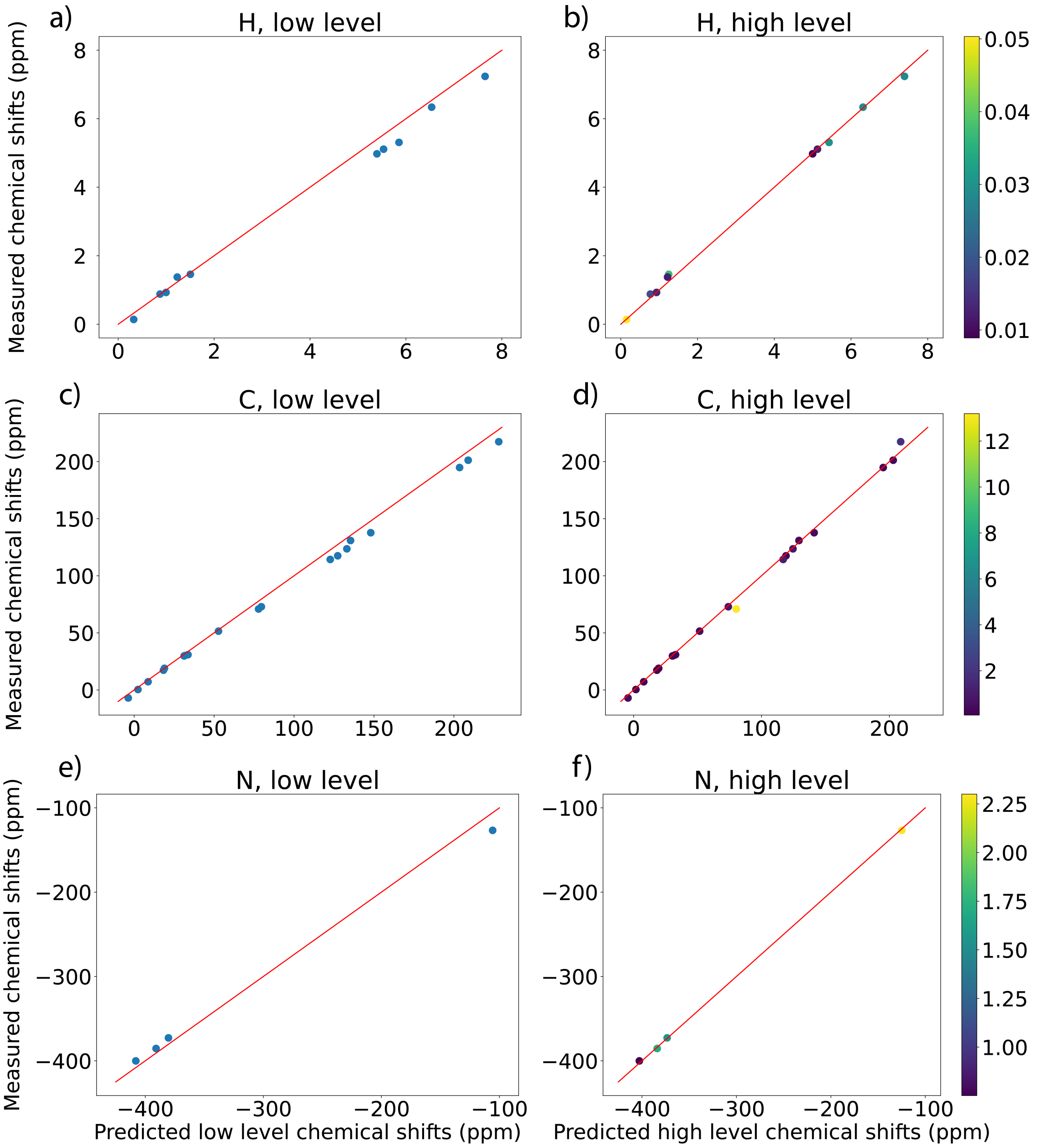


Figure 3.B.1: *Scatter plots on low-level and iShiftML predicted high-level chemical shifts compared to experimentally measured gas phase chemical shifts for different atom types.* a), c) and e) show the calculated chemical shifts using low level ωB97X-V/pcSseg-1 DFT methods, while b), d) and f) show the predicted chemical shifts using iShiftML targeting CCSD(T)(1)$\cup$RIMP2(3) composite method accuracy, with data points colored by the standard deviation from the ensemble using color codes on the right. Red lines in the figures represent y=x.

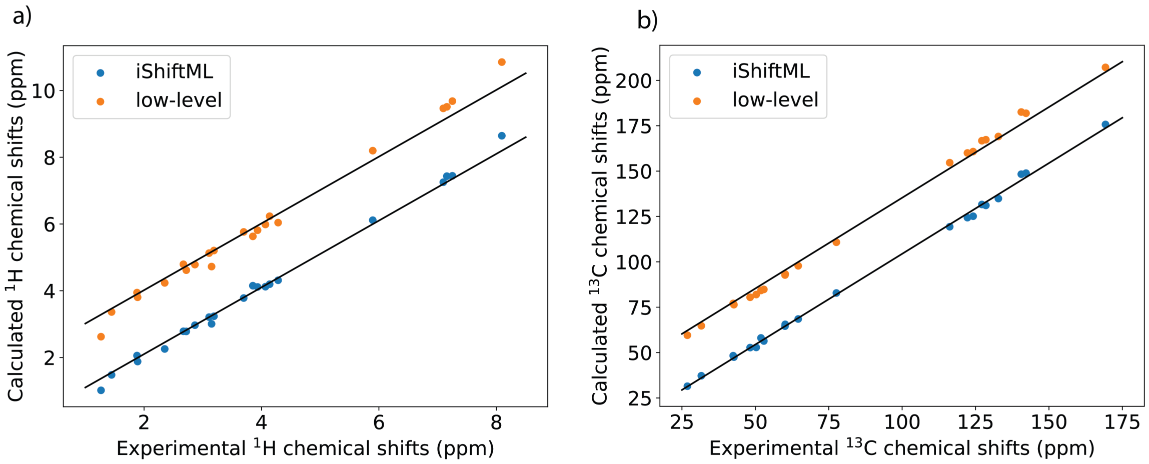


Figure 3.B.2: *Scatter plots on calculated chemical shifts compared to experimental chemical shifts using low-level DFT and iShiftML for strychnine.* a) Scatter plots for hydrogens. b) Scatter plots for carbons. Black lines represent y=x. Data for low-level DFT are shifted on the y axis for 2 ppm (a) and 30 ppm (b) to make comparisons more clear.

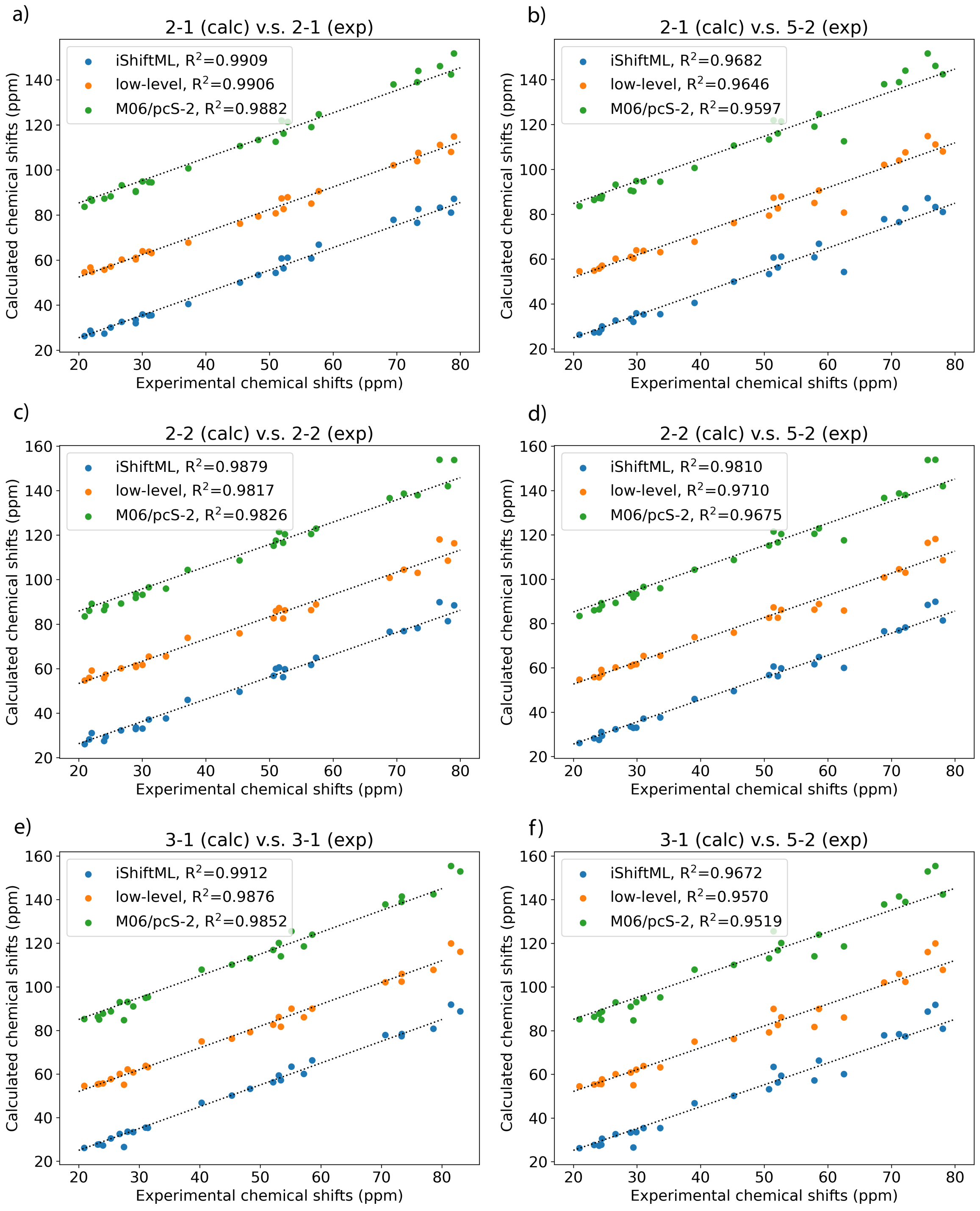

a)
2-1 (calc) v.s. 2-1 (exp)
iShiftML, R^2=0.9909
low-level, R^2=0.9906
M06/pcS-2, R^2=0.9882
b)
2-1 (calc) v.s. 5-2 (exp)
iShiftML, R^2=0.9682
low-level, R^2=0.9646
M06/pcS-2, R^2=0.9597
c)
2-2 (calc) v.s. 2-2 (exp)
iShiftML, R^2=0.9879
low-level, R^2=0.9817
M06/pcS-2, R^2=0.9826
d)
2-2 (calc) v.s. 5-2 (exp)
iShiftML, R^2=0.9810
low-level, R^2=0.9710
M06/pcS-2, R^2=0.9675
e)
3-1 (calc) v.s. 3-1 (exp)
iShiftML, R^2=0.9912
low-level, R^2=0.9876
M06/pcS-2, R^2=0.9852
f)
3-1 (calc) v.s. 5-2 (exp)
iShiftML, R^2=0.9672
low-level, R^2=0.9570
M06/pcS-2, R^2=0.9519
Calculated chemical shifts (ppm)
Experimental chemical shifts (ppm)
20 40 60 80 100 120 140 160
20 30 40 50 60 70 80

Figure 3 (continued)

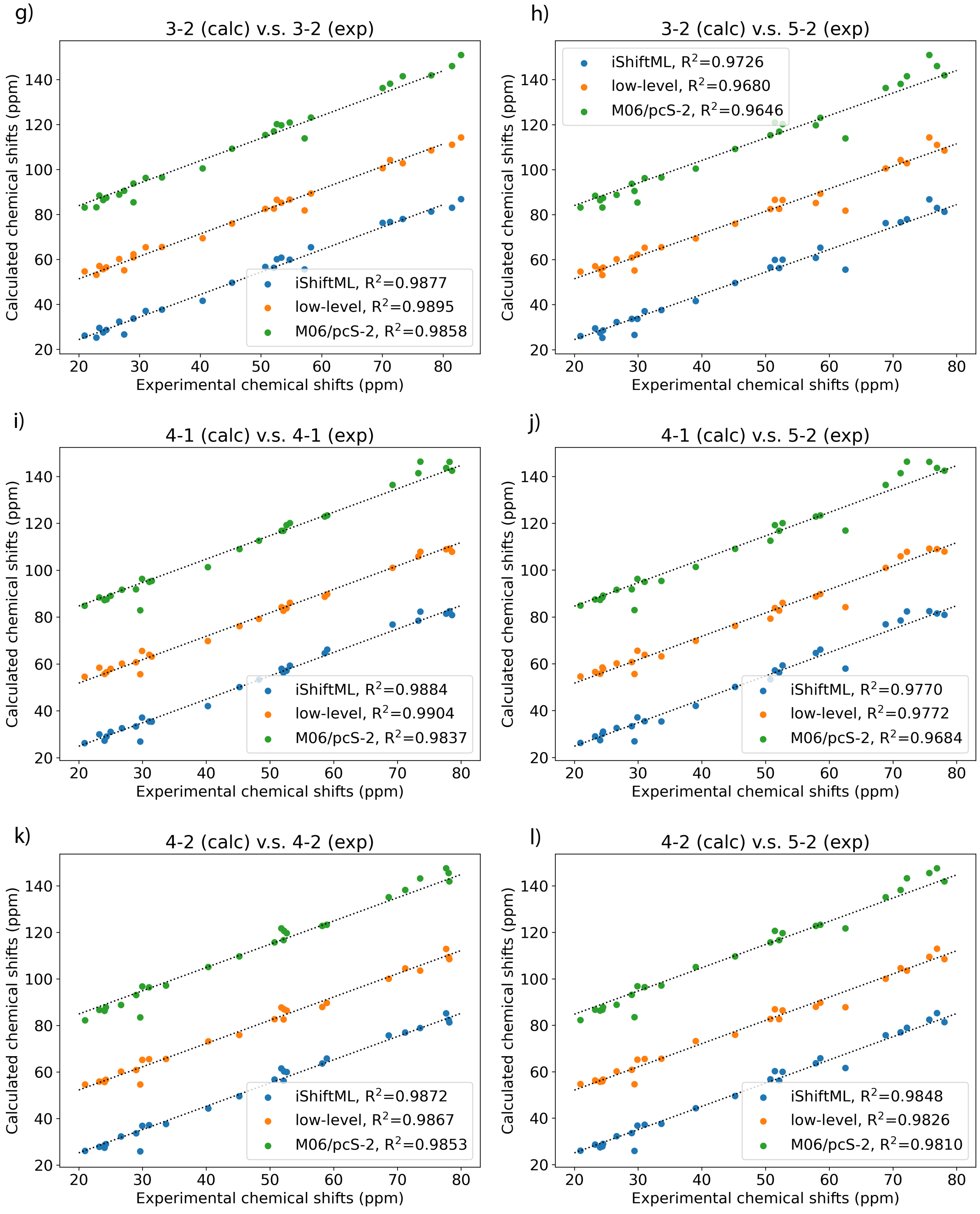


Figure 3 (continued)

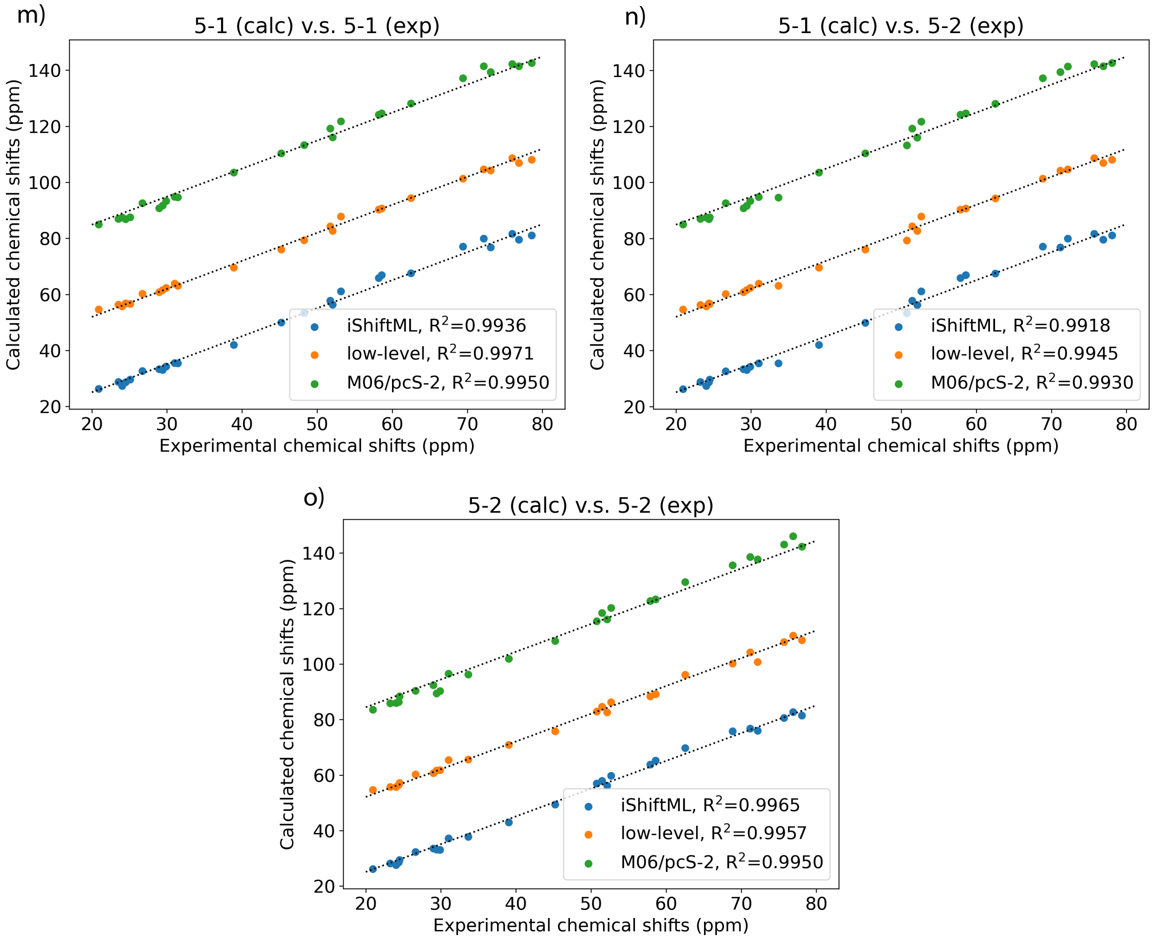


Figure 3.B.3: Scatter plots on calculated carbon chemical shifts compared to experimental chemical shifts using low-level DFT, iShiftML and M06/pcS-2 for pairs of comparison between diastereomers of vannusal B.

3.C Supporting Tables

Table 3.C.1: Prediction root mean square errors (RMSE) on in-distribution dataset (ID, randomly selected non-equilibrium geometries from training dataset containing up to 5 heavy atoms which are excluded from training), and out-of-distribution dataset (OOD, 5 heavy atom dataset with equilibrium geometries) from models trained using QM features from different low-level methods using pcSseg-1 basis sets. All units in ppm.

Low level QM method	**H**		**C**		**N**		**O**	
	ID	**OOD**	**ID**	**OOD**	**ID**	**OOD**	**ID**	**OOD**
B97-D[81]	0.05	0.14	0.78	2.97	2.4	6.0	6.8	12.7
B97M-V[82]	0.05	0.14	1.08	2.88	2.3	6.3	6.0	13.0
KT3[83]	0.05	0.14	0.88	2.96	2.5	7.2	7.6	13.0
SCAN[84]	**0.04**	0.14	0.66	2.81	1.4	5.1	4.3	56.6
ωB97X-V[40]	**0.04**	**0.12**	**0.34**	**2.66**	**1.0**	**4.0**	**4.2**	**10.1**

Table 3.C.2: RMSE on the 8 heavy atom test dataset for models trained with active learning under training dataset containing molecules with different sizes, and also a baseline model that is trained using linear regression (LR). Models trained with DS-AL-N are called 1-N HA models in the table. Outliers (data with predicted standard deviations>30) are excluded from the RMSE calculation, and the proportion of outliers is provided in Table 3.C.3

Model type	**Training dataset size**	**H**	**C**	**N**	**O**
LR(baseline)	17178	0.255	3.11	7.26	8.46
1-4HA	12677	0.144	2.17	8.30	7.57
1-5HA	14177	0.125	1.81	5.49	8.15
1-6HA	15676	0.117	1.76	5.03	6.92
1-7HA	17178	0.112	1.54	3.90	6.33

Table 3.C.3: Proportion of outliers (data with predicted standard deviations>30) for chemical shielding predictions on the 8 heavy atom test dataset using different models during AL training.

Model type	**H**	**C**	**N**	**O**
1-4HA	0	0	2.1%	10.1%
1-5HA	0	0	1.4%	3.9%
1-6HA	0	0	0.7%	0.7%
1-7HA	0	0	0	0

Table 3.C.4: Calculated and predicted NMR chemical shifts for small organic molecules and comparison to experimental gas phase chemical shift measurements. All units in ppm.

Molecule	Atom	Exp. CS	Calc. low level CS	Pred. high level CS	Pred. std
CH4	H	0.14	0.32	0.15	0.05
C2H6	H	0.88	0.87	0.77	0.02
C2H4	H	5.31	5.86	5.43	0.03
C2H2	H	1.46	1.51	1.25	0.04
C3H8	H (CH3)	0.93	1.00	0.93	0.01
C3H8	H (CH2)	1.38	1.23	1.22	0.01
Butadiene	H (CH2, trans)	4.98	5.40	5.00	0.009
Butadiene	H (CH2, cis)	5.11	5.53	5.13	0.01
Butadiene	H (CH)	6.34	6.54	6.31	0.03
Benzene	H	7.24	7.65	7.39	0.03
CH4	C	-7.0	-3.7	-4.4	0.4
C2H6	C	7.2	8.7	7.9	0.2
C2H4	C	123.6	133.0	124.6	0.9
C2H2	C	70.9	77.8	80.2	13.2
C3H8	C (CH3)	17.3	18.3	18.2	0.1
C3H8	C(CH2)	19.0	18.9	19.6	0.1
Butadiene	C (CH2)	117.5	127.3	119.2	0.3
Butadiene	C (CH)	137.7	148.0	141.2	0.4
Benzene	C	130.9	135.3	129.3	0.5
CH3OH	C	51.5	52.7	51.6	0.5
CH3NH2	C	29.8	31.2	30.4	0.2
CH3CHO	C (CH3)	30.9	33.7	32.8	0.4
CH3CHO	C (CHO)	194.8	203.5	195.2	0.2
CH3COCH3	C (CH3)	30.1	31.6	31.2	0.3
CH3COCH3	C (CO)	201.2	208.9	203.0	0.3
CH2CCH2	C (CH2)	72.9	79.6	74.1	0.8
CH2CCH2	C	217.4	228.1	208.8	1.9
CH3CN	C (CH3)	0.4	2.4	1.8	0.2
CH3CN	C (CN)	114.3	122.6	117.0	1.2
CH3NH2	N	-385.4	-390.9	-383.9	1.7
CH3CN	N	-126.7	-105.8	-124.9	2.3
N(CH3)3	N	-372.8	-380.4	-373.4	1.6
NH3	N	-400.1	-408.0	-402.8	0.8

CHAPTER 4

Learning Correlations between Internal Coordinates to Improve 3D Cartesian Coordinates for Proteins[†]

4.1 Introduction

Biomolecular structures are described using two widely used mathematical representations: internal coordinates and Cartesian coordinates. The internal coordinate representation is defined by a set of bond lengths, bond angles, and dihedral or torsion angles, and provides a compact description in terms of the Z-matrix. In contrast, a Cartesian representation defines all of the atomic positions in Euclidean x,y,z coordinates and additionally captures the orientation of a molecule in space. Both representations are useful in certain contexts and applications. Internal coordinates can be beneficial for geometry optimizations[1] and are the preferred description for NMR structure determination and refinement as an intermediate step towards an atomistic structure. The bond lengths and bond angles are typically taken as fixed[2] in these scenarios. Cartesian coordinates are the preferred format of molecular dynamics simulations[3] and X-ray crystallography, NMR, and cryo-EM structures deposited in the Protein Data Bank (PDB) repository[4].

Figure 4.1 considers the internal coordinates of a protein backbone that contains the three torsion angles ϕ $(C - N - C_\alpha - C)$, ψ $(N - C_\alpha - C - N)$, and ω $(C_\alpha - C - N - C_\alpha)$, bond lengths $N - C_\alpha$ (d_1), $C_\alpha - C$ (d_2), and $C - N$ (d_3), and bond angles $N - C_\alpha - C$ (θ_1), $C_\alpha - C - N$ (θ_2), and $C - N - C_\alpha$ (θ_3); side chain information that may affect the backbone could also include $C_\alpha - C_\beta$ (r_1) and $N - C_\alpha - C_\beta$ (α_1) for example. When all of these quantities are specified exactly, the back-transformation from internal coordinates will also result in a perfect 3D Cartesian reconstruction of the protein backbone structure, using algorithms

[†] Reproduced with permission from: Li J, Zhang O, Lee S, Namini A, Liu ZH, Teixeira JM, Forman-Kay JD, Head-Gordon T. Learning Correlations between Internal Coordinates to Improve 3D Cartesian Coordinates for Proteins. *Journal of Chemical Theory and Computation.* 2023 Feb 7.

such as the natural extension reference frame (NeRF)[5]. However, in certain areas of protein modelling, such as fragment-based protein folding and loop modelling [6, 7], the Cartesian reconstruction is almost universally defined by only the backbone torsions while holding the bond lengths and angles fixed at mean values to decrease the complexity of the problem. Sometimes the variations on the ω torsion angles are also ignored and taken as fixed values of 0° or 180°, due to the planar nature of the peptide bond.[8] One might assume that a protein structure can be reconstructed in Cartesian coordinates quite well utilizing fixed bond lengths and angles since they typically have quite small variations around their means. However, even small deviations from the mean of the stiff degrees of freedom can strongly influence the Cartesian reconstruction. The origin of this error arises especially clearly from the nature of chain molecules: as the protein chain gets longer, small errors in bond lengths and bond angles can quickly accumulate and result in significant differences in the final back-transformed structure. According to a study by Holmes et al.[9] on globular proteins, the RMSD errors incurred in the internal coordinate back-transformations to C$_\alpha$ Cartesian positions under fixed bond lengths and angles is ∼6 Å for an average 150-amino-acid protein, and can be as high as 40 Å for larger proteins.

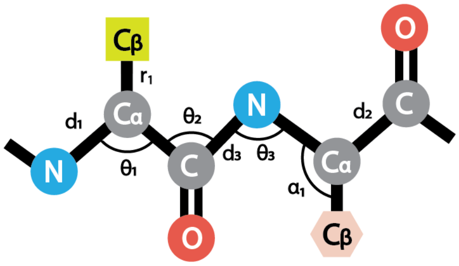


Figure 4.1: *Schematic of the polypeptide backbone and internal degrees of freedom.* Definition of the prediction targets: backbone bond angles $\theta_1 - \theta_3$, backbone bond lengths $d_1 - d_3$, $C_\alpha - C_\beta$ sidechain bond lengths r_1 and $N - C_\alpha - C_\beta$ sidechain bond angles α_1.

Alternatively, one could replace the assumption of fixed bonds and bond angles with a statistical approach that uses variable bond lengths and bond angles according to sequence or structural correlations in the PDB. Given the many types of correlations that exist between the internal coordinates of globular proteins, such as the ϕ and ψ torsion angles of the Ramanchandran plot[10], restraints on ω torsion angles as a function of ϕ and ψ[11], and the correlation between backbone and sidechain torsion angles used in the Dunbrack rotamer library[12], the correlations among the stiff bond lengths and angles with the flexible torsions should not be surprising. Earlier studies on the relationship between bond angles and ϕ, ψ torsion angles or amino acid types were mostly focused on the $N - C_\alpha - C$ bond angle, using both statistical methods and quantum mechanics calculation on model dipeptides.[13, 14, 15, 16]. The work by Berkholz[17] found that by using a static library for backbone bond angles dependent on backbone ϕ, ψ angles and residue types, the median RMSDs of protein reconstruction normalized to 100 amino acids is 2.85 Å. Following this pioneering study, more recent work by Roberto and co-workers have extended the correlation to all bond lengths and bond angles centered on backbone N, C_α and C atoms.[18, 19] Similarly, Lundgren et al. studied the correlation between protein backbone angles, secondary structure, and sidechain orientations.[20], and Ashraya et al. evaluated the steric-clash Ramanchandran maps conditioned on bond geometries. [21] However, none of these studies have considered the correlations in internal coordinates beyond local amino acid context and backbone geometries.

This work provides a more comprehensive machine learning approach that both quantifies and learns internal coordinate correlations within a deeper amino acid sequence context, that in turn provides a more accurate prediction of the 3D Cartesian coordinates relative to the errors incurred under the standard assumptions of fixed bond lengths and angles. By capturing the subtle correlations observed among internal coordinates, the Int2Cart (Internal to Cartesian) algorithm reduces the reconstruction RMSD error to ∼2.07 Å for test proteins normalized to 100-amino-acids, and an average RMSD of ∼3.74 Å over the entire test set for globular proteins as large as 599 amino acids. While many current protein modelling algorithms have adopted pairwise distance-based constraints[22, 23, 24] or directly output 3D coordinates, thereby bypassing the internal-to-Cartesian conversions[25, 26], the applicability of the Int2Cart algorithm is multi-fold.

First, our bond geometry prediction module is capable of providing more accurate references for internal coordinates, making our method a helpful tool for structural validation and refinement. We demonstrate this Int2Cart application by showing that the agreement between bond lengths and bond angles from AlphaFold2 (AF2) predicted structures [27] and Int2cart predictions is a strong indicator of AF2 model quality. Second, torsion-angle based approaches are still widely used in loop modelling [28] and in generating conformational ensembles of intrinsically disordered proteins (IDPs).[29] We find that Int2Cart is able to reproduce a structural ensemble of the disordered Sic-1 IDP with lower RMSD error when back-calculated to experimental observables, and generates fewer undesirable steric clashes. We also envision that Int2Cart should be applicable in the development of protein force fields that could benefit from more accurate valence models of backbone bond lengths and bond angles conditioned on other geometrical or sequence features [30].

4.2 Methods

Dataset preparation. We have adopted SidechainNet[31] as a preprocessed dataset that uses clustering techniques to extract protein sequences and structures with defined similarity cutoffs, to reduce bias in the original PDB structures, and to prevent information leakage from the training set to the test set relevant to assessing the machine learning generalization.[32, 31] The SidechainNet dataset represents each protein by its amino acid sequence, backbone and sidechain torsion angles (ϕ, ψ, ω, χ_1, χ_2, etc), backbone bond angles

θ_1-θ_3, as well as the all-atom 3D coordinates. For this study we ignore the sidechain torsions as we are only reconstructing backbones, and supplement the protein dataset with backbone bond lengths d_1-d_3 calculated from the 3D coordinates for training, validation and test sets. We also identified some θ_2 and θ_3 bond angles that were incorrect due to missing atoms in the next residue, and they were masked out along with the residue at the end of the protein chain. We used the latest available version of the SidechainNet dataset (CASP12) under 70% thinning and combined validation sets from 10% to 50% similarity cutoffs with the test set. This then defines the final test data for our algorithm while keeping track of the similarity for each individual test data point. When needed, we separated test set proteins at any broken chain positions and only retained chains longer than 50 consecutive amino acids.[32]

Our final training dataset contains 41,380 proteins with a minimum sequence length of 20 and maximum length of 4914 amino acids. Most structures in the training set have reported structural resolution < 4Å. The test set was comprised of 182 protein or protein fragments with sequence lengths between 23 and 599 amino acids. We have additionally compiled an IDP structural ensemble comprised of 1000 conformations for the N-terminal 92 residues of the Sic1 protein[33, 34] to validate the transferability of our model in a more challenging application scenario. We extracted the backbone torsions and rebuilt the Cartesian structures for each conformer under different assumptions about the bond lengths and bond angles as reported in Results. In addition, 20 randomly selected proteins from the human proteome were downloaded from the AlphaFold2 database [27] to illustrate the application of Int2Cart in validating protein structure models. The identification codes for these 20 proteins are provided in Appendix.

Neural network design. The structure of the deep neural network, Int2Cart, is depicted in Figure 4.2. The recurrent neural network architecture is chosed due to its capability to capture long-range correlations in internal coordinates, such as torsion angles that are exemplified in applications including protein folding[35] and IDP modelling[36]. We utilized 3 layers of stacked bidirectional gated recurrent units (GRU) as the central component, each of which contains a hidden state h_t with its information updated by the reset and update mechanisms for each element in the input sequence through the following set of equations[37]:

$$r_t = \sigma(W^r x_t + U^r h_{t-1} + b^r) \tag{4.1}$$

$$z_t = \sigma(W^z x_t + U^z h_{t-1} + b^z) \tag{4.2}$$

$$\tilde{h}_t = \tanh\left(W^n x_t + b^{nx} + r_t \odot (U^n h_{t-1} + b^{nh})\right) \tag{4.3}$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{4.4}$$

where $[W^r, W^z, W^n, U^r, U^z, U^n, b^r, b^z, b^{nx}, b^{nh}]$ are the trainable parameters of the model, x_t is the input to the cell at the current timestep, and r_t and z_t represent the reset and update gates, which are numbers between $(0, 1)$ that control how much information to retain in the new update vector $\tilde{h}_t$ and how the new hidden state vector h_t is composed from the update vector $\tilde{h}_t$ and the old hidden state h_{t-1}. σ denotes the sigmoid function, and $\odot$ represents element-wise multiplication. Dropout was applied to the hidden states between layers, so

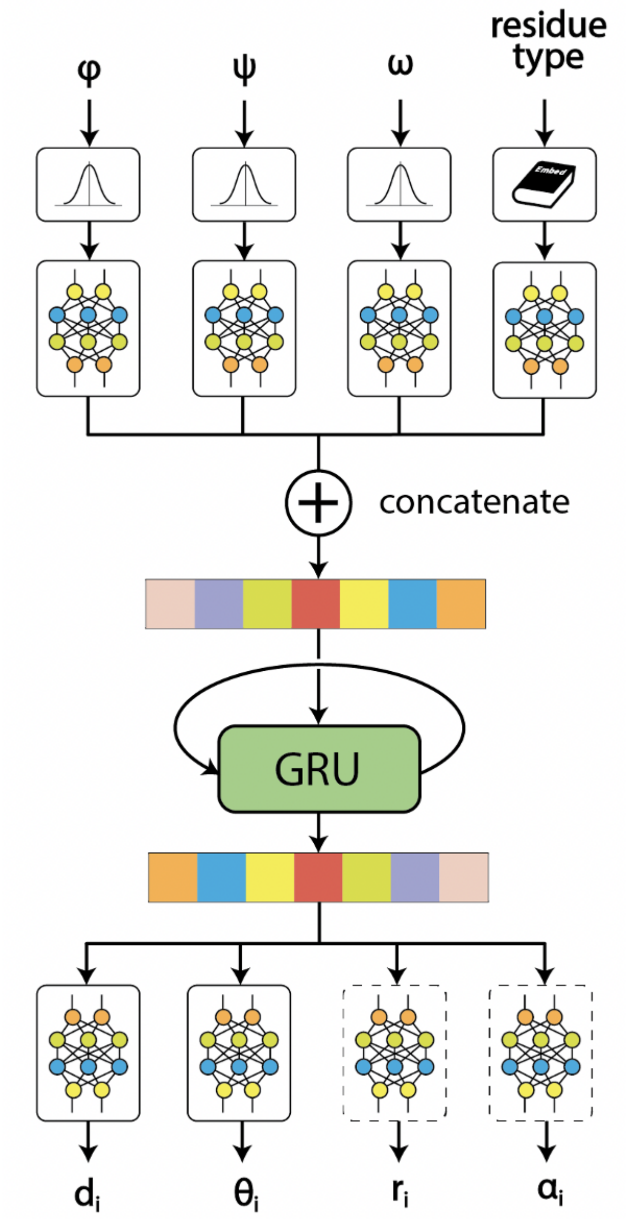


Figure 4.2: *Schematic of the Int2Cart neural network architecture.* The neural network is a gated recurrent unit (GRU) recurrent neural network. The inputs at each timestep are the concatenated latent vectors from Gaussian-smeared ϕ, ψ and ω torsion angles and embedded residue types; variations on the Int2Cart network can include the use of χ sidechain angles as well. The latent vector output from GRU are connected with multiple output networks to predict different targets.

that $x_t^{(l)} = h_t^{(l-1)} \odot \delta_t^{(l-1)}$, where each $\delta_t^{(l-1)}$ is a Bernoulli random variable that zeros out elements in the hidden state vector with a probability defined by the dropout rate.

The inputs into the first layer of GRU cells are the ϕ, ψ and ω torsion angles and the amino acid type. Since we are using a bidirectional recurrent neural network architecture, information about previous/following residues should already be included in the hidden state at any "timestep" in an implicit way in the GRU, which is sufficient information to allow the network to make predictions accurately enough without formulating it explicitly as the input. Each torsion angle, a, was represented by a Gaussian smearing function discretized to a vector of length 180 to account for uncertainty in the data, denoted x_{ia}

$$x_{ia} = \exp\left(-\frac{\text{diff}(\alpha_i, \hat{x}_a)}{2\sigma^2}\right) \tag{4.5}$$

where α_i is the actual ϕ, ψ or ω angle and $\hat{x}_a = (-180 + 2 * a)$ (both in degrees), and in this work we used σ =0.5°. The custom diff function

$$\text{diff}(\alpha_i, \hat{x}_a) = \min(|\alpha_i - \hat{x}_a|, \min(|\alpha_i - \hat{x}_a - 360|, |\alpha_i - \hat{x}_a + 360|)) \tag{4.6}$$

ensures that the periodicity of the angles is taken into account. Each smeared torsion angle vector is further transformed through two fully-connected layers with 90 and 64 units each and Rectified Linear Units (ReLU) activation[38] to generate latent representations of the torsion angles. The residue types are encoded by a trainable embedding dictionary and formulated into latent vectors of length 64; the hidden dimension size of 64 was chosen after a careful hyperparameter search and found to be the optimal value. The torsion angle latent vectors and the embedded residue types are then concatenated and transformed together through 2 fully-connected layers with 128 and 64 units and ReLU activation and constitute the inputs into the GRU cells.

The hidden state output from the last GRU layer is connected with multiple outputs to predict the backbone bond lengths and bond angles, or optionally sidechain bond lengths and bond angles as well. Each output is a fully-connected neural network with a hidden layer of size 100 and activation ReLU, and the output has size of 1 without any activation. The raw outputs are scaled by the standard deviation and translated by the mean value of that data type in the training dataset. The means and standard deviations we used are provided in Appendix Table 4.B.1.

Training details of the Int2Cart machine learning method. The neural network was trained by minimizing the weighted mean square error loss function

$$\mathcal{L} = \sum_i w_i(y_i - \hat{y}_i)^2 \tag{4.7}$$

where w_i controls the weighting for different data types in the loss function, y_i are the predictions from the model and $\hat{y}_i$ are the actual values from the data set. In practice we used the same weighting for all the data types. Missing data targets were masked out during the training. We used the Adam optimizer[39] with an initial learning rate of 0.001 and an exponentially decayed learning rate schedule, so $\text{lr}_i = \exp(-i * \alpha)$ where i is epoch number and $\alpha = 0.05$ in our case. The model was trained for a total of 100 epochs using a batch size of 128.

Building all-atom Cartesian structures from internal angle model predictions. With the full profile of backbone torsion angles and predictions of bond lengths and bond angles from the model, the 3D Cartesian structure of the protein containing all backbone atoms is reconstructed using the SidechainNet package.[31] It utilizes the natural extension reference frame (NeRF) algorithm[5] to sequentially calculate the position of the next atom with the positions of three previous atoms and the new bond length, bond angle and torsion angle. The all-atom backbone Cartesian structures for all the protein fragments in our test data set are built from either the Int2Cart algorithm vs. a standard baseline of using fixed bond lengths and bond angles (Fixed), or using bond lengths and bond angles from the Protein Geometry Database (PGD)[17] which uses bond geometries that depend on local torsions or amino acid type.

4.3 Results

Statistical analysis of the protein training set. Given the large collection of deposited protein structures in the PDB, we first consider a statistical analysis of protein bond lengths and bond angles when analyzed over the training set. Overall the distributions of these internal coordinate values are mostly Gaussian with relatively small standard deviations of ~ 0.01 Å for bond lengths and $\sim 2.6°$ for bond angles. Figure 4.3 (a-f) depicts the deviations from the mean bond length and angle values for a given (ϕ, ψ) combination, and confirms the existence of strong correlations among the internal coordinates averaged over the training data set. Specifically, the θ_1 angle is larger for the right-hand and left-hand helix regions in the Ramanchandran plot, while the beta-sheet regions have more narrow θ_1 angles, with deviations from the mean as large as 7.5°.

The θ_2 values are strongly correlated with the ψ torsion angle, with larger angles when ψ is between -100 and 0 degrees, and smaller angles than the mean otherwise. The θ_3 values for nearly all of the (ϕ, ψ) combinations are larger than average, but have smaller angles for helix regions. Similarly, the d_1 and d_2 bond lengths show greater correlations with the ϕ torsion angle, with a preference for larger values when ϕ is between -50 and +50 degrees, in which the bond lengths change by as much as 0.02 Å. Finally the correlation for the peptide d_3 bond with the backbone torsions is weak, consistent with its partial double bond character, except for a few hot spots where it can vary up to 0.04 Å. These correlations are statistically meaningful, because the standard deviations in each bin are smaller than the mean value differences (Appendix Figure 4.C.1), which means the statistical bias is more significant than the variance.

We have also considered the relationships between backbone ω torsion angles with bond angles (Figure 4.3 g-i) and bond lengths (Appendix Figure 4.C.2), and found interesting correlations between internal coordinates and ω torsion angles. The majority of peptide

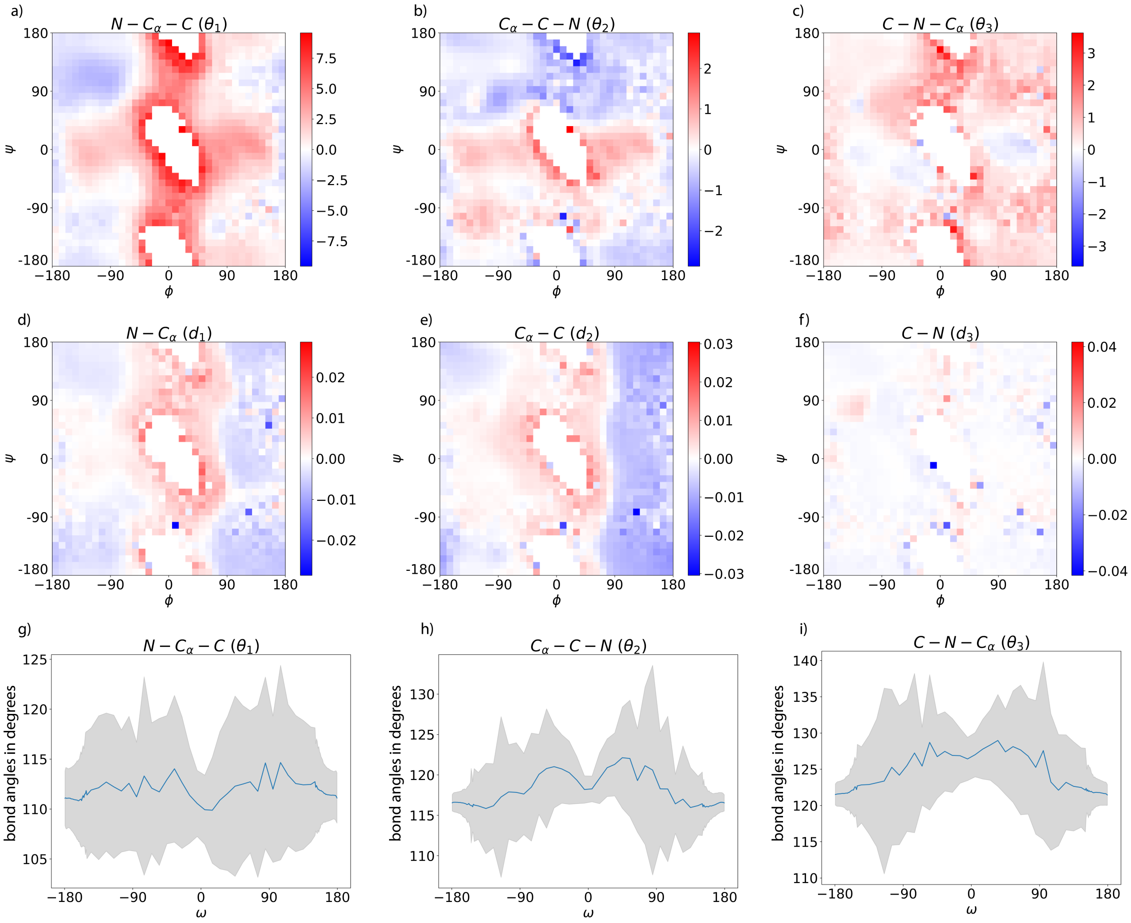


Figure 4.3: *Variations of bond angles and bond lengths as a function of* (ϕ, ψ), *or* ω *torsion angles.* a-f) Bond angle and bond length deviations from the mean values averaged over ϕ and ψ angles of the training set. The regions of red correspond to wider angles and longer bonds while the region in blue show reduced angle and bond values relative to the mean. The bond lengths and bond angles were categorized according to ϕ and ψ angles rounded to the closest tens, and the data are aggregated by calculating the means and standard deviations in each bin. The standard deviations are provided in Figure 4.C.1. g-i) Mean values and standard deviations of bond angles as a function of ω. The blue solid line represents mean values of bond angles at specific ω torsion angles, and the gray regions correspond to one standard deviation.

bonds in proteins are in the *trans-* conformation, with ω torsion angles close to 180°. However, *cis-* peptide bonds tend to be associated with smaller θ_1 angles and larger θ_2 and θ_3 angles. This result also makes structural sense since *cis-* peptide bonds incur more steric repulsion between sidechains of two consecutive residues, and larger θ_2, θ_3 and smaller θ_1 values allow the sidechains to be more separated. On the other hand, the correlation between bond lengths and ω torsion angles are not obvious (Appendix Figure 4.C.2). These correlations dependent on ω are also important for accurately predicting internal coordinates from backbone torsion angles as we will show later.

When we consider the observed distributions of all six internal coordinates as a function of the residue type (Appendix Figure 4.C.3), we find that the distributions are quite similar between amino acids with only subtle differences in the shape of the peaks, with the exception of glycine, which tends to have d_1 and d_2 values that are smaller, and θ_1 angles that are larger than other residues. Proline also defines an exception, with larger d_1 values due to the formation of the five-membered ring that requires longer bond lengths. However the bond length and angle distributions as a function of backbone torsions and residue type exhibit notable variations across *all* twenty amino acids as seen in Figure 4.4 for the θ_1 bond angle, as well as for the other backbone bonds and angles shown in Appendix Figures 4.C.4-4.C.8. To test whether structural resolution quality has an effect on the conclusion drawn from the statistical analysis, we further separated the training dataset by structure resolution categories of higher quality (resolution $\leq$ 2Å) and lower quality (2Å$<$ resolution $\leq$ 4Å) and compared the dependence of bond angles on ω torsion angles. The results are provided in Appendix Figure 4.C.9. No significant discrepancies exist on the correlations between two groups of structures with different qualities, which supports utilizing the whole training dataset without filtering based on resolution.

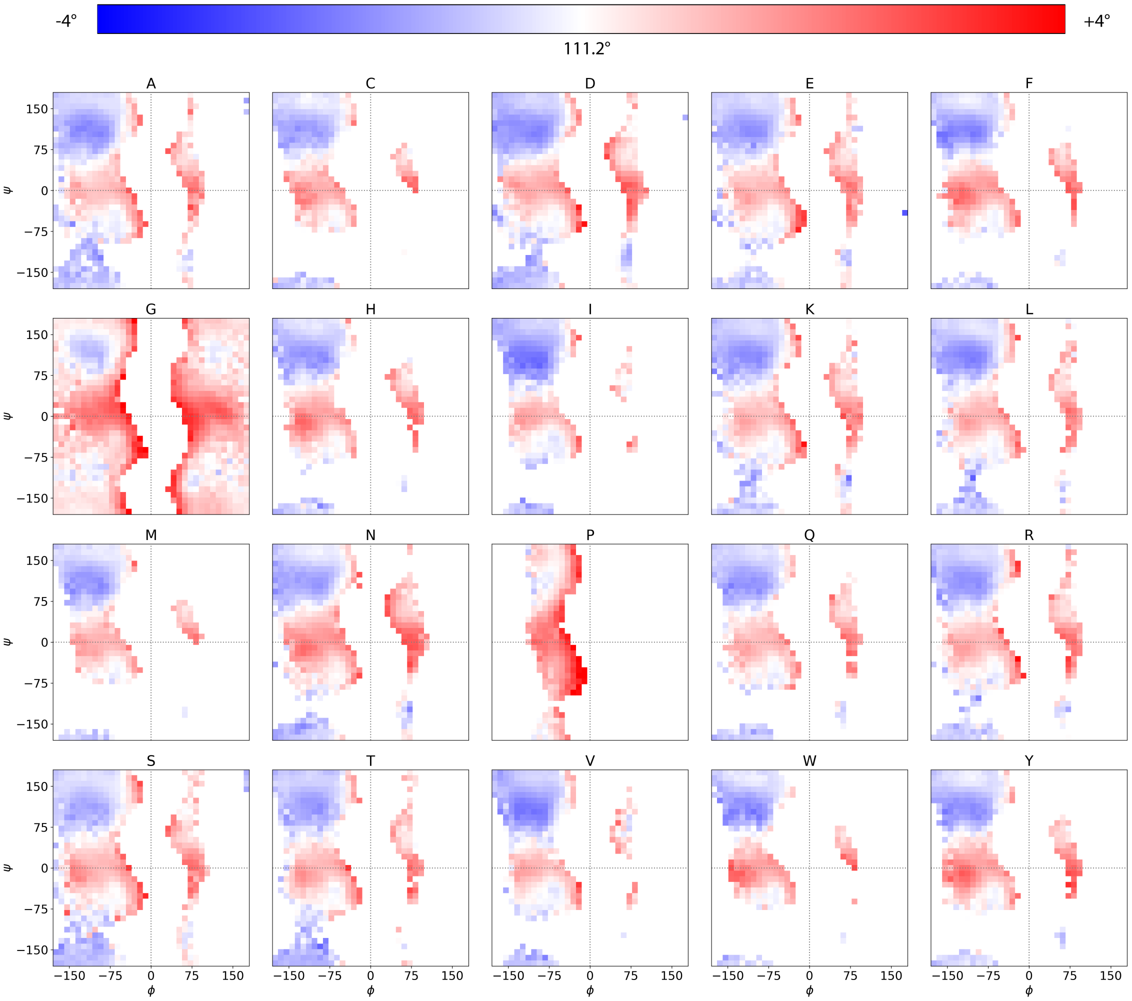


Figure 4.4: *$N - C_\alpha - C$ bond angle deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to large bond angles while the region in blue show reduced bond angles relative to the mean. The $N - C_\alpha - C$ bond angles were categorized according to ϕ and ψ angles rounded to the closest tens.

Machine learning of sequence and structural correlations. While the correlation graphs just described could serve as a source for bond lengths and angles when backbone torsion angles and residue types are provided, we are still missing the sequence-dependent correlations that are buried beneath the statistics of the single residue results. Therefore, we trained a deep neural network on the same data in order to capture the more subtle correlations among the internal coordinates conditioned on amino acid sequence. After training, the Int2Cart neural network was used to predict the test set which has low sequence and structural similarity with the training proteins. The root-mean-square error (RMSE) and Pearson correlation coefficients (R) on the test set are summarized in Appendix Table 4.B.2. We find that the RMSE in bond length predictions are within the variance determined from the data set, while predictions on the bond angles are more successful in terms of the RMSEs that are smaller than the dataset variance.

Table 4.1: *Quality of Cartesian reconstructed structures using Int2Cart, Fixed, and PGD methods normalized by sequence length, and Int2Cart results on different test data categories.* Accuracy is assessed in terms of the median and mean $C_\alpha(RMSD_{100})$, the root-mean-square error of the predicted C_α positions to the reference PDB structure normalized to 100 amino acids based on the test dataset. The second half of the table shows the breakdown of Int2Cart results in different similarity categories of data in the test dataset including CASP12 (which were after the time cutoff for proteins in the training dataset). All units in Å.

Method	Median	Mean±std
Fixed	3.22	3.47±1.83
PGD	2.92	3.32±1.87
Int2Cart	**2.07**	**2.38±1.36**
Test data category	Median	Mean±std
10% similarity	2.32	2.87±2.07
20% similarity	2.22	2.44±1.15
30% similarity	1.79	1.96±0.84
40% similarity	1.89	2.11±1.35
50% similarity	2.47	2.36±0.94
CASP12	2.06	2.39±1.22

Cartesian coordinate reconstructions. Given the three torsion angles $[\phi, \psi, \omega]$ for each residue over the entire protein sequence as input, we next consider how well the Cartesian coordinates are reconstructed based on whether bond and angle geometries are held fixed, using PGD, or learned from Int2Cart. Table 4.1 provides a general overview of the performance of the three approaches using a $C_\alpha(RMSD_{100})$ metric, which is the C_α RMSD values divided by the length of the protein and then multiplied by 100, as well as the C_α RMSD over all test set proteins regardless of length.

The reconstructed RMSDs for the Int2Cart structures are centered around lower median values of $C_\alpha(RMSD_{100})$ of 2.07 Å, and C_α RMSD of 3.74 Å over all test proteins. By contrast the Fixed model yields a median RMSD of 3.22 Å when all proteins are normalized to 100 amino acids, and the average over the entire test set is 5.39 Å. Table 4.1 also shows that the Int2Cart results are notably better than the PGD method which provides bond lengths and bond angles as a function of local ϕ, ψ and amino acid type, in which the median (2.92 Å) and mean (3.32 Å) $C_\alpha(RMSD_{100})$ are much higher than that found with Int2Cart. Furthermore, to investigate the tranferability of the Int2Cart model, the test dataset was broken down into subsets that have 10%-50% sequence similarity to any protein in the training dataset, and proteins from CASP12. The results reported in Table 4.1 indicate that the sequence similarity to the training dataset has little effect on the reconstruction quality of the proteins. Therefore, our model is expected to be generalizable to proteins it has not seen.

To provide a more statistical view of the predictions, Figure 4.5a reports the distribution of RMSDs for all backbone atoms with respect to the actual PDB structure for all proteins in the test set using Int2Cart and the Fixed method, as well as the pairwise RMSDs for the test proteins (Figure 4.5b), and the RMSD difference between the two methods as a function of sequence length (Figure 4.5c). It is evident that the vast majority of the test set proteins benefit from the machine learned bond lengths and bond angles, with an average improvement of $2-4$ Å RMSD over using Fixed bond lengths and bond angles. There is no obvious correlation between the RMSD improvements made by Int2Cart over Fixed with respect to sequence length, although the largest improvements occur in those proteins with longer amino acid sequences.

Figure 4.5d illustrates that proteins reconstructed by assuming fixed bond lengths and bond angles have lost significant secondary structure integrity compared to the reference structures, whereas the Int2Cart structures retain a much higher proportion of intact secondary structural elements. Beyond this anecdotal case, we performed a more extended analysis of Int2Cart and Fixed performance regarding the radius of gyration (R_g) and secondary structure recovery rate (SS-match) over the whole test dataset. Although we find that the Int2Cart Cartesian predictions have closer R_g values to the ground truth structures,

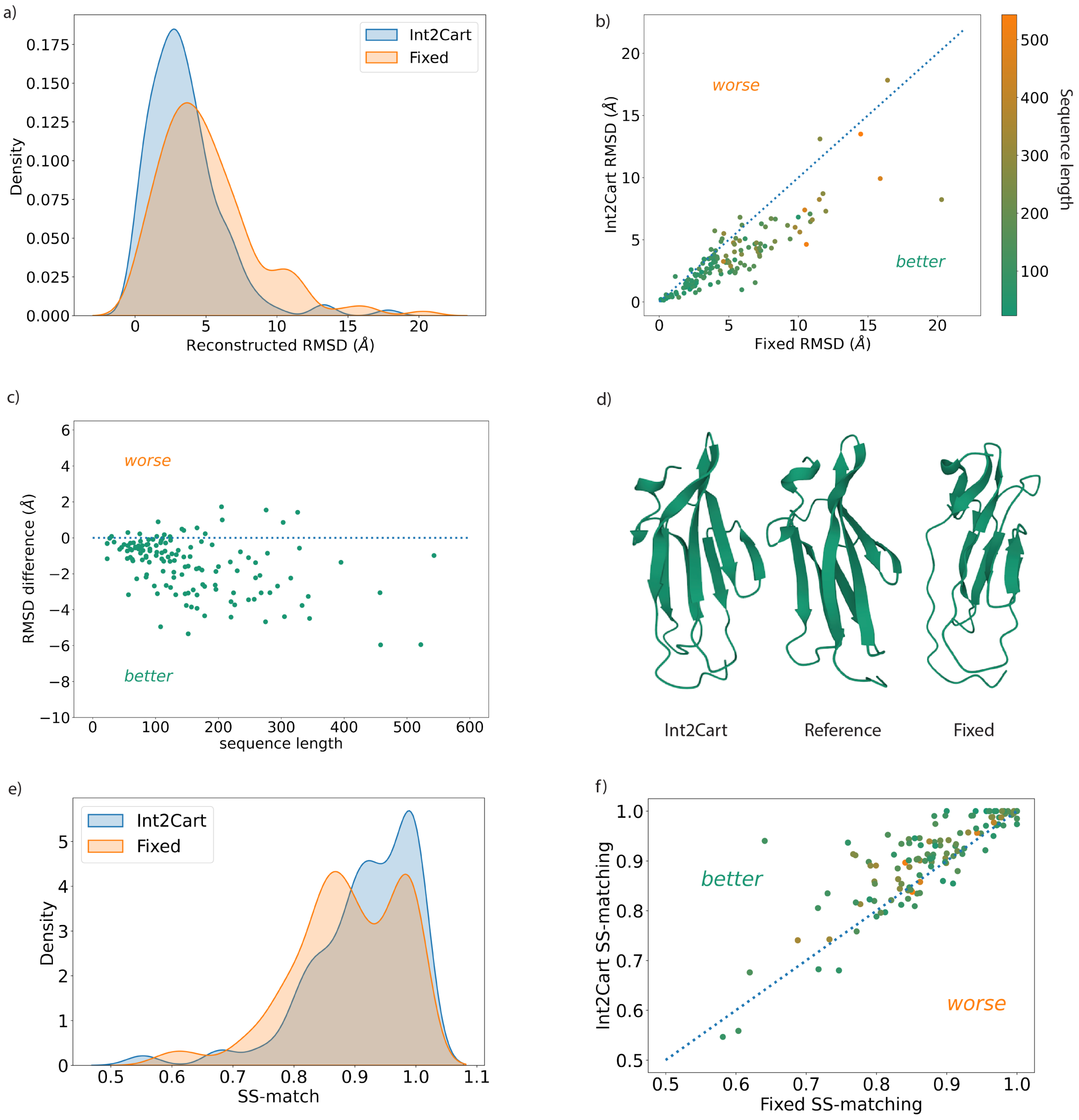


Figure 4.5: *Comparison of 3D Cartesian reconstructions of test set proteins using Int2Cart and compared to Fixed bonds and angles.* (a) Distribution of the RMSD in reconstructed Cartesian coordinates using Int2Cart and Fixed. (b) Comparison of Cartesian reconstruction error between Int2Cart and Fixed relative to the reference structure. (c) Improvement of Int2Cart over Fixed as a function of amino acid length. (d) An example of the backbone representation using Int2Cart and Fixed for the CASP12 TBM0872 protein[40], (e) The SS-match distribution and (f) comparison of SS-match for Int2Cart vs. Fixed across the test set.

the Fixed Cartesian structures still yields a comparably good result as seen in Appendix Figure 4.C.10. Figure 4.5e shows that Int2Cart systematically improves upon Fixed in regards the SS-match values, defined as the proportion of helix, strand, and coil DSSP assignments[41] for each residue that matches the reference structure. It is seen that Int2Cart has a higher proportion of test set proteins that have SS-match values larger than 0.8 (Figure 4.5f), which translates to more than 80% of the residues having correct secondary structure assignments.

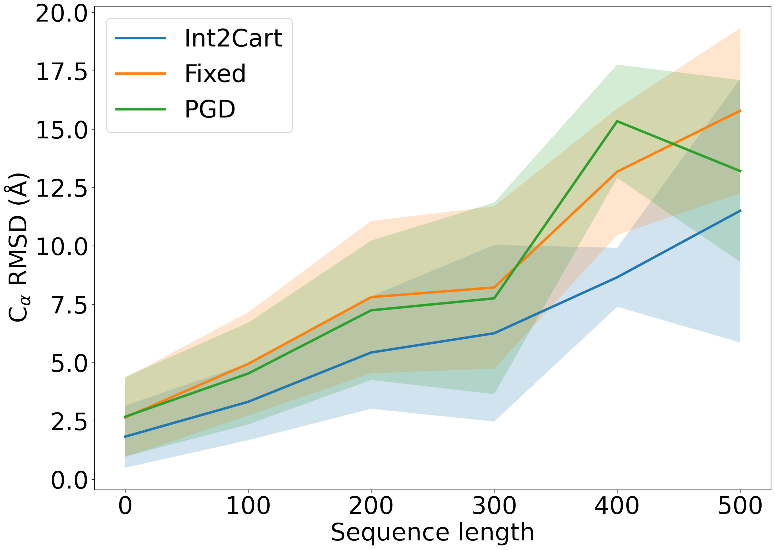


Figure 4.6: *Comparison of reconstructed structure C_α RMSD values in the test set as a function of sequence length using different sources of bond lengths and bond angles.* The C_α RMSDs were calculated against ground truth structures after using only their torsion angles for reconstruction. Shaded regions represent 1 standard deviation. The blue line represents Int2Cart, the orange line represents fixed bond lengths and angles, and the green line is the PGD method[17].

Comparison of sequence-length-dependent reconstruction quality among methods. Due to the sequential nature of the process of modelling protein 3D structures with internal coordinates, the reconstruction error is expected to increase as the protein sequence increases in length. In Figure 4.6 the reconstruction error evaluated as the RMSD on the C_α atoms compared to the initial structures from the PDB are plotted as a function of sequence length, in which proteins were reconstructed using either Int2Cart-predicted bond geometries, using fixed bond lengths and bond angles, or using the local-conformation dependent Protein Geometry Database (PGD) as described in Ref [17]. Test proteins are grouped by sequence lengths with increments of 100 amino acids, and the standard deviations in each group are described by the shaded regions in Figure 4.6. Compared to using fixed bond lengths and bond angles, the PGD method has slight improvements in almost all sequence length ranges except around 400 amino acids. Even so, the Int2Cart has a more significant improvement in C_α RMSD compared to PGD, suggesting its superiority is likely due to the fact that Int2Cart is able to learn deeper sequence correlations.

Ablation studies. To understand the importance of various inputs for prediction accuracy of Int2Cart and how accuracy affects reconstructing the Cartesian structures, we performed an ablation study by training separate deep learning models using subsets of the inputs, and reconstructing structures using only predicted bond lengths, only predicted bond angles, or using both. We have also trained models with additional inputs of χ_1 torsion angles, along with r_1 and α_1 sidechain bond lengths and bond angles as additional outputs, to evaluate how including sidechain information could improve prediction and reconstruction of backbone structures. All ablation trials are reported in Table 4.2.

We see that the differences in predictions of the backbone bond lengths from different deep learning models are not significant, but prediction accuracy for backbone bond angles RMSE and reconstructed Cartesian structure RMSD are highly dependent on what information is available to the model. Specifically, a machine learning model that only knows about the residue types performs the worst with >5 Å in the reconstruction RMSD. Unsurprisingly based on statistical analysis of the PDB, backbone ϕ and ψ torsion angles provide more information than residue types alone, and allows the reconstruction RMSD to decrease to 4.56 Å on average. Including both ϕ, ψ and residue types further decreases the average reconstruction error to 4.29 Å. As expected from the correlation of bond lengths and bond angles with ω torsion angles as well, including exact values for ω torsion angles also significantly improves the model and allows the reconstructed structure RMSD to decrease further to 3.77 Å across the whole test set, and to 2.38 Å for proteins normalized to 100 amino acids. When we tested the inclusion of sidechain χ_1 torsion angles, we find that the 3D reconstruction model is even better, achieving an average reconstruction structure RMSD of 3.30 Å regardless of protein length. This is probably due to the fact that χ_1 torsion angles are indicative of avoidable steric clashes between protein backbones and side chains to create more accurate descriptions of subsequent backbone bond geometries, even though side chain atoms are not explicitly treated during structure reconstruction in this work.

Table 4.2: *Ablation studies of internal coordinate inputs and Cartesian coordinate reconstructions.* Upper table: predicted bond lengths and bond angles RMSEs of Int2Cart taking different internal coordinate inputs, and corresponding RMSD of the reconstructed Cartesian structure. Each ablation of the input is repeated 3 times with different initializations of the machine learning model to obtain statistically meaningful results. Standard deviations reflect fluctuations of mean values among 3 parallel experiments. Lower table: cartesian structure reconstruction RMSD using different Int2Cart predicted and Fixed combinations of bond lengths and angles and binary ω torsion angles. Standard deviations reflect range of reconstructed structure RMSDs among different proteins.

Training Model inputs	**<d> RMSE(Å)**	**<θ> RMSE (°)**	**Reconstructed RMSD**(Å)
Residue type	0.010±1E-5	1.84±0.0008	5.21± 0.04
$\phi + \psi$	0.010±1E-4	1.69±0.02	4.56±0.07
$\phi + \psi$ + Residue type	0.010±5E-5	1.63±0.001	4.29±0.02
$\phi + \psi + \omega$ + Residue type	0.010±1E-4	1.50±0.006	3.77±0.03
$\phi + \psi + \omega + \chi_1$ + Residue type	0.009±1E-4	1.37±0.004	3.30±0.03

Source of bond geometries	**Reconstructed RMSD (Å)**
Predicted bond lengths and bond angles	3.74±2.94
Fixed bond lengths and predicted bond angles	3.74±2.94
Predicted bond lengths and fixed bond angles	5.38±3.70
Fixed bond lengths and angles	5.39±3.71
Fixed bond lengths, bond angles and using 0°/180° ω angles	9.52±6.49

To bolster these conclusions, Table 4.2 shows that the reconstruction quality does not depend on the direct prediction of bond lengths, as it essentially has no effect on the reconstructed structures, which may have been anticipated from the fact that bond length errors are on par with the variance. But this final ablation study provides direct evidence that accurate predictions of bond angles are of primary importance for the quality of the reconstructed Cartesian structures. In addition, using accurate ω torsion angles in the reconstruction is of great importance, since treating ω as binary greatly deteriorates the quality of reconstructed structures.

Using Int2Cart internal coordinate agreements to validate AlphaFold2 structures. AlphaFold2 (AF2) has been a huge success in predicting atomic structures of proteins with astonishing accuracy.[26] Nevertheless its predictions have variety of quality, which is also reflected in its internal confidence estimations for each residue called the predicted local distance difference test (pLDDT) score, with values greater than 90 indicating high confidence, and values below 50 indicating low confidence. To investigate the relationship between AF2 model quality and how much the bond lengths and bond angles in these AF2 models agree with the same Int2Cart quantities, we randomly collected 20 AF2 predicted protein structures from the human proteome, and calculated the bond lengths and bond angles using Int2Cart and the AF2 torsion angles. The results are summarized in Figure 4.7 and Appendix Figures 4.C.11-4.C.13.

On a per-residue basis, we observe strong correlation between the agreement of AF2 and Int2Cart bond geometries, and AF2 prediction confidence, as we illustrate in Figure 4.7(a). We see that the most confident residue predictions in AF2 models have better correlation in θ_1 values between AF2 models and Int2Cart predictions, compared to the residues with lower confidence. Figure 4.7(b) further discretizes the absolute differences into bins of 1° increments and shows that the residues that have a larger discrepancy between bond geometries in AF2 structures and Int2Cart predictions have on average lower quality in terms of pLDDT scores. Similar plots are generated for the θ_2, θ_3 d_1, d_2 and d_3 data where the Int2Cart and AF2 agreement is less good, but still exhibit strong correlations between geometry differences and pLDDT values (Appendix Figures 4.C.11 and 4.C.12).

Finally, we aggregrate all three bond angle results into correlations and mean absolute differences over the entirety of all 20 AF2 protein models we have tested, and compared with their average structure confidence score. Figure 4.7(c-d) indicate that the agreement between Int2Cart predicted bond angle geometries and the AF2 model strongly correlates with overall model quality, thus supporting using Int2Cart for structure validations. Similar conclusions are reached for the bond lengths as given in Appendix Figure 4.C.13.

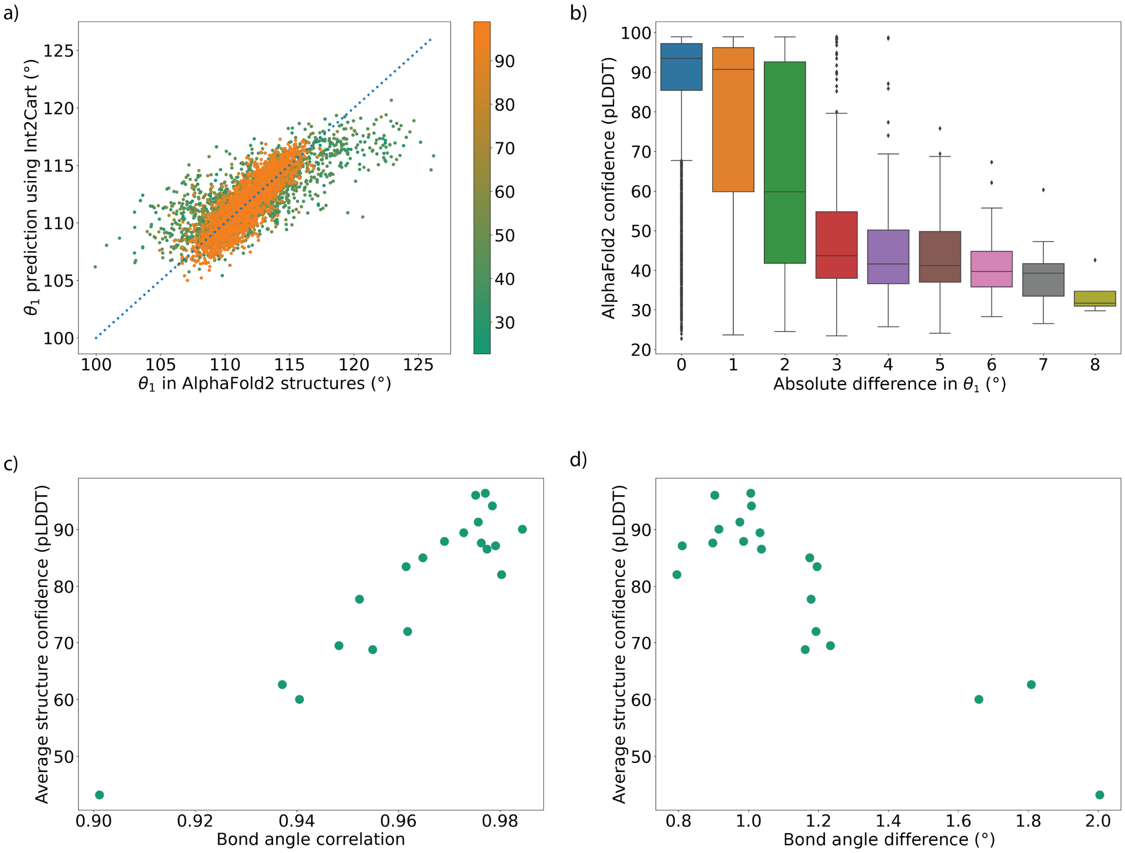


Figure 4.7: *Correlation between AlphaFold2 (AF2) structure quality and the agreement between bond geometries from the AF2 predicted structures and Int2Cart predicted values using torsion angles from AF2 structures* (a) Correlation between θ_1s ($N - C_\alpha - C$ bond angles) from AF2 structures and Int2Cart predictions colored by AF2 pLDDT scores of the relevant residues. (b) Box plot showing distribution of AF2 pLDDT scores of individual residues based on absolute difference in θ_1 between AF2 structures and Int2Cart predictions. The boxes represent the quartiles of the distribution and the whiskers represent the rest of the distribution. Individual data points are outliers identified from the inter-quartile range. (c) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond angle correlations between AF2 and Int2Cart in an AF2 predicted protein structure (d) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond angle absolute difference between AF2 and Int2Cart in an AF2 predicted protein structure.

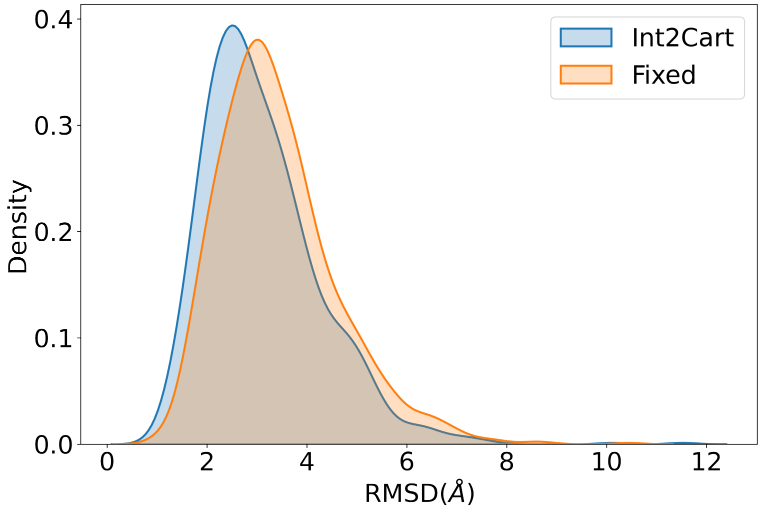


Figure 4.8: *Comparison of distribution of reconstruction RMSD for individual conformaions in the Sic1 IDP ensemble.* Structures reconstructed with Int2Cart method on average has lower RMSD to their original structures compared with using fixed bond lengths and bond angles.

Using Int2Cart to rebuild an IDP ensemble. Finally we consider a test case that is quite different from the originally defined test set from SidechainNet, in which we show that our Int2Cart method can improve upon the Cartesian reconstruction of an ensemble of structures of a disordered protein compared to Fixed bond lengths and angles. Figure 4.8 compares the Cartesian reconstruction RMSD distributions for Int2Cart and Fixed for the Sic1 IDP ensemble, in which we find that the Int2Cart method is overall closer to the original ensemble, with a 3.1 Å average RMSD compared to the Fixed method that has a mean RMSD of 3.4 Å. We have also checked the number of steric clashes in the structures generated from these two methods. A steric clash is defined as two atoms in the structure that are closer to 0.6 times the sum of the van der Waals radii of the two atoms.[42] Out of the 1000 conformations, 73 structures generated from Int2Cart contained steric clashes, which means 92.7% of the structures are clash-free. By comparison, 102 structures generated using fixed bond lengths and bond angles contained steric clashes, which translates to 89.8% of clash-free structures. A higher proportion of clash-free structures is meaningful because typically structures containing clashes are discarded, and a method with higher proportion of clash-free structures wastes less computational resources, and supports the application of the Int2Cart algorithm to the modelling of disordered protein ensembles.

4.4 Discussion and conclusion

In this work we have developed a new machine learning approach to the generic representation problem of internal coordinates (bond lengths, valence angles, and dihedral angles) and how to increase the fidelity of the back-transformation to 3D Cartesian coordinates. The Int2Cart algorithm utilizes a gated recurrent unit neural network to predict real-valued backbone bond lengths and bond angles for each residue of a complete protein sequence given its torsion angle profile. In summary, Int2Cart can reconstruct the Cartesian structure of proteins with RMSDs that are significant improvements over the fixed backbone bond lengths and bond angles that are the standard practice in a large variety of protein modelling approaches, or some recent approaches such as the Protein Geometry Database. The success of our algorithm across IDP ensembles further validates that the Int2Cart algorithm is transferable among different types of proteins, and can consistently improve the quality of Cartesian structure reconstruction. We have also exposed the potential of Int2Cart in validating structure quality by showing the agreement on bond geometries between Int2Cart predictions and values in an AlphaFold2 model has strong correlation with the AlphaFold2 pLDDT confidence metric. Possibilities in refining AF2 structures using Int2Cart will be investigated in the future.

In its current form the Int2Cart algorithm only generates backbone structures for the target proteins, although we can improve Cartesian reconstruction performance with the inclusion of the χ_1 torsion and predicting r_1 and α_1. Theoretical approaches such as the Monte Carlo Side Chain Ensemble (MC-SCE) method can utilize the backbone from Int2Cart to calculate side chain ensembles in order to complete the full structure.[42] It is also clear that there is still room for improvement in the Cartesian reconstruction of larger proteins, and the inherent scaling of error with respect to sequence length is inevitable for a deep learning model that predicts internal coordinates in a sequential manner (i.e., a GRU model). Therefore, it may be possible to improve the quality of Cartesian structure reconstruction with a distance-based neural network model, i.e., by representing the 3D coordinates of the structure directly.

Nevertheless, the model in its current form already provides a useful computational tool to greatly improve the quality of protein structures reconstructed from backbone torsion angles alone, whether globular folded proteins or disordered protein ensembles. We envision Int2Cart should see broad use in structure refinement and validation[43, 44] and development of protein force fields that could benefit from more accurate valence models of backbone bond lengths and bond angles conditioned on other geometrical or sequence features.[45] Finally, the Int2Cart GRU neural network model could also be useful for other chain molecules, only requiring retraining with new data if available for systems such as nucleic acids and lipids.

4.5 Author contributions

J.L. and T.H.-G. designed the project. J.L. designed and wrote the Int2Cart software. O.Z. also provided input on the neural network design and tuning of the model and valuable critiques including testing. J.L. and T.H.-G. wrote the paper and all authors provided valuable input and discussion.

4.6 Acknowledgements

T.H.-G. and J.D.F.-K. acknowledge funding from the National Institute of Health under Grant 5R01GM127627-04. J.D.F.-K. also acknowledges support from the Natural Sciences and Engineering Research Council of Canada (2016-06718) and from the Canada Research Chairs Program.

4.7 References

[1] Jon Baker, Don Kinghorn, and Peter Pulay. Geometry optimization in delocalized internal coordinates: An efficient quadratically scaling algorithm for large molecules. *J. Chem. Phys.*, 110(11):4986–4991, 1999.

[2] Charles D. Schwieters and G.Marius Clore. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.*, 152(2):288–302, 2001.

[3] Stewart A. Adcock and J. Andrew McCammon. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106(5):1589–1615, 2006. PMID: 16683746.

[4] wwPDB consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.*, 47(D1):D520–D528, 10 2018.

[5] Jerod Parsons, J Bradley Holmes, J Maurice Rojas, Jerry Tsai, and Charlie EM Strauss. Practical conversion from torsion space to cartesian space for in silico protein synthesis. *J. Comput. Chem.*, 26(10):1063–1068, 2005.

[6] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier, 2004.

[7] Rhiju Das and David Baker. Macromolecular modeling with rosetta. *Annu. Rev. Biochem.*, 77:363–382, 2008.

[8] Pierrick Craveur, Agnel Praveen Joseph, Pierre Poulain, Alexandre G de Brevern, and Joseph Rebehmed. Cis–trans isomerization of omega dihedrals in proteins. *Amino acids*, 45(2):279–289, 2013.

[9] J Bradley Holmes and Jerry Tsai. Some fundamental aspects of building protein structures from fragment libraries. *Protein Sci.*, 13(6):1636–1650, 2004.

[10] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, 7(1):95–99, 1963.

[11] Donald S Berkholz, Camden M Driggers, Maxim V Shapovalov, Roland L Dunbrack Jr, and P Andrew Karplus. Nonplanar peptide bonds in proteins are common and conserved but not biased toward active sites. *Proc. Natl. Acad. Sci. USA*, 109(2):449–453, 2012.

[12] Maxim V Shapovalov and Roland L Dunbrack Jr. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure*, 19(6):844–858, 2011.

[13] P Andrew Karplus. Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci.*, 5(7):1406–1420, 1996.

[14] Xiaoqin Jiang, Ming Cao, Brian Teppen, Susan Q Newton, and Lothar Schaefer. Predictions of protein backbone structural parameters from first principles: Systematic comparisons of calculated nc (. alpha.)-c'angles with high-resolution protein crystallographic results. *J. Phys. Chem.*, 99(26):10521–10525, 1995.

[15] Lothar Scháfer, Ming Cao, and Mary Jane Meadows. Predictions of protein backbone bond distances and angles from first principles. *Biopolymers: Original Research on Biomolecules*, 35(6):603–606, 1995.

[16] Ching-Hsing Yu, Mya A Norman, Lothar Schäfer, Michael Ramek, Anik Peeters, and Christian Van Alsenoy. Ab initio conformational analysis of n-formyl l-alanine amide including electron correlation. *J. Mol. Struct.*, 567:361–374, 2001.

[17] Donald S Berkholz, Maxim V Shapovalov, Roland L Dunbrack Jr, and P Andrew Karplus. Conformation dependence of backbone geometry in proteins. *Structure*, 17(10):1316–1325, 2009.

[18] Roberto Improta, Luigi Vitagliano, and Luciana Esposito. The determinants of bond angle variability in protein/peptide backbones: A comprehensive statistical/quantum mechanics analysis. *Proteins: Structure, Function, and Bioinformatics*, 83(11):1973–1986, 2015.

[19] Roberto Improta, Luigi Vitagliano, and Luciana Esposito. Bond distances in polypeptide backbones depend on the local conformation. *Acta Crystallogr. D Biol. Crystallogr.*, 71(6):1272–1283, 2015.

[20] Martin Lundgren and Antti J Niemi. Correlation between protein secondary structure, backbone bond angles, and side-chain orientations. *Phys. Rev. E.*, 86(2):021904, 2012.

[21] Ashraya Ravikumar, Chandrasekharan Ramakrishnan, and Narayanaswamy Srinivasan. Stereochemical assessment of (φ, ψ) outliers in protein structures using bond geometry-specific ramachandran steric-maps. *Structure*, 27(12):1875–1884, 2019.

[22] Wei Zheng, Yang Li, Chengxin Zhang, Robin Pearce, SM Mortuza, and Yang Zhang. Deep-learning contact-map guided protein structure prediction in casp13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1149–1164, 2019.

[23] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demisothers Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[24] Zongyang Du, Hong Su, Wenkai Wang, Lisha Ye, Hong Wei, Zhenling Peng, Ivan Anishchenko, David Baker, and Jianyi Yang. The trrosetta server for fast and accurate protein structure prediction. *Nat. Protoc.*, 16(12):5634–5651, 2021.

[25] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[26] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[27] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, 50(D1):D439–D444, 2022.

[28] Julian Lee, Dongseon Lee, Hahnbeom Park, Evangelos A Coutsias, and Chaok Seok. Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins: Structure, Function, and Bioinformatics*, 78(16):3428–3436, 2010.

[29] João M. C. Teixeira, Zi Hao Liu, Ashley Namini, Jie Li, Robert M. Vernon, Mickaël Krzeminski, Alaa A. Shamandy, Oufan Zhang, Mojtaba Haghighatlari, Lei Yu, Teresa Head-Gordon, and Julie D. Forman-Kay. Idpconformergenerator: A flexible software

suite for sampling the conformational space of disordered protein states. *J. Phys. Chem. A.*, 126(35):5985–6003, 2022. PMID: 36030416.

[30] Wouter G. Touw and Gert Vriend. On the complexity of Engh and Huber refinement restraints: the angle τ as example. *Acta Crystallogr. D.*, 66(12):1341–1350, Dec 2010.

[31] J. E. King and D. R. Koes. Sidechainnet: An all-atom protein structure dataset for machine learning. *Proteins*, 89(11):1489–1496, 2021.

[32] Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC Bioinform.*, 20(1):311, 2019.

[33] Tanja Mittag, Stephen Orlicky, Wing-Yiu Choy, Xiaojing Tang, Hong Lin, Frank Sicheri, Lewis E. Kay, Mike Tyers, and Julie D. Forman-Kay. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. USA*, 105(46):17772–17777, 2008.

[34] Gregory-Neal W. Gomes, Mickaël Krzeminski, Ashley Namini, Erik W. Martin, Tanja Mittag, Teresa Head-Gordon, Julie D. Forman-Kay, and Claudiu C. Gradinaru. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. *J. Am. Chem. Soc.*, 142(37):15697–15710, 2020. PMID: 32840111.

[35] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Syst.*, 8(4):292–301, 2019.

[36] Oufan Zhang, Mojtaba Haghighatlari, Jie Li, Joao Miguel Correia Teixeira, Ashley Namini, Zi-Hao Liu, Julie D Forman-Kay, and Teresa Head-Gordon. Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *arXiv preprint arXiv:2206.12667*, 2022.

[37] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[38] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323. JMLR Workshop and Conference Proceedings, 2011.

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[40] K. Tan, M. Gu, R. Jedrzejczak, and A. Joachimiak. The crystal structure of the n-terminal domain of a novel cellulases from bacteroides coprocola. *PDB*, 2016.

[41] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[42] Asmit Bhowmick and Teresa Head-Gordon. A monte carlo method for generating side chain structural ensembles. *Structure*, 23(1):44–55, 2015.

[43] KS Wilson, S Butterworth, Z Dauter, VS Lamzin, M Walsh, S Wodak, J Pontius, J Richelle, A Vaguine, C Sander, et al. Who checks the checkers? four validation tools applied to eight atomic resolution structures. *J. Mol. Biol.*, 276(2):417, 1998.

[44] Gerard J Kleywegt. On vital aid: the why, what and how of validation. *Acta Crystallogr. D Biol. Crystallogr.*, 65(2):134–139, 2009.

[45] Patrick Conway, Michael D Tyka, Frank DiMaio, David E Konerding, and David Baker. Relaxation of backbone bond geometry improves protein energy landscape modeling. *Protein Sci.*, 23(1):47–55, 2014.

Appendix

4.A Identifiers in the AlphaFold Protein Structure Database for all AlphaFold2 proteins used in the analysis

Q3LI64, Q5VTL8, Q15287, Q86TL2, O43316, Q8NBN3, Q8IVB5, L0R819, Q9Y216, Q8NH92, Q9Y259, A6NM03, Q8WXK1, P62070, P12104, P21145, Q9NW38, O60547, Q9NR45, P13716

4.B Supplementary Tables

Table 4.B.1: Mean and standard deviations used to rescale model predictions for bond lengths (Å) and bond angles (**rad**)

Data type	Mean	Standard deviation
N-C$_\alpha$ bond length	1.460	0.0118
C$_\alpha$-C bond length	1.525	0.0123
C-N bond length	1.331	0.0095
N-C$_\alpha$-C bond angle	1.941	0.0472
C$_\alpha$-C-N bond angle	2.034	0.0413
C-N-C$_\alpha$ bond angle	2.122	0.0480

Table 4.B.2: *Int2Cart prediction accuracy on backbone bond lengths and bond angles.* Accuracy is assessed in terms of root-mean-square error (RMSE) and Pearson correlation coefficients (R).

Data type	RMSE	R
N-C$_\alpha$-C (θ_1)	1.87°	0.71
C$_\alpha$-C-N (θ_2)	1.06°	0.49
C-N-C$_\alpha$ (θ_3)	1.46°	0.50
N-C$_\alpha$ (d_1)	0.010 Å	0.38
C$_\alpha$-C (d_2)	0.011 Å	0.40
C-N (d_3)	0.008 Å	0.45

4.C Supplementary Figures

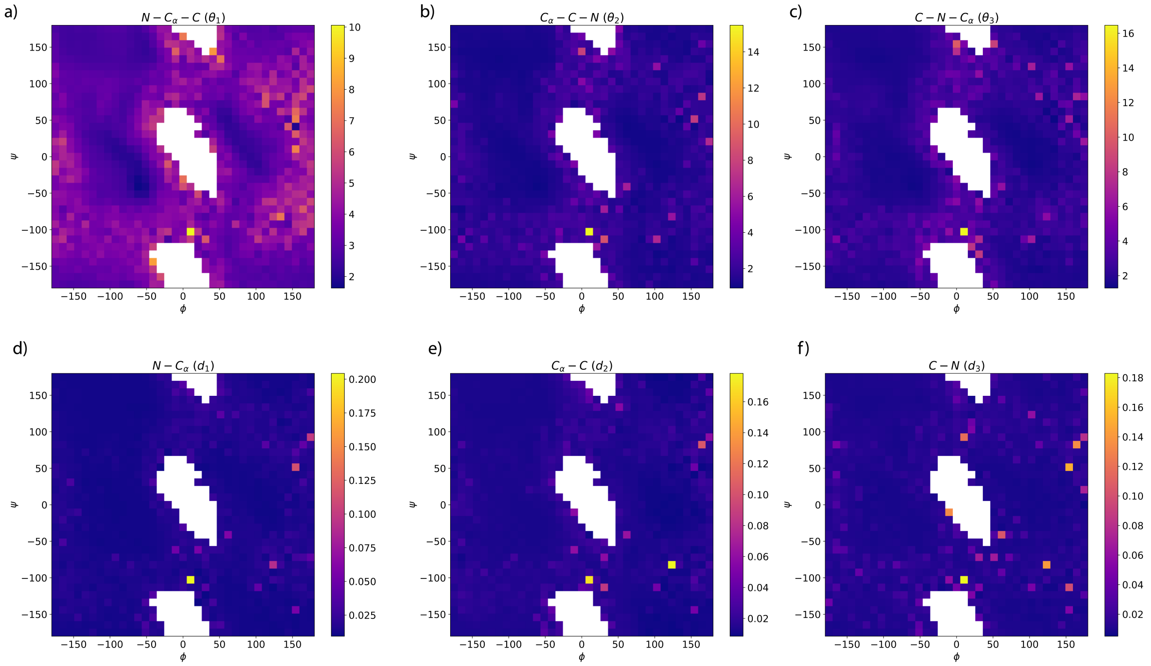


Figure 4.C.1: *Variations in the standard deviation (STD) of bond angle and bond lengths as a function of* ϕ,ψ. The regions of red correspond to larger STD while the region in blue have much smaller STD.

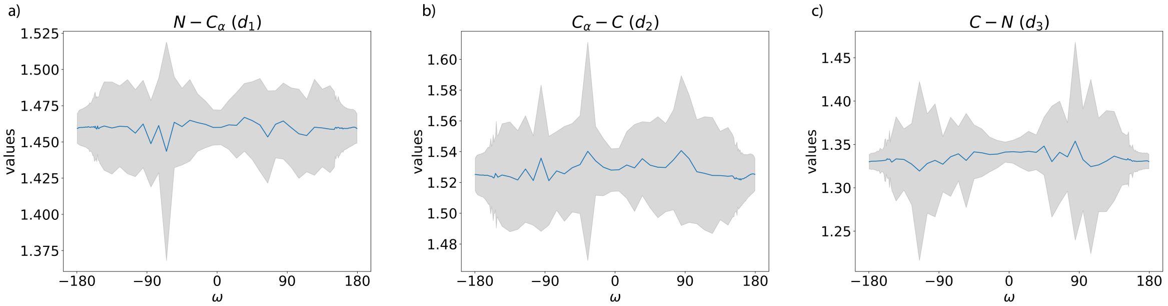


Figure 4.C.2: *Mean values and standard deviations of bond lengths as a function of* ω*.* The blue solid lines represent mean values of bond lengths at specific ω torsion angles, and the gray regions correspond to one standard deviation.

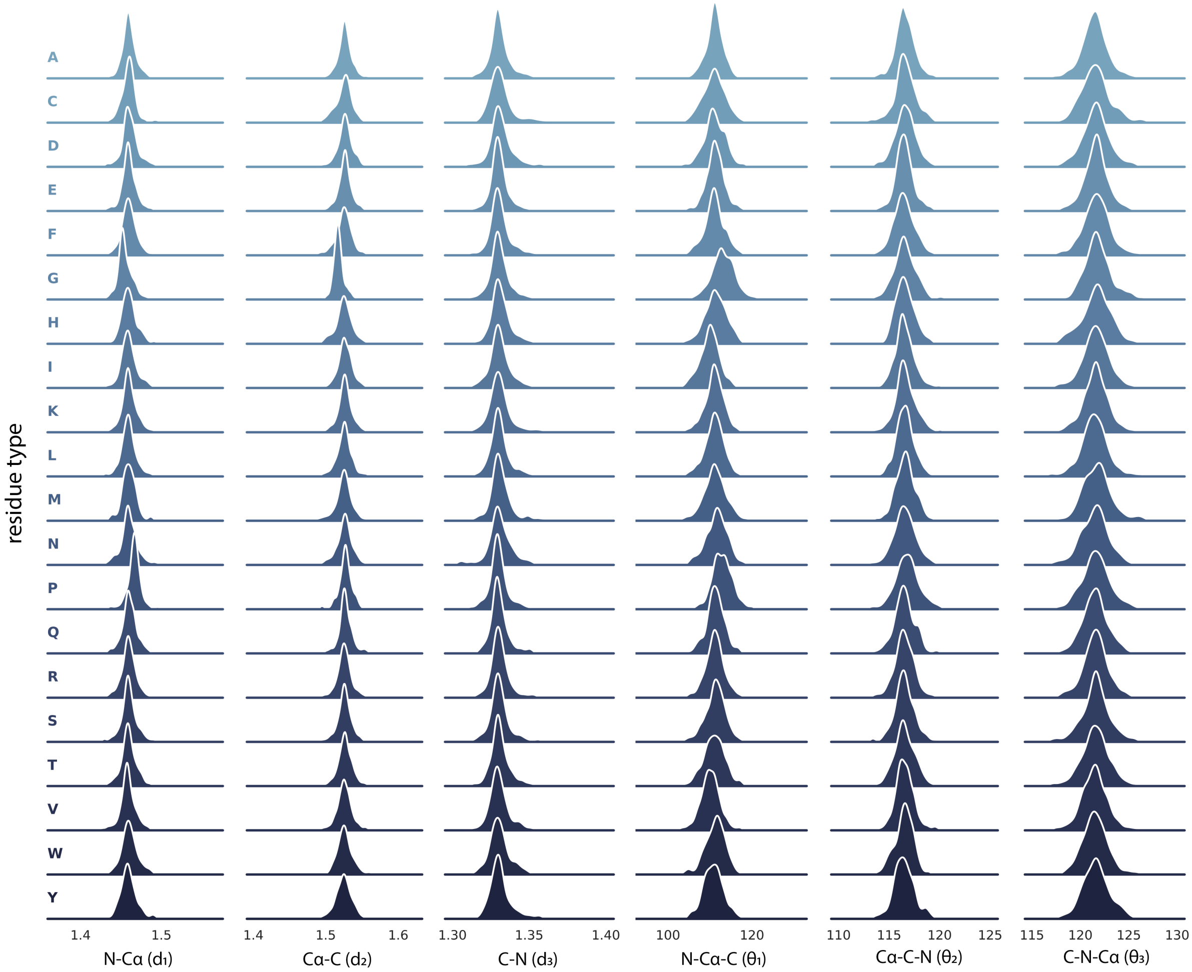


Figure 4.C.3: *Distributions of bond lengths and bond angles as a function of residue type.* Shown for all twenty amino acids.

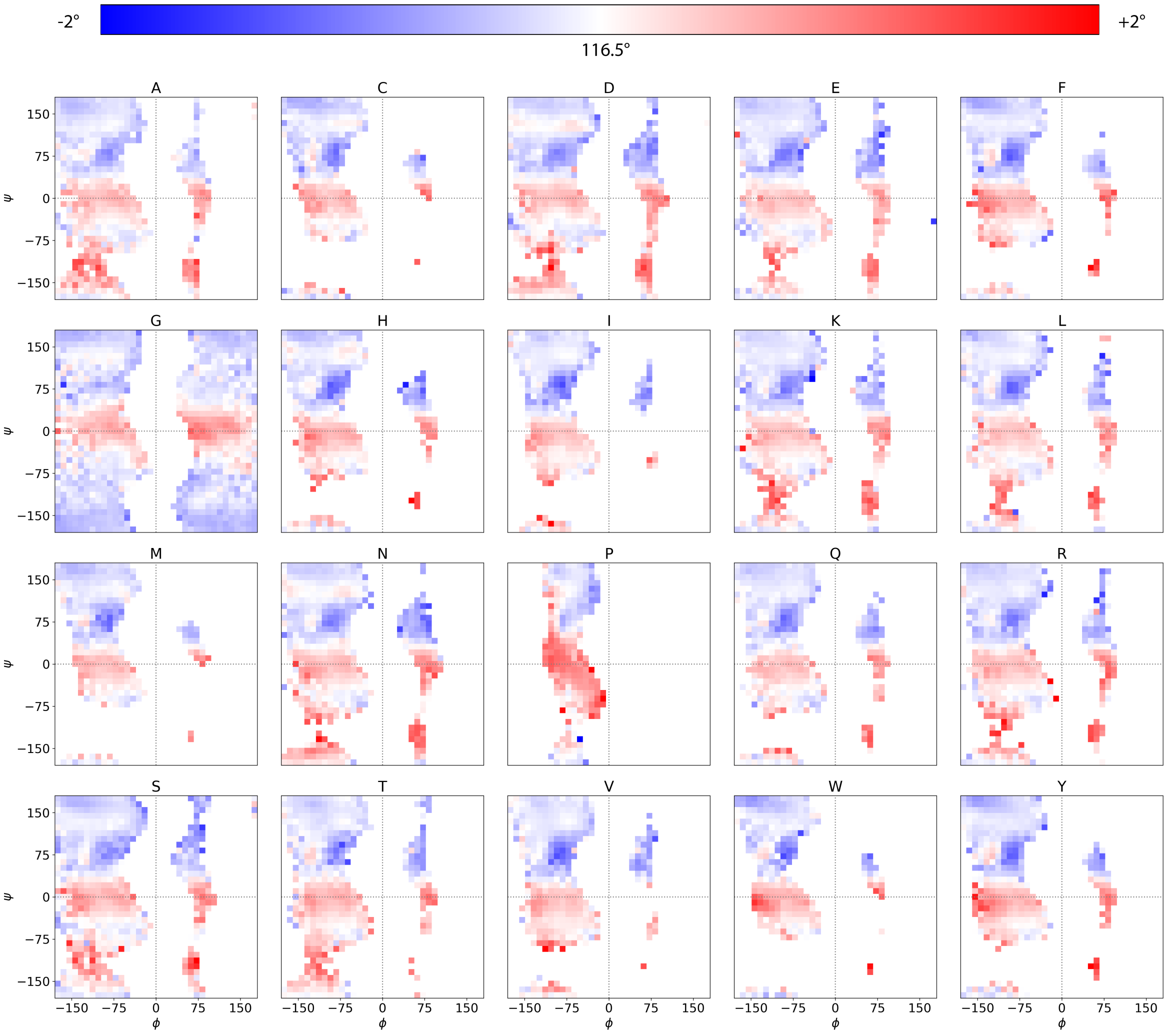


Figure 4.C.4: *$C_\alpha - C - N$ bond angle deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to larger bond angles while the region in blue show reduced bond angles relative to the mean. The $C_\alpha - C - N$ bond angles were categorized according to ϕ and ψ angles rounded to the closest tens.

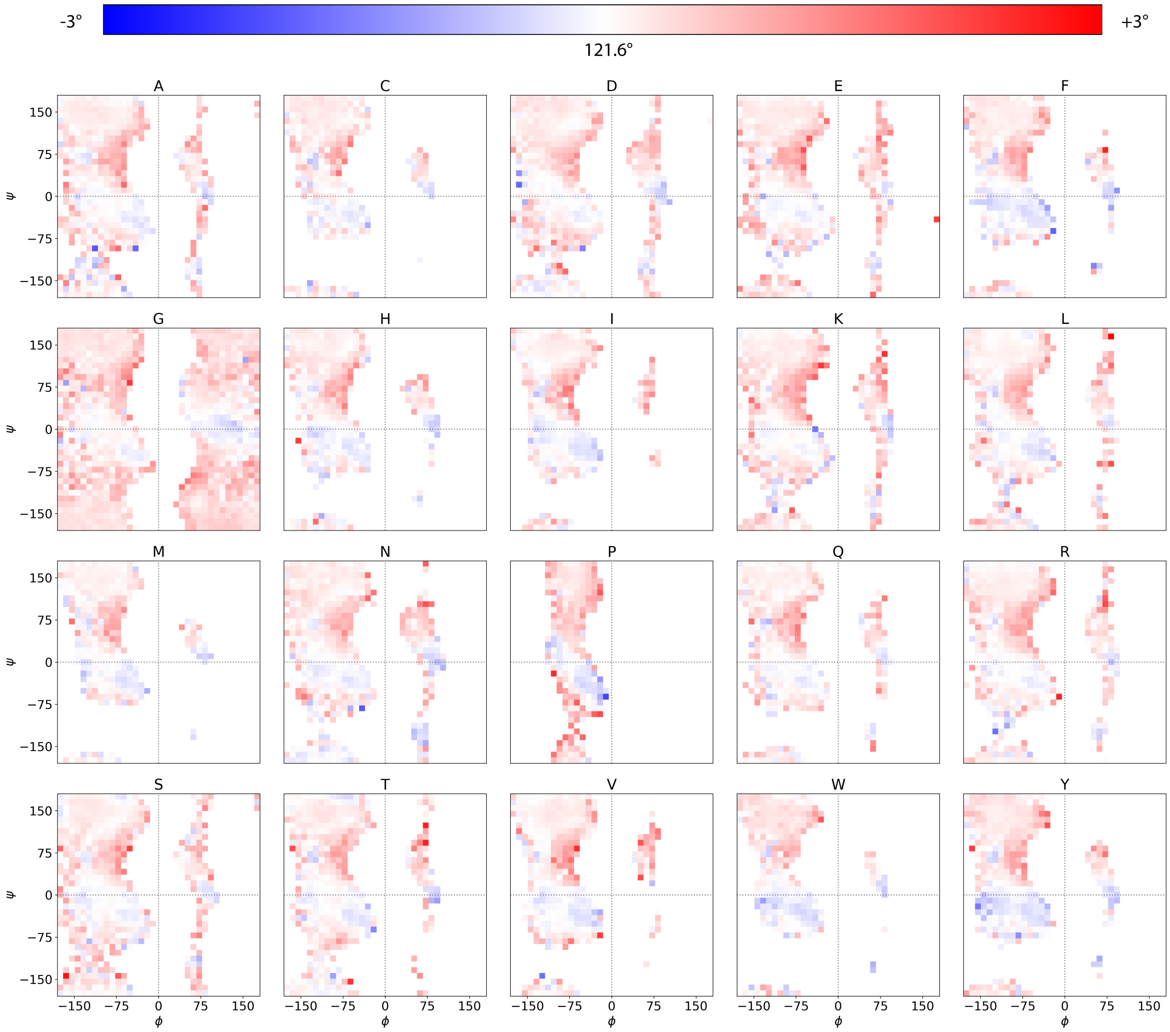


Figure 4.C.5: *$C - N - C_\alpha$ bond angle deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to larger bond angles while the region in blue show reduced bond angles relative to the mean. The $C - N - C_\alpha$ bond angles were categorized according to ϕ and ψ angles rounded to the closest tens.

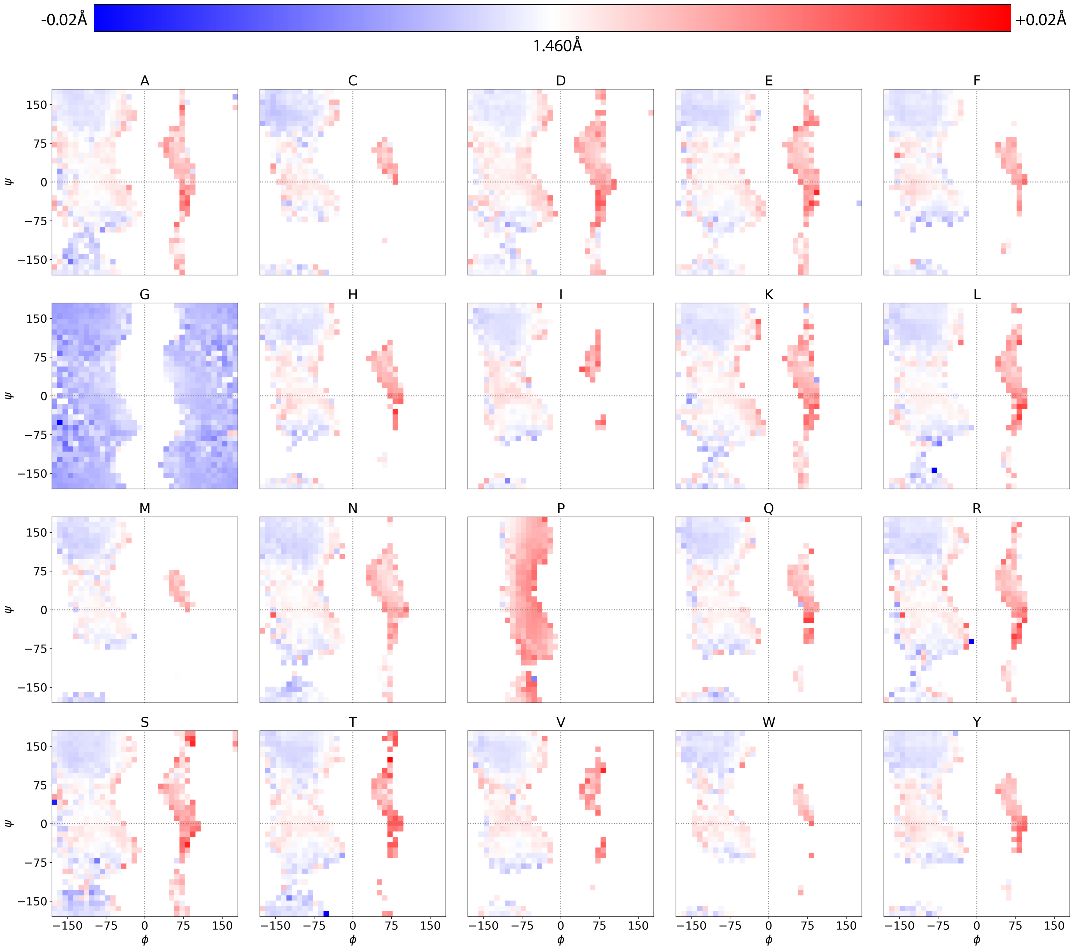


Figure 4.C.6: *$N - C_\alpha$ bond length deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to longer bonds while the region in blue show reduced bond values relative to the mean. The $N - C_\alpha$ bond lengths were categorized according to ϕ and ψ angles rounded to the closest tens.

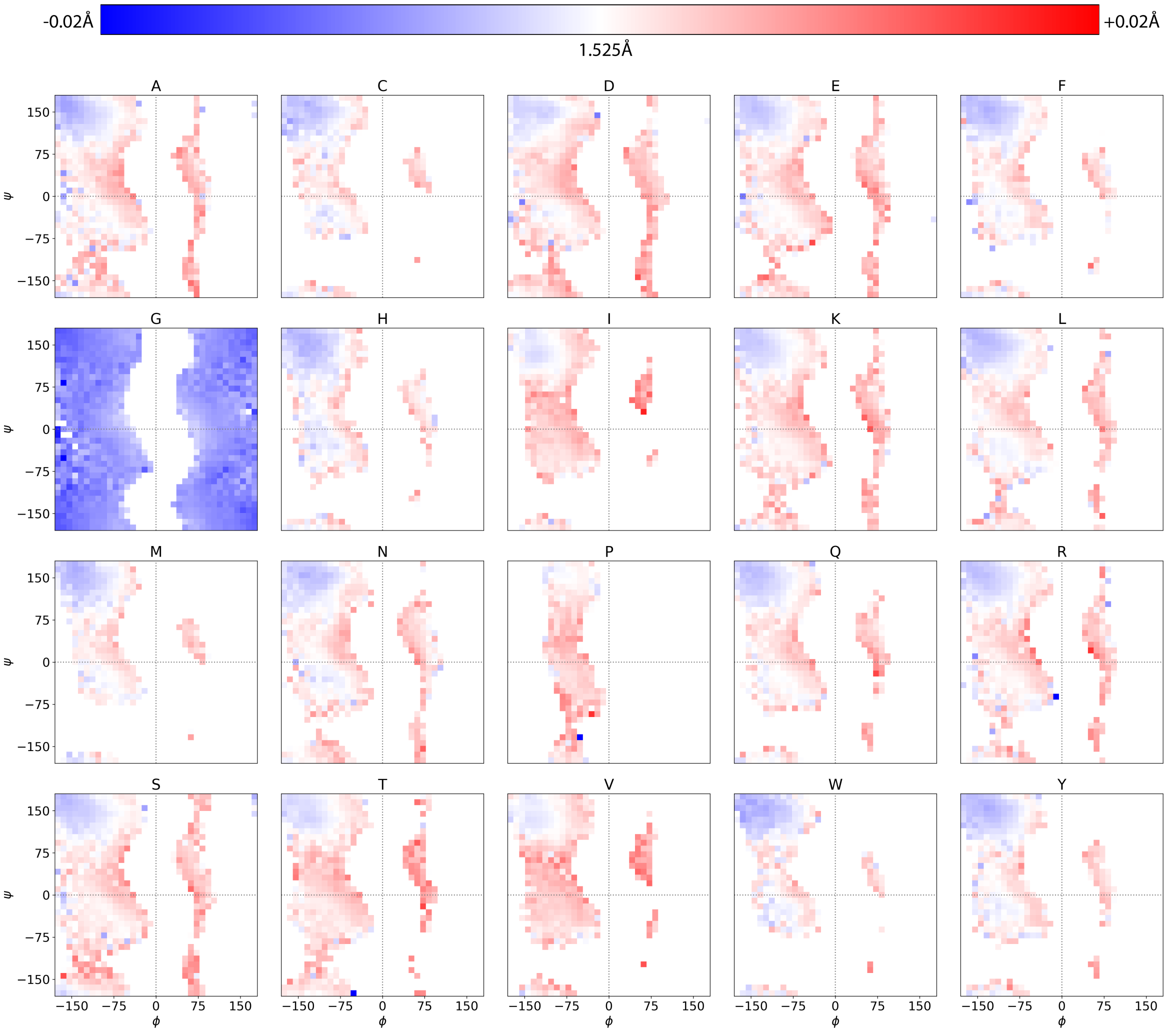


Figure 4.C.7: *$C_\alpha - C$ bond length deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to longer bonds while the region in blue show reduced bond values relative to the mean. The $C_\alpha - C$ bond lengths were categorized according to ϕ and ψ angles rounded to the closest tens.

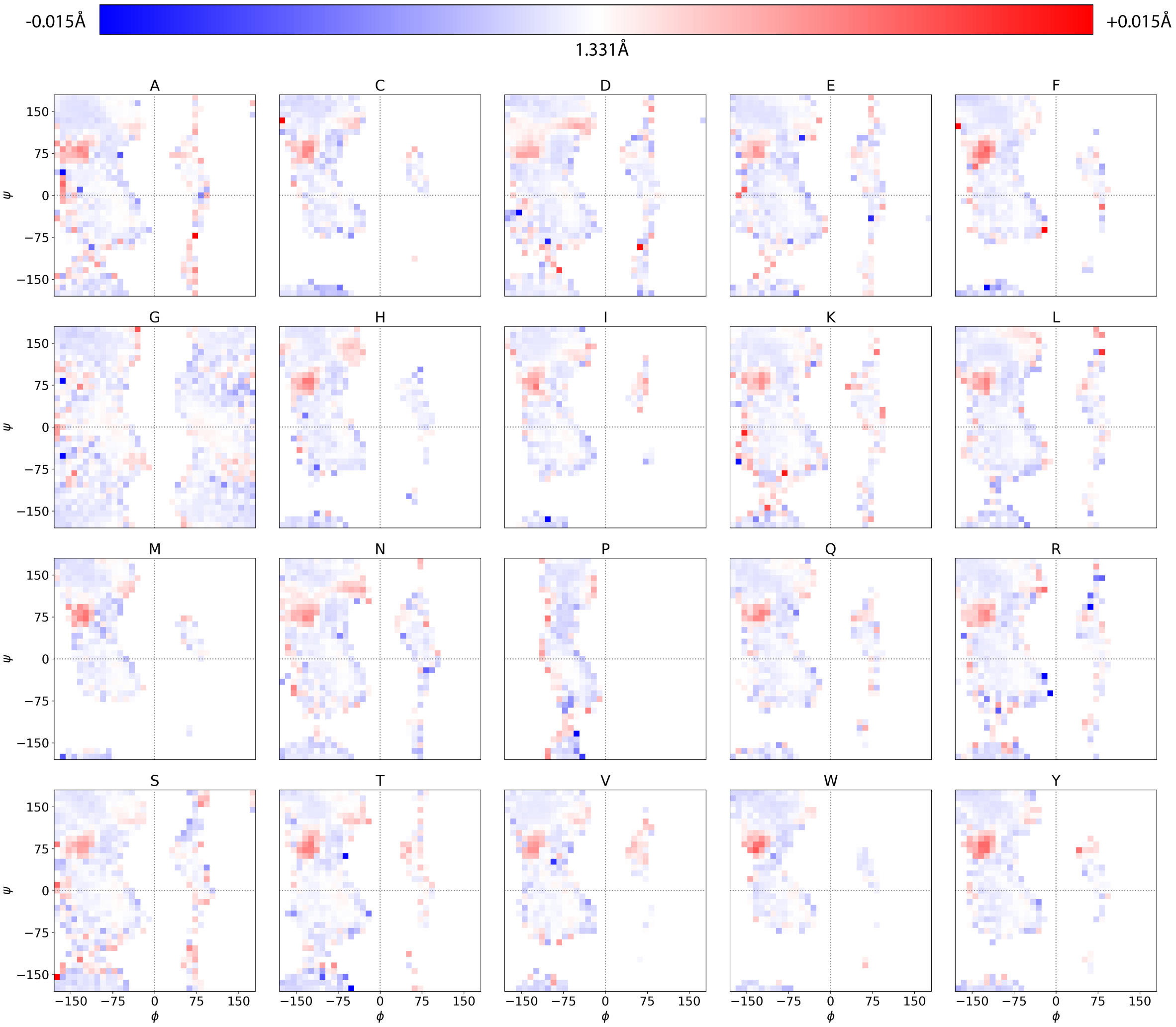


Figure 4.C.8: *$C-N$ bond length deviations from the mean values averaged over ϕ and ψ angles as a function of residue type.* The regions of red correspond to longer bonds while the region in blue show reduced bond values relative to the mean. The $C-N$ bond lengths were categorized according to ϕ and ψ angles rounded to the closest tens.

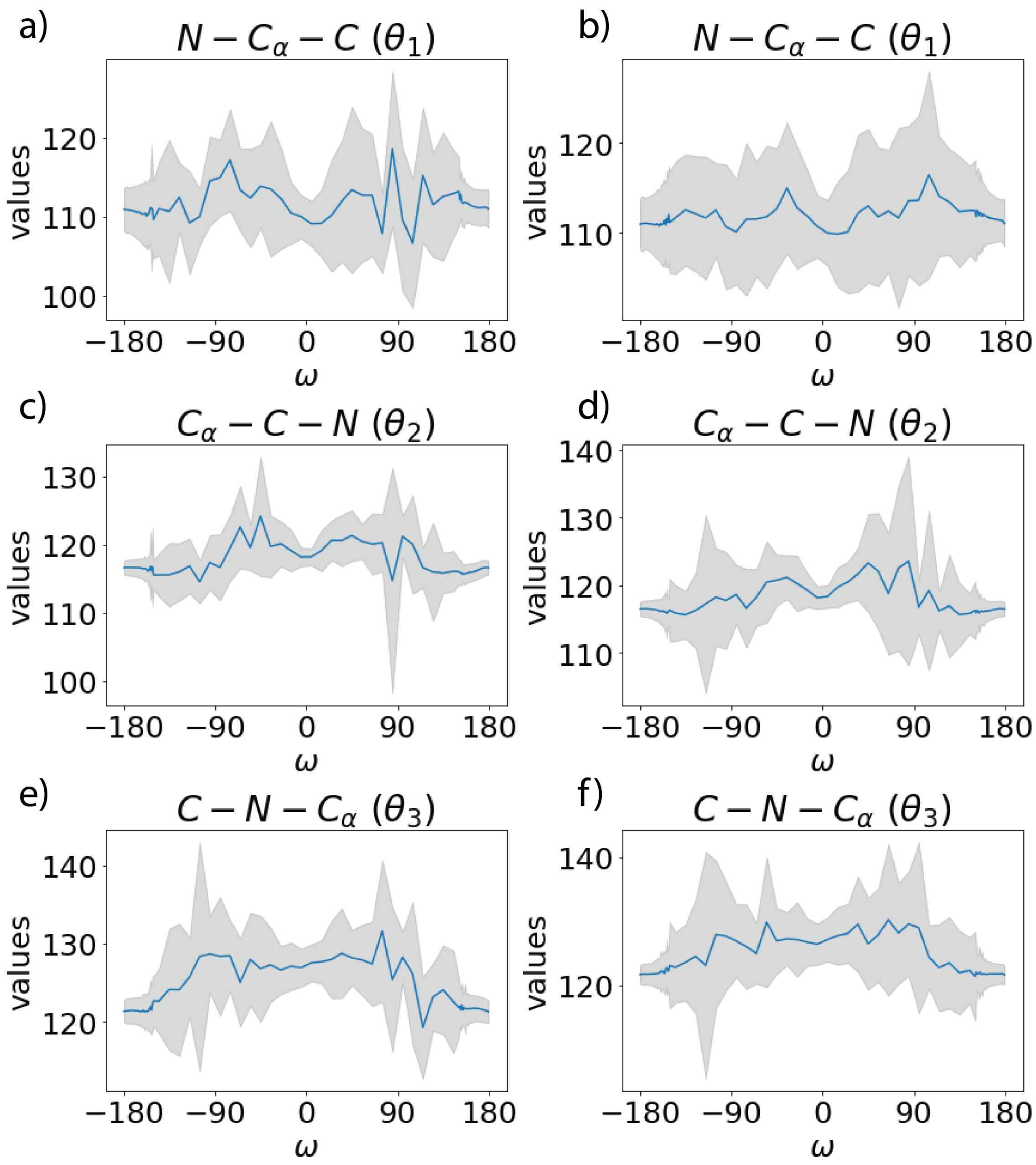


Figure 4.C.9: *Comparison of bond angle - ω dependence among structures with different qualities.* (a, c, e) Mean values and standard deviations of bond angles (a: θ_1, c: θ_2, e: θ_3) as a function of ω for structures with resolution < 2Å(b, d, f) Mean values and standard deviations of bond angles (b: θ_1, d: θ_2, f: θ_3) as a function of ω for structures with resolution between 2 and 4Å

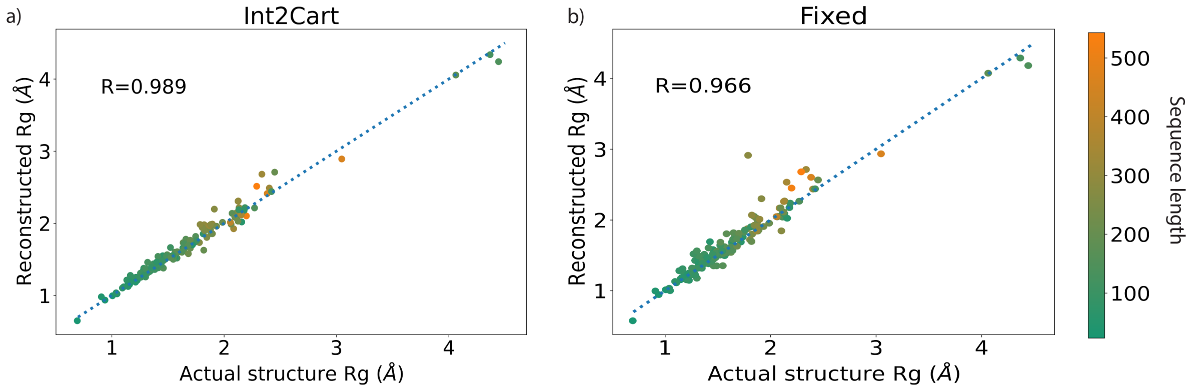


Figure 4.C.10: *The accuracy of radius of gyration when internal coordinates are back-transformed to Cartesian coordinates using Int2Cart or Fixed.* The R_g-match calculates the correlation of radius-of-gyration of individual proteins with the reference proteins. Correlations of R_g-match values for (a) Int2Cart and (b) Fixed over the test set.

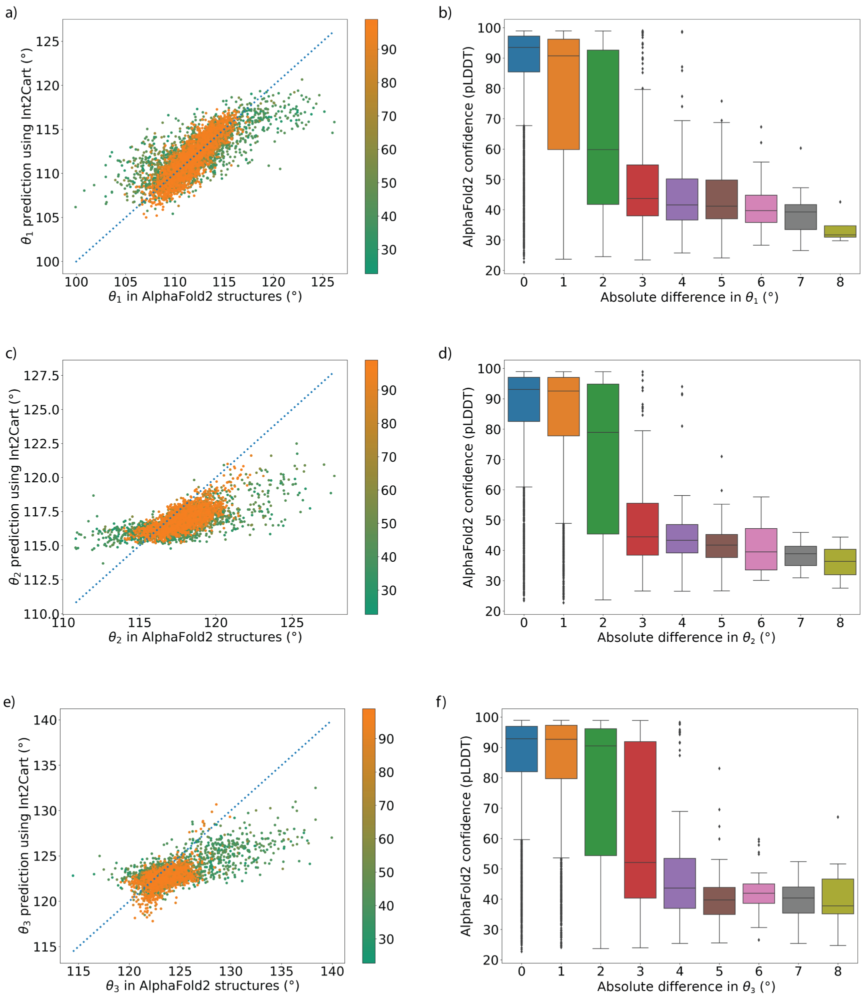


Figure 4.C.11: *Correlation between AlphaFold2 (AF2) structure quality and the agreement between bond angles from the AF2 predicted structures and Int2Cart predicted values using torsion angles from AF2 structures.* (a, c, e) Correlations between θ_1s ($N - C_\alpha - C$ bond angles, a), θ_2s ($C_\alpha - C - N$ bond angles, c) and θ_3s ($C - N - C_\alpha$ bond angles, e) from AF2 structures and Int2Cart predictions colored by AF2 pLDDT scores of the relevant residues. (b, d, f) Box plot showing distribution of AF2 pLDDT scores of individual residues based on absolute differences in θ_1 (b), θ_2 (d) and θ_3 (f) between AF2 structures and Int2Cart predictions. The boxes represent the quartiles of the distribution and the whiskers represent the rest of the distribution. Individual data points are outliers identified from the inter-quartile range.

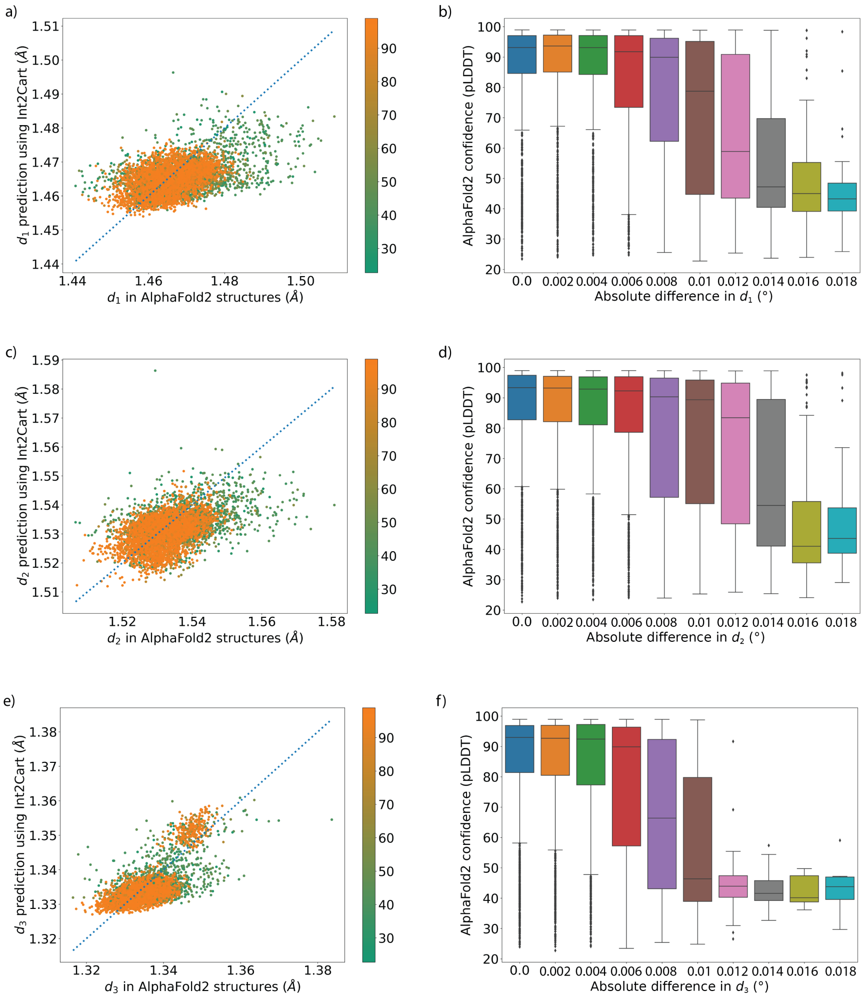


Figure 4.C.12: *Correlation between AlphaFold2 (AF2) structure quality and the agreement between bond lengths from the AF2 predicted structures and Int2Cart predicted values using torsion angles from AF2 structures.* (a, c, e) Correlations between d_1s ($N - C_\alpha$ bond lengths, a), d_2s ($C_\alpha - C$ bond lengths, c) and d_3s ($C - N$ bond lengths, e) from AF2 structures and Int2Cart predictions colored by AF2 pLDDT scores of the relevant residues. (b, d, f) Box plot showing distribution of AF2 pLDDT scores of individual residues based on absolute differences in d_1 (b), d_2 (d) and d_3 (f) between AF2 structures and Int2Cart predictions. The boxes represent the quartiles of the distribution and the whiskers represent the rest of the distribution. Individual data points are outliers identified from the inter-quartile range.

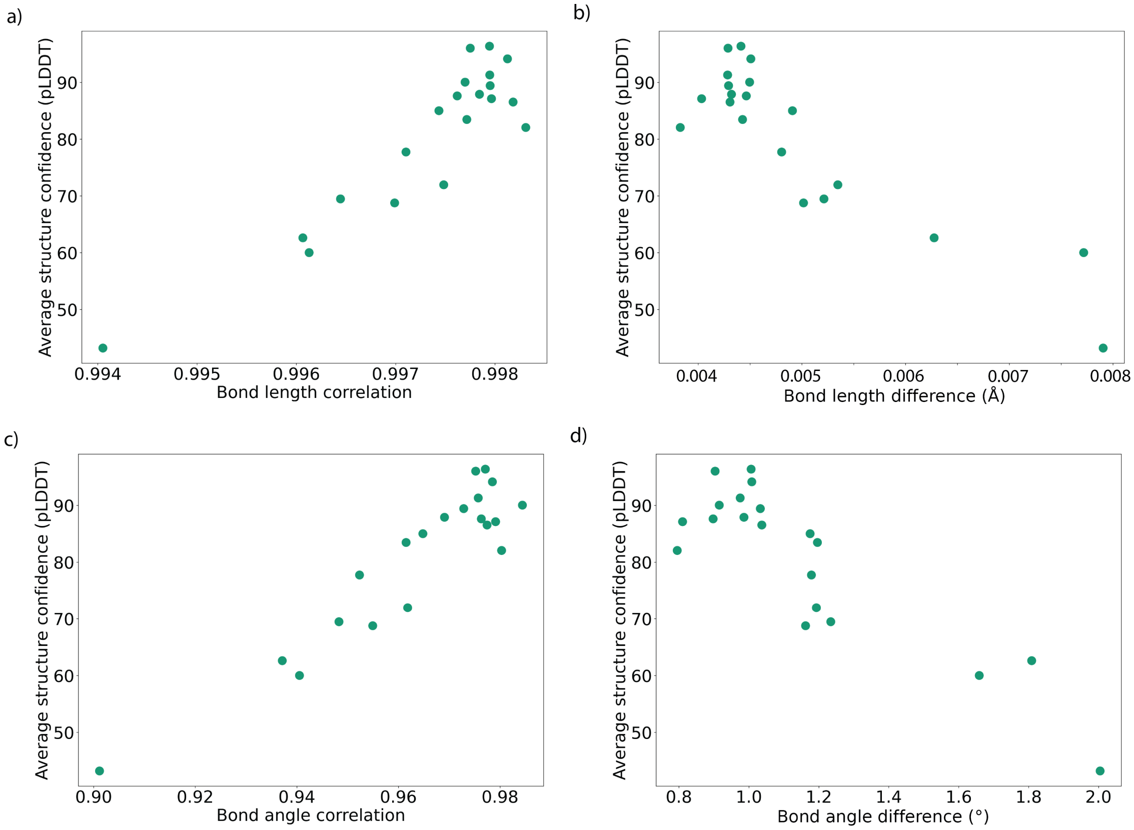


Figure 4.C.13: *Correlation between AlphaFold2 (AF2) structure quality and the agreement between bond angles from the AF2 predicted structures and Int2Cart predicted values using torsion angles from AF2 structures.* (a, b) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond length correlations (a) and absolute differences (b) between AF2 and Int2Cart in an AF2 predicted protein structure (c, d) Relationship between the average AF2 structure prediction confidence (pLDDT score) and all bond angle correlations (c) and absolute differences (d) between AF2 and Int2Cart in an AF2 predicted protein structure

CHAPTER 5

Mining for Potent Inhibitors of Protein Targets with Reinforcement Learning and Real-time Docking of 3D Structures[†]

5.1 Introduction

Existing high-throughput virtual screening approaches to identify protein inhibitors often rely on evaluating existing drug databases such as CHEMBL[1], PubChem[2], and ZINC[3] among others to identify promising small molecule therapeutics. At the same time, the number of potential drug candidates in so-called chemical space is practically infinite, and even the very recent Enamine REAL library of 1.4 billion molecules are still dwarfed by estimates for the total number of possible synthesizable small molecules that range from 10^{24}-10^{60}.[4] Unfortunately, due to the size of such established or expanded databases, screening all compounds according to sufficiently sophisticated structure-based methodologies such as flexible ligand docking can be intractable. Instead simpler methods such as pharmacophore modeling or rigid body docking are often used for navigating through the chemically feasible space, with a tendency towards false-positives being ruled in while false-negatives, i.e. potential optimum lead molecules, can be ruled out.[5, 6]

With the advent of modern machine learning, deep learning models have been proposed that can generate new molecules for multiple viral diseases[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17], and the distribution can be skewed towards molecules with specific properties such as drug likeness using techniques such as variational autoencoders (VAE)[8, 9], transfer learning

[†] Reproduced with permission from: Li J, Zhang O, Kearns FL, Haghighatlari M, Parks C, Guan X, Leven I, Amaro RE, Head-Gordon T. Reinforcement Learning with Real-time Docking of 3D Structures to Cover Chemical Space: Mining for Potent SARS-CoV-2 Main Protease Inhibitors. *arXiv preprint* arXiv:2110.01806. 2021 Oct 5.

[10] and reinforcement learning (RL)[11, 12, 13, 14, 15, 16, 17]. However, most deep learning methods rely on 1-dimensional sequence or 2-dimensional chemical representations of the drug and protein, and do not take full advantage of 3-dimensional structural information of the putative drug, thereby constraining the ability to *generate* drugs with shape and molecular compatibility with the target active site. Recent work has also explored chemical space in the vicinity of some starting molecular scaffold and running docking simulations on these derived molecules[18], however there has been no method that develops new drug molecules with real-time 3d structural docking to guide the efficient exploration of an immense chemical space with the aid of machine learning.

In this work, we propose a novel workflow dubbed "iMiner" that mines chemical space for new tight binding inhibitors by combining deep RL with real-time flexible ligand docking against a protein binding site (Figure 5.1). We represent putative inhibitors as Self-Referencing Embedded Strings (SELFIES)[19] that are generated from an Average Stochastic Gradient Descent Weigh-Dropped Long Short Term Memory (AWD-LSTM)[20] recurrent neural network (RNN), allowing wide coverage of chemical space. We illustrate the RL training procedure of iMiner that uses on-the-fly AutoDock Vina[21] with the 3d structures of the protein with the predefined binding pocket and the generated inhibitors. The Vina docking scores are used to adjust the RNN so that the distribution of generated inhibitor molecules are shifted towards those that more strongly interact with the protein. We perform docking with a second docking software, Schrödinger's Glide SP[22], to build consensus for a drug's strong binding affinity to the target protein, and a final filtering based on synthetic accessibility (SA), druglikeness, and elimination of PAINS [23] molecules.

As the COVID-19 pandemic continue to be a global crisis, finding effective antiviral drugs to treat patients infected with the SARS-COV-2 virus is still of pressing importance. Among all the proteins related to the SARS-COV-2 virus, the main protease (Mpro) has arguably received the most attention with respect to drug development[24], in part because it is one of the earliest SARS-COV-2 proteins in which the 3d structure has been fully determined experimentally.[25] We use this Covid-19 relevant example to illustrate the iMiner workflow, in which we ultimately propose 51 molecules as potential Mpro inhibitors that are worthy of experimental validation (work in progress). Furthermore, we compare our top hits generated with the iMiner workflow with the molecules submitted to the COVID moonshot project[26], a crowdsourcing effort aimed at developing a novel inhibitor for Mpro, and show that we achieve a broader coverage of the inhibitor drug chemical space. We also find excellent shape and molecular attributes of the inhibitors generated by our model in regard their compatibility with the actual target cavity in Mpro, which is a direct consequence of the real-time docking with actual 3d structures during the training procedure. Further analysis of non-bonded interactions between the found inhibitors with specific binding pocket residues in Mpro also create testable hypotheses in regards their potential role as antivirals to treat COVID-19.

Although we have illustrated the workflow's first use on a pressing and timely test case – i.e., inhibition of SARS-CoV-2 Mpro due to the desperate need for antiviral treatments of COVID-19 – the iMiner method is highly generalizable. As our workflow only requires a 3D structure of the target protein with a pre-defined binding site, iMiner can be readily adapted to generate small molecules for other protein targets. Thus, we believe our ML algorithm will be of great interest to the drug design community to rapidly screen novel regions of chemical space in real-time for other anti-virals or small molecule therapeutics in the future. All scripts required to run our workflow on an arbitrary protein target can be found on a public GitHub repository*.

5.2 The iMiner Machine Learning Workflow

Here we describe the entire iMiner life cycle for generating new inhibitor molecules in more detail.

SELFIES representation of inhibitor molecules. An arbitrary molecule can be represented as a topological graph using two main approaches: adjacency matrix based methods and string based methods. The former uses an N by N matrix to encode a molecule, where N is the number of atoms in the molecule, and the values of the matrix are typically bond orders between atoms. An adjacency matrix is not ideal for generative tasks, because the size of the molecule that can be generated should not be fixed, and the learning of chemical knowledge by a ML model through adjacency matrix can be difficult. Instead, string based methods are more suited for molecular generation tasks, and SMILES strings have been the standard for molecular representation due to its conciseness and readability. However, SMILES strings have relatively complex syntax, require matching of open and close brackets for branching, and ring modeling/modification is not trivial. Therefore, generating novel, chemically correct compounds through use of SMILES strings can be challenging.

The SELFIES molecular representation[19] is specifically designed to ensure that all generated strings correspond to valid molecules. By utilizing `[Branch]` and `[Ring]` tokens with predefined branch lengths and ring sizes, as well as generating symbols using derivation rules, the SELFIES representation guarantees that valence bond constraints are met, and any combinations of tokens from its vocabulary corresponds to a valid molecule. Therefore, we have used SELFIES in our generative model to encode molecules since it does not need to learn chemical syntax rules, and can allocate more of its learning capacity towards generating valid molecules with properties of interest as shown in Figure 5.1.

Pre-training the inhibitor molecule generation. Conceptually, generating molecules using string representation is similar to how text is generated in a natural language processing task. Our method starts with a specific `[Break]` token, and for each molecule, we utilized an RNN that takes in the last token in the string, together with the hidden state from last step to predict a distribution of tokens following the current string. In this work a specific variant of the RNN, known as the AWD-LSTM, was used due to its exceptional performance in similar generative tasks (Figure 5.1).[20] The network was pre-trained using supervised-learning (SL) of all molecules from the ChEMBL database to learn the conditional probability distributions of tokens that correspond to drug-like molecules. When our trained

*https://github.com/THGLab/iMiner

generative model is used for generating new molecules, a new token is sampled according to the predicted probabilities, and this new token is concatenated to the input string to sample the next token, until the `[Break]` token is sampled, in which case a complete molecule has been generated.

The performance of our pre-trained generative model was evaluated using the GuacaMol benchmarks[27], which probe 5 different aspects of the distribution of generated molecules with respect to the training dataset (Table 5.1). Model "validity" reports the proportion of molecules that are syntactically correct. Because we generated molecules via SELFIES representations, we achieved close to 100% validity for all generated molecules. Invalid molecules were either empty strings, or molecules for which the SELFIES package failed to convert into a SMILES string, and therefore were discarded before the next workflow steps.

Table 5.1: GuacaMol benchmarks for the pretrained generative model and after RL training

Benchmark	Pretrained model	After RL
Validity	0.998	0.998
Uniqueness	0.999	0.983
Novelty	0.867	0.999
KL divergence	0.985	0.791
Frechet ChemNet Distance	0.870	0.007

Model "uniqueness" reports how many generated molecules are duplicates vs. those which are genuinely distinct. Our pretrained models illustrated high uniqueness, indicating the model is able to generate a wide variety of non-redundant molecules. Model "novelty" is defined as the proportion of generated molecules that do not exist in the training dataset. Our model's high novelty indicates that it is not memorizing molecules from the training dataset, but is indeed generating molecules that it has not seen before. Kullback–Leibler (KL) divergence measures differences in probability distributions of various physicochemical descriptors for the training set and the model generated molecules. As defined by GuacaMol, a high KL divergence benchmark such as predicted for our model suggests that our generated molecules have similar physicochemical properties to that of training dataset. Finally, Frechet ChemNet Distance (FCD) utilizes a hidden representation of molecules in a previously trained NN to predict biological activities, and thus captures both chemical and biological similarities simultaneously for two sets of molecules.[28] Molecules generated by our pre-trained model also have high FCD values, indicating that our molecules are expected to have similar biological activities as molecules from the ChEMBL training dataset.

We then validated our pre-trained distributions using 13 drug-likeliness properties between our generated molecules and randomly sampled molecules from ChEMBL database that we used for training. The molecular properties considered are well-recognized chemical features related to the drug-likeliness of a molecule which can be obtained through 2D topological connectivity of the molecule: fraction of sp^3 hybridized carbons, number of heavy

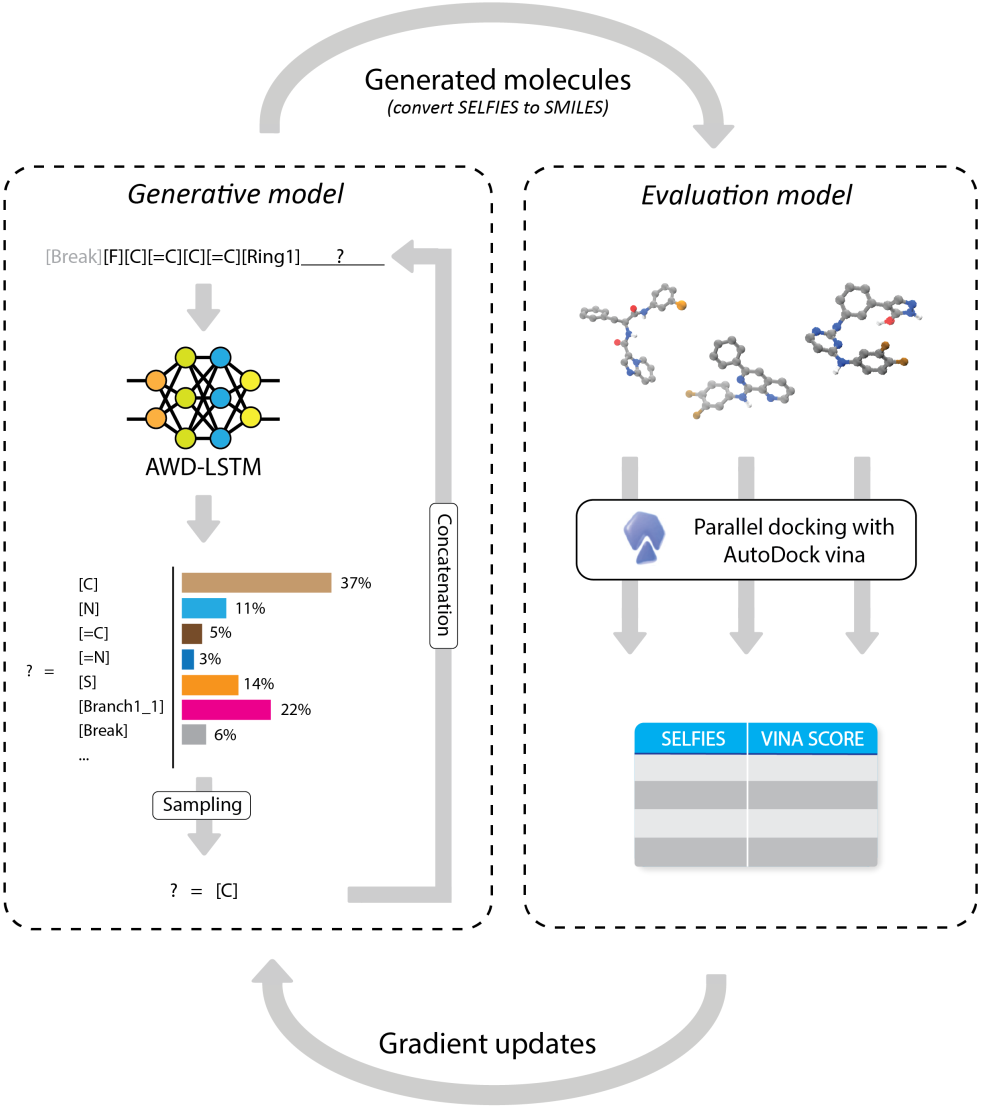


Figure 5.1: *Illustration of the overall structure of the iMiner workflow which highlights the two major machine learning components, the generative and evaluative models and their interplay.* The generative model uses SELFIES representations for molecules and a recurrent neural network to "mine" for new molecules that are presented to the evaluative model for 3D docking using AutoDock vina. Vina scores are used in the loss function to drive gradient updates of the neural network.

atoms, fraction of non-carbon atoms in all heavy atoms, number of hydrogen bond donors and acceptors, number of rotatable bonds, number of aliphatic and aromatic rings, molecular weight, quantitative estimate of drug-likelihood (QED) value[29], approximate log partition coefficient between octanol and water (alogP)[30], polarizable surface area (PSA), and the number of structural alerts.[31] Despite the fact that during pre-training only token distributions were used as training targets, all distributions collected from our generated molecules closely follow the distributions from the ChEMBL database (Appendix Figure 5.C.1). This result suggests our pre-trained model has learned key concepts of “drug-likeness” and provides a good starting point for the RL procedure.

The evaluation module. After our generative model was pre-trained, we employed an RL workflow to bias the distribution of generated molecules towards specific properties of interest. RL training allows the model to interact with an environment by performing actions according to a policy model, and uses the feedback from the environment to provide training signals to improve the model. In this work, the pre-trained generative model is taken as the policy, and in each iteration 2000 molecules were generated and sent to the evaluation module (Figure 5.1).

The central component of our evaluation model is docking with AutoDock Vina executed through cloud computing in parallel with the RL. Within our evaluation model, the Vina score calculator is set up to take a SMILES string representing the ligand, and the 3D structure of the protein target, together with a predefined docking region as input. AutoDock Vina then explores dihedral degrees of freedom and identifies the optimal conformation of the input inhibitor for placement in the designated protein binding site. Finally, AutoDock Vina returns the Vina score as an approximation of the binding energy between the ligand and the protein. Multiple instances of the Vina score calculator tasks were established through Microsoft Azure Batch to allow high-throughput evaluation of the generated molecules. Vina scores were then cycled back to the generative model to improve molecule generation through proximal policy optimization (PPO)[32], as will be discussed in next section. We emphasize that by using a physics-based docking model which utilizes full 3D structure of our target protein and generated molecules as the critic, the training of the policy model is less likely to be contaminated due to exploiting failure modes of a neural-network based critic, an issue called *wireheading*[33]. Instead, we benefit from a more reliable training signal and reduce the false positive and false negative rates of the generated molecules.

Vina scores alone are not sufficient to reliably train a molecule generator, as shown in the Supporting Information (Appendix Figure 5.C.2) because it will not always satisfy requirements for drug-likeness. To ensure that our generated molecules still bear drug-like properties, we incorporated an additional metric into the reward, S_{DL}, which is a weighted average of the log likelihood for the 13 different drug-like properties used in pre-training assessment. Formally, our drug-likeliness score S_{DL} is defined as:

$$S_{DL}(X) = \sum_i \sigma_i \log p_i(\mathrm{prop}_i(X)) \tag{5.1}$$

where $\mathrm{prop}_i(X)$ calculates the ith property of a molecule X and p_i is defined by the probability distribution of property i by all molecules in the ChEMBL database. The parameter σ_i is defined as:

$$\sigma_i = S_i^{-1} / \sum_j S_j^{-1} \tag{5.2}$$

where S_i is the entropy of the distribution of property i,

$$S_i = -\sum_x p_i(x) \log p_i(x) \tag{5.3}$$

such that a narrower distribution from the ChEMBL database contribute more to the drug likeliness score, and defines the weights for each property as proportional to the inverse of the entropy. Introducing this additional reward ensures our model also accounts for similarity of generated molecules to the drug-likeliness present in the ChEMBL database, and ensures that our generated molecules are more likely to be optimal leads for further drug design endeavors.

Reinforcement learning with multiple rewards. Our pretrained policy model defines a probability distribution for an arbitrary sequence of tokens from the SELFIES vocabulary, since the generation of the sequence is a Markovian decision process (MDP), and can be written as:

$$p_\Theta(s_T) = p_\Theta(s_1|s_0) p_\Theta(s_2|s_1) ... p_\Theta(s_T|s_{T-1}) \tag{5.4}$$

where s_0 corresponds to a starting state with `[Break]` as the only token in the string, s_t corresponds to an intermediate state with a finite length string of SELFIES tokens not ended with the `[Break]` token, and s_T corresponds to the terminal state, with the last token being `[Break]`, or the length of the string exceeding a predefined threshold. $p(s_t|s_{t-1})$ is the transition probability at timestep t, which is the probability distribution of the next token from the generative RNN with network parameters Θ. For each terminal state not exceeding the length limit, a corresponding molecule can be decoded, and the Vina score S_{vina} and drug-likeliness score S_{DL} can be calculated. The total reward for a terminal state with a decoded molecule X is then defined as:

$$r(s_T) = \lambda \max(S_{DL}(X), 0) - \min(S_{vina}(X), 0) \tag{5.5}$$

since the drug-likeliness score needs to be maximized and Vina score needs to be minimized. The λ parameter controls the balance between the physical Vina score and the drug-likeliness score in the reward function, but in this work we simply used $\lambda = 1$. Negative S_{DL} is upward clipped to 0 and positive S_{vina} is downward clipped to 0 to ensure the reward is non-negative. The expected reward under the MDP is then

$$J(\Theta) = E_{s_T \sim p_\Theta(s_T)}[r(s_T)] \tag{5.6}$$

Further details of the RL training procedure are given in the Methods section.

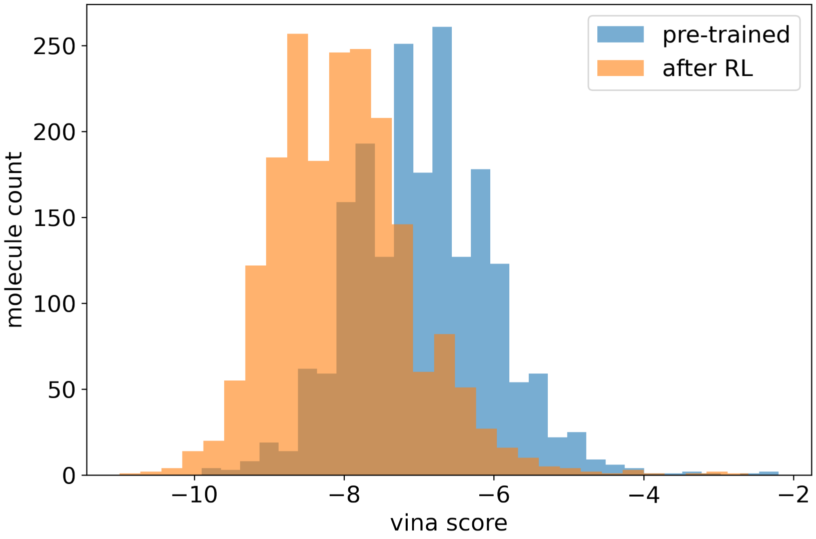


Figure 5.2: Comparison of AutoDock Vina score distributions for the pre-trained model (blue) and the model trained by reinforcement learning (orange). The mean vina score decreased from -6.95 kcal/mol to -8.01 kcal/mol after RL training.

Figure 5.2 compares the distribution of Vina docking scores for molecules generated from the model prior (the pre-trained model) and after RL training which shows a clear shift towards more favorable vina scores. The average Vina score of molecules decreased from -6.95 ± 0.94 kcal/mol to -8.01 ± 0.94 kcal/mol, showing that on average more molecules have stronger interactions with the predefined Mpro docking region. In addition, the GuacaMol benchmarks were also evaluated for the model after RL training, which are also shown in Table 5.1. Except the Frechet ChemNet Distance (FCD), all other benchmarks are relatively close to the pretrained model, indicating that the RL training does not hurt the quality of the generated molecules, and they are still similar to the structures from ChEMBL database. However, FCD has changed significantly, which means the newly generated molecules have different biological activities than the molecules from ChEMBL database. The changes seen in FCD are expected, since, after training, the generated molecules should target a specific cavity of SARS-CoV-2 Mpro, a target for which there are currently no FDA approved treatments. Thus, the FCD differences validate that RL is properly steering the distributions of generated molecules away from its initial distribution.

Validation and filtering of new inhibitor molecules. Validating results from, or checking for consensus between, one docking program with another is often considered standard practice as scoring functions from different programs may have limited accuracy or be parameterized for differing test cases.[34] Furthermore it is desirable to filter out molecules that are non-specific binders (Pan-assay interference compounds or PAINS) in which we use swissADME[35] to check for any PAINS alerts[23], as well as Lipinski rule violations[36], and to evaluate the synthetic accessibility (SA) scores of these molecules. Figure 5.3 illustrates the procedures for post-filtering using these additional metrics, which we describe in more detail here.

We start by collecting all molecules from intermediate RL training iterations with a Vina score <-9.0, arriving at our "vina-selected set" containing 33,105 molecules (the number of molecules from each training iteration is provided in Figure 5.C.3). As expected, more molecules from later iterations were selected, since molecules from later iterations were driven towards having lower Vina scores. Glide Standard Precision (SP)[22] docking was performed on all molecules in our vina-selected set with the flexibility to optimize the conformation again with respect to the Glide scoring function. This way we could fully exploit Glide docking as a cross-validation for the generated molecules. Even though the molecules were all good candidates according to Vina score, their glide docking score still showed a wide distribution. We then applied a filter with Glide Gscore (Glide Score) <-8.0 and a drug-likeliness score filter of >2.7 to exclude any structure that is not sufficiently drug-like. After applying these filters we obtained the glide-selected set with 240 molecules in total. The final step was to run these 240 molecules through a final set of filters requiring no PAINS alerts, no Lipinski rule violations and SA scores <3.5.

5.3 Results

The outcome of the iMiner workflow formulated a final set of 51 molecules for the Mpro catalytic site shown in Figure 5.4. These molecules are predicted to be consensus Mpro inhibitors by both AutoDock Vina and Glide SP, they satisfy drug-likeliness criterion, and are relatively easy to synthesize due to their predicted low SA scores. The full SMILES representations and Vina scores, Glide Gscores, and SA scores are provided in Appendix Table 5.B.1.

Comparison of chemical diversity of inhibitors discovered by iMiner. Figure 5.5 compares the total chemical space coverage of molecules generated using iMiner and the COVID-moonshot project for MPro[26], as well as molecules from ChEMBL for reference, by performing dimension reduction on the hidden representation of these molecules encoded through ChemNet. ChemNet is a deep network trained on canonized SMILES strings of molecules as input and encodes each molecule into a 512-dimensional latent vector to predict their chemical and biological properties[28], and the dimensions were further reduced to 2 through the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm[37] for better visualization. Plot points on the resultant figure indicate individual molecules, and points are drawn close to or far from one another based on their degree of chemical similarity: points closer to one another indicate chemically similar molecules and therefore correspondingly low coverage of chemical space, while widely dispersed points indicate dissimilar molecules and therefore broad coverage of chemical space. The nearly 1500 COVID-moonshot molecules

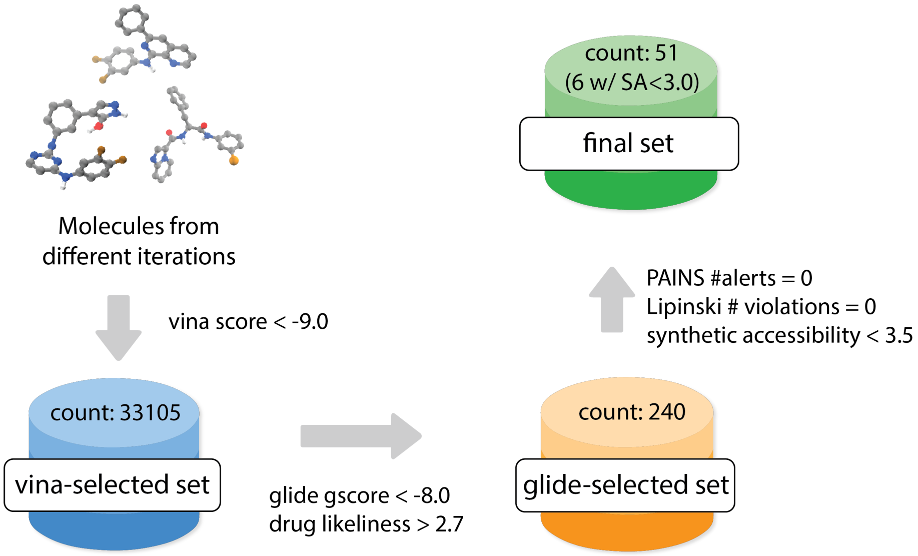


Figure 5.3: *Illustration of the filtering process for chemical and biological feasibility of iMiner generated molecules.* The filtering procedure from molecules collected from intermediate training iterations is based on both favorable Vina and Glide SP docking scores, high drug-likeliness scores, no PAINS and no Lipinski's Rule violations.

are also color-coded with their experimentally determined pIC_{50} values, and our generated molecules in the vina-selected set are color-coded with their Vina scores.

The visualization clearly shows that the molecules generated by iMiner covers a broader chemical space and are spread evenly within plotting range than those molecules published on the COVID-moonshot website which form several tight clusters. We recognize that one of the reasons for the COVID moonshot molecules to be clustered in chemical space is that many of these molecules are generated through an inspirational approach, i.e., later molecules are borrowing designing ideas and sub-structures from molecules submitted earlier. By comparison, our final-51 set of molecules are dispersed throughout chemical space, which is an important characteristic of our workflow, since it provides a wide variety of structures as candidates for lead optimization. Interestingly, even compared to samples from the training dataset (ChEMBL), the molecules in the vina-selected set are still more diverse, which suggests the model was encouraged to explore more of chemical space during RL training while still reporting low Vina scores. Finally, we also see that the 51 molecules from iMiner are coming from different regions of the chemical space spanned by molecules generated from the model. As drug leads built on single or closely related scaffolds might be ruled out entirely during drug development, a wider coverage of the chemical space gives us a better chance of developing an effective lead for an Mpro inhibitor for treating SARS-CoV-2.

The molecular interactions between generated inhibitors of Mpro's catalytic

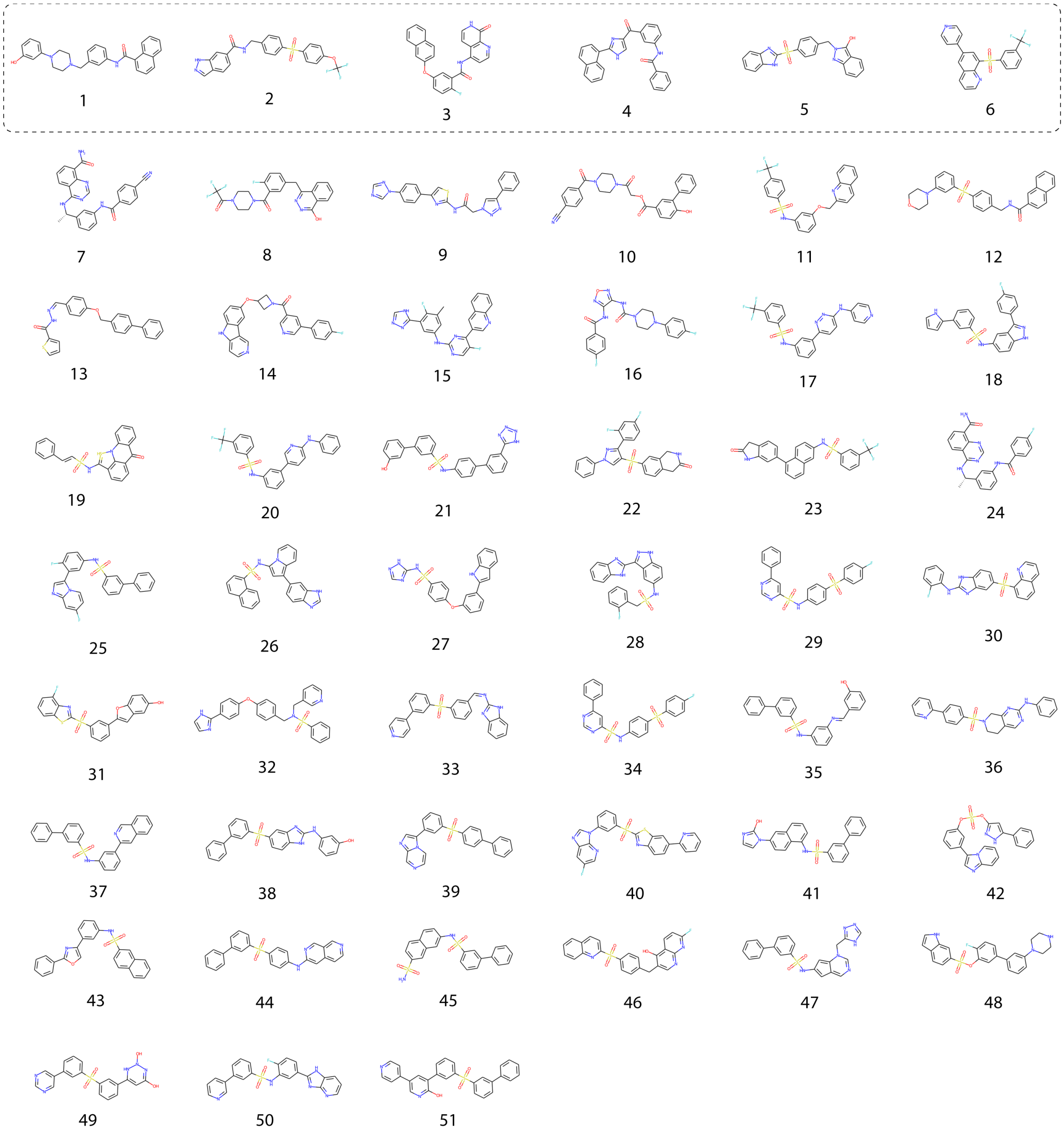


Figure 5.4: *Prediction of 51 molecules that are tight binding inhibitors of Mpro in the final set generated from the iMiner workflow.* We propose further experimental validations on these molecules as potential SARS-CoV-2 Mpro inhibitors (work in progress). The first 6 molecules in the dashed frame have better synthetic accessibility scores than the rest. The diversity over chemical space of these proposed inhibitors is evident from metrics described in Table 5.1 and Figure 5.5.

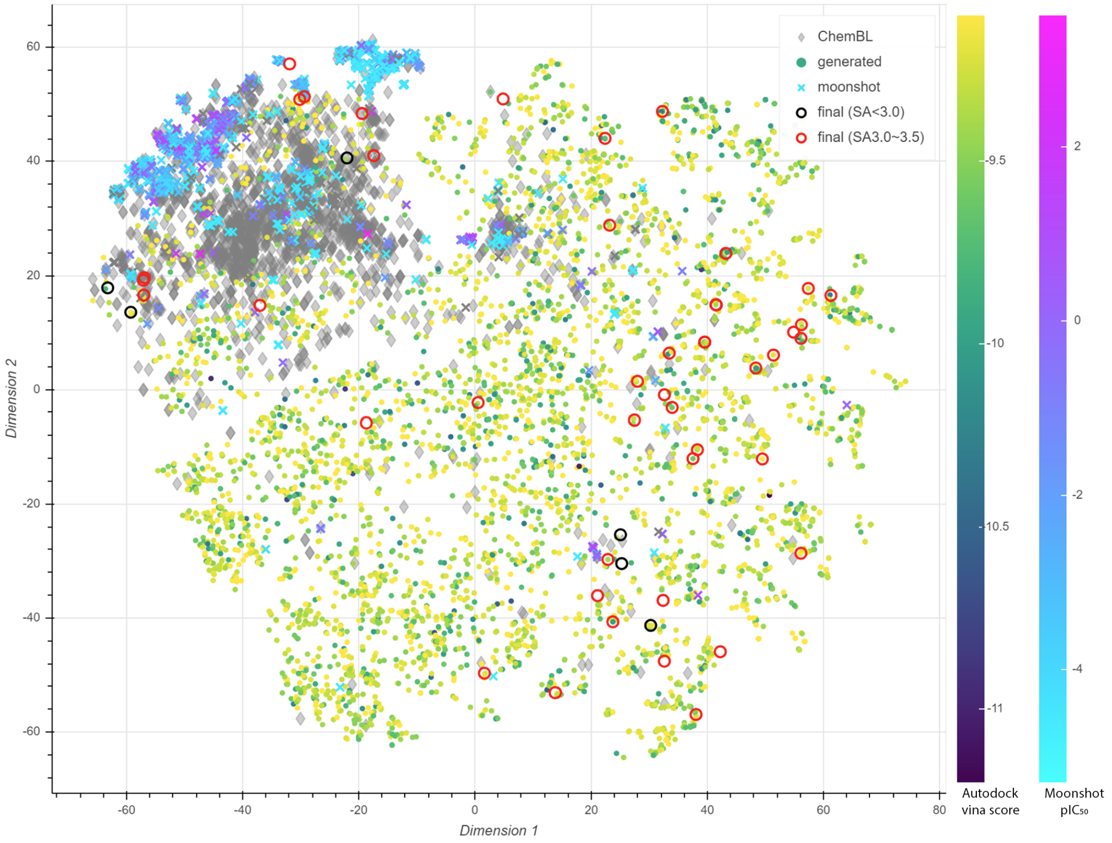


Figure 5.5: *Dimensionality-reduced latent space scatter plot for molecules from the COVID-moonshot project (crosses), generated molecules from the RL model (dots), molecules randomly sampled from the ChEMBL database (diamonds) and molecules in the final-51 set (circles).* Molecules on the figure are encoded by ChemNet[28] and the latent space vectors undergo dimensionality reduction by principal component analysis (PCA)[38] and t-distributed stochastic neighbor embedding (t-SNE)[37]. Molecules from the COVID-moonshot project are color coded by their experimental pIC_{50} values according to the color bar on the right, and molecules generated by our model are color coded by Vina docking scores according to the color bar on the left.

site. In Figure 5.6A, we show an overlay of several iMiner generated molecules in their optimal binding conformations determined through AutoDock Vina with respect to the surface of the binding pocket in which the predicted binding orientations fit nicely into the Mpro's catalytic pocket. Additionally, ligand functional groups mirror the hydrophobicity requirements imposed by the Mpro binding site topography, meaning the generated molecules indeed have optimized interactions with the pocket. These results are no doubt due to our inclusion of real-time, explicit, flexible ligand docking in our evaluation model as well as a result of requiring minimization of Vina score distributions. Through this visualization we also see an interesting and encouraging result: although our final set of 51 molecules represent vastly different regions of chemical space, these molecules are relatively similar in size (i.e., similar number of heavy atoms), and the optimal docking conformations adopt similar shapes. These results illustrate the true power of our model, that we can quickly enumerate and expand upon the searched chemical space while still ensuring all generated molecules appropriately fit in the target protein pocket.

Figure 5.6B-E provide two representative examples of the molecular interactions between an iMiner predicted inhibitor and the Mpro binding site residues. Many and various types of favorable ligand-target interactions are observed, including hydrogen bonds, halogen bonds, and different types of π interactions. For example, CYS145 contributes to the π-Sulfur interaction in the first molecule illustrated in Figure5.6B and C, but it participates in a conventional hydrogen bond to the SO_2 group in the second molecule illustrated in Figure5.6D and E. Furthermore, when comparing the two proposed inhibitors each molecule exhibits unique interaction types to a different or complementary set of MPro protein residues. This variety in intermolecular interaction types stemming from the same protein binding site is a direct result of our enumeration of chemical space and our construction of novel ligand scaffolds.

5.4 Conclusions

In this work we have shown that by combining real-time docking of 3D structures with state-of-the-art reinforcement learning algorithms, we can efficiently navigate through uncharted regions of chemical space while maintaining good metrics for synthetic feasibility and drug-likeness. As illustrated using the exemplar target, the Mpro catalytic site, the ultimate final set of 51 inhibitor molecules proposed by our model are optimized with respect to shape and intermolecular interactions to the target protein, but are also diverse enough when compared to other predicted Mpro inhibitor datasets, i.e. molecules submitted to the COVID-moonshot project[26]. We understand the true effectiveness of these molecules as Mpro inhibitors can only be determined through experimental screening. Nevertheless, as we have seen agreement between AutoDock Vina and Glide SP results, and since we have visually inspected the predicted binding modes revealing consistency in intermolecular interactions to the Mpro pocket, we strongly believe there is good evidence that these molecules may be potent Mpro inhibitors.

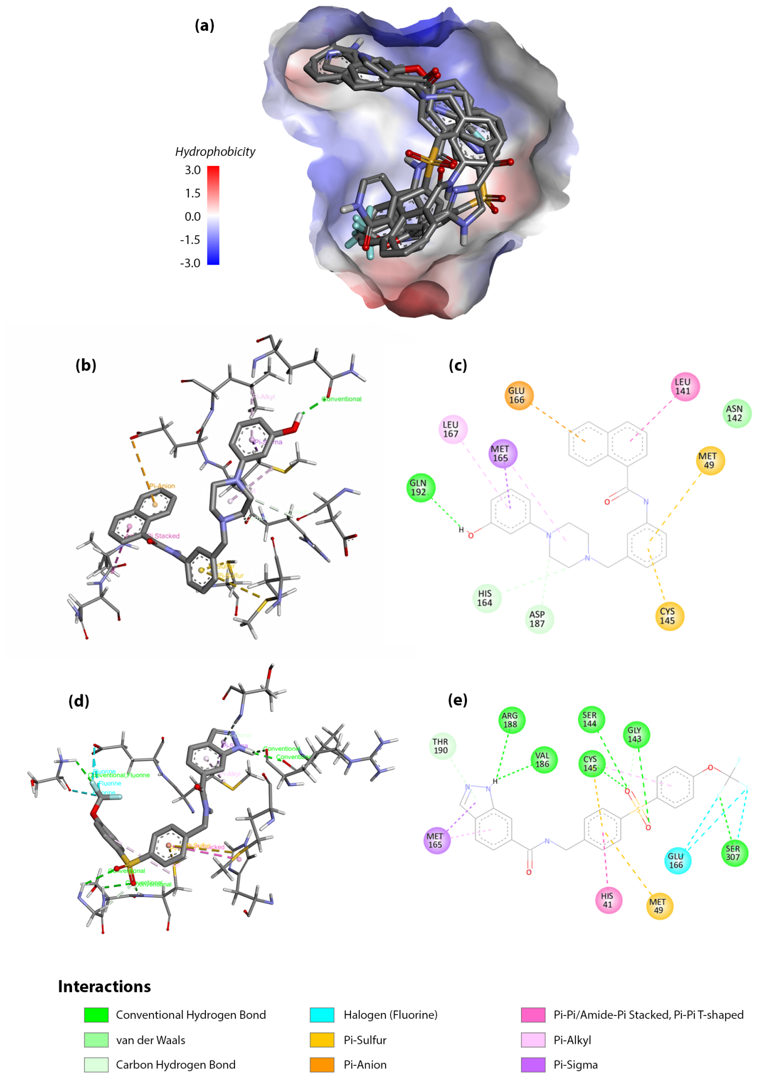


Figure 5.6: *Conformational and chemical compatibility of inhibitors predicted from iMiner for the MPro catalytic pocket.*(A) Randomly selected molecules from the final set of 51 inhibitors with their docking conformation determined by AutoDock Vina overlayed onto the surface of the binding pocket, with the surface color coded by hydrophobicity. Blue parts are hydrophilic and red parts are hydrophobic. (B) 3D interactions between molecule 1 and residues near the binding pocket (C) 2D illustrations for the interactions between molecule 1 and residues near the binding pocket (D) 3D interactions between molecule 2 and residues near the binding pocket (E) 2D illustrations for the interactions between molecule 2 and residues near the binding pocket. All figures generated by BIOVIA Discovery Studio Visualizer[39].

Furthermore, every aspect of this work is generalizable. There are many well defined proteins vital to the replication of SARS-CoV-2 with 3D structures available including the RNA-dependent, RNA polymerase protein (RdRp)[40], the Papain-like protease (PLpro)[41], and the exonuclease (ExoN)[42]. Although we have focused our current work on targeting SARS-CoV-2's Mpro, extension of this work to these other targets would be relatively trivial. Although identifying antiviral treatments for SARS-CoV-2 is of pressing concern at the time of this publication, our model could be quickly applied to design novel inhibitors for proteins relevant to other global diseases. For example, bacterial resistance to antibiotics is of preeminent concern in the medical community[43], and our iMiner workflow approach could be used to target novel bacterial biomolecules, such as bacterial Ribosomes, or target resistance conferring bacterial proteins such as β-lactamase.[43]

Overall, we believe our tool will be of great benefit to the computational and medicinal chemistry fields at large, and potentially aid traditional drug-design workflows as well. For example, molecules that are experimentally validated through a traditional HTVS approach as good binders could utilize the iMiner algorithm as an optimization or refinement step for elaborating on these existing leads or scaffolds. The potential of the method in this direction will be explored in future work.

5.5 Methods

Neural network architecture. The generative model employed in this study was an ASGD Weight-Dropped LSTM (AWD-LSTM)[20], which is a specific variant of the Long Short Term Memory (LSTM) recurrent neural network with shared DropConnect for weight regularization, and was trained through a non-monotonically triggered average stochastic gradient descent (NT-ASGD) algorithm.[20, 44] The basic LSTM cell contains two internal states, the hidden state h_t and the cell state c_t, and can be described through the following set of equations:

$$
\begin{aligned}
i_t &= \sigma(W^i x_t + U^i h_{t-1}) && (5.7)\\
f_t &= \sigma(W^f x_t + U^f h_{t-1}) && (5.8)\\
o_t &= \sigma(W^o x_t + U^o h_{t-1}) && (5.9)\\
\tilde{c}_t &= \tanh\left(W^c x_t + U^c h_{t-1}\right) && (5.10)\\
c_t &= i_t \odot \tilde{c}_t + f_t \odot c_{t-1} && (5.11)\\
h_t &= o_t \odot \tanh c_t && (5.12)
\end{aligned}
$$

where $[W^i, W^f, W^o, W^c, U^i, U^f, U^o, U^c]$ are the trainable parameters of the model, x_t is the input to the cell at the current timestep, $\tilde{c}_t$ contains the information to be added to the cell state, and i_t, f_t, o_t represent the update gate, forget gate and output gate respectively, which are numbers between $(0, 1)$ that controls how much information should be updated, discarded or retrieved from the cell state. σ denotes the sigmoid function, and $\odot$ represents element-wise multiplication. The DropConnect mechanism[45] was applied to the hidden-to-hidden weight matrices $[U^i, U^f, U^o, U^c]$ by randomly zeroing out a small portion of the parameters in these weight matrices to prevent overfitting and ensured that the same positions in the hidden vectors were treated consistently throughout the forward and backward pass in regards to whether or not to be dropped.

The inputs into the RNN cells were tokens embedded as vectors of length 400, and 3 LSTM cells were stacked sequentially, that had 1152, 1152 and 400 units each. The hidden state from the last timestep of the last RNN cell was then connected to a linear decoder with output size of 56 and softmax activation, representing the probabilities of the 56 possible tokens from the vocabulary. The dropout values used in the model were: embedding dropout=0.002, LSTM weight dropout=0.02, RNN hidden state dropout=0.015 and output dropout=0.01. The neural network was implemented using pyTorch[46] and the fastai package[47].

Supervised pretraining of the network The generative model was pretrained using molecules from ChEMBL 24[1], and a total of 1,440,263 molecules were selected for training. All molecules were first converted to SELFIES strings using the selfies python package[19], and the tokens were extracted from the SELFIES strings with fastai language model. We used categorical cross entropy loss:

$$L_\Theta = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t_i}\hat{p}(t_i|t_1, t_2, ..., t_{i-1})\log p_\Theta(t_i|t_1, t_2, ..., t_{i-1}) \tag{5.13}$$

where N represents the number of tokens in a molecule, $\hat{p}(t_i|t_1, t_2, ..., t_{i-1})$ represents the actual probability of a specific token in the string at position i and with all previous defined tokens t_1 through t_{i-1}, and $p_\Theta(t_i|t_1, t_2, ..., t_{i-1})$ the probability predicted by the neural network with parameters Θ. The model was trained using Adam optimizer[48] in batches of size 512, and we employed the "one cycle" learning rate policy[49] with the maximum learning rate of 0.0005 to achieve superconvergence[50]. During this pretraining stage we also used weight decay=0.01 and the dropout multiplier of 0.2. The model was pretrained for 30 epochs.

Reinforcement learning procedure. Our RL training target goal is to maximize $J(\Theta)$ from formula(5.6) by taking steps along $\partial_\Theta J(\Theta)$. The exact value for $J(\Theta)$ is intractable to evaluate, but can be approximated through sampling the distribution of s_T, which gives

$$J(\Theta) \approx \sum_{S_T} p_\Theta(s_T)r(s_T) \tag{5.14}$$

and then

$$\partial_\Theta J(\Theta) = \sum_{s_T} [\partial_\Theta p_\Theta(s_T)] r(s_T) \tag{5.15}$$

$$= \sum_{s_T} p_\Theta(s_T) [\sum_{t=1}^{T} \partial_\Theta \log p_\Theta(s_t|s_{t-1})] r(s_T) \tag{5.16}$$

Directly taking gradients according to (5.16) corresponds to the REINFORCE algorithm[51]. In this work we further utilized the PPO algorithm[32], which estimated the gradients through a clipped reward and with an extra entropy bonus term:

$$J'(\Theta) = \sum_{s_T} p_\Theta(s_T) [\sum_{t=1}^{T} J_t^{\text{CLIP}}(\Theta) + \alpha S[p_\Theta(s_t|s_{t-1})]] \tag{5.17}$$

where

$$J_t^{\text{CLIP}}(\Theta) = \min(R_t(\Theta) r(s_T), \operatorname{clip}(R_t(\Theta), 1-\epsilon, 1+\epsilon) r(s_T)) \tag{5.18}$$

with

$$R_t(\Theta) = \frac{p_\Theta(s_t|s_{t-1})}{p_{\Theta_{\text{old}}}(s_t|s_{t-1})} \tag{5.19}$$

denoting the ratio between the probability distribution in the current iteration and the probability distribution from the previous iteration (the iteration before last gradient update). A PPO algorithm reduces variance in the gradient, stabilizes training runs, and also encourages the model to explore a wider region of the chemical space through the introduction of an entropy bonus term. The two hyperparameters in the algorithm, α and ϵ, were taken as $\alpha = 0.02, \epsilon = 0.1$ in this work.

After the pretraining finished, we copied the weights to a separate model with identical architecture and trained with reinforcement learning using PPO. In each iteration 2000 molecules were sampled, and model weights were updated by taking gradient steps on the target function through formula (5.17), using a batch size of 1024 and Adam optimizer with fixed learning rate of 0.0001. In each iteration, all collected data were used for training the model for a maximum of 10 epochs. The trainer would continue into next iteration and collect new molecules for training if the K-L divergence between the latest predicted probability and the old probability exceeded 0.03.

The model was trained with RL for 400 iterations, until the mean entropy of the predicted probability of the tokens from the RNN started to decrease drastically. The change of mean entropy and mean vina score during the RL training can be found in Appendix Figure 5.C.4.

5.6 Acknowledgements

The authors thank Greg Merritt, Ian Lumb, and Mike Smith for their generous help in setting up the MicroSoft Azure cloud computing infrastructure and credits for the project. The authors thank the National Institutes of Health for support under Grant No 5U01GM121667. This work was also supported in part by the C3.ai Digital Transformation Institute. This research used Microsoft Azure through C3.ai DTI award, and also computational resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

5.7 Author Contributions

T.H-G. and J.L. conceived the scientific direction, designed the experiments, analyzed results, and wrote the manuscript. F.L.K. prepared Mpro structures for docking in Glide and AutoDock Vina, conducted Glide SP docking, and read and edited the manuscript. C.P. wrote the code for pre-training the network. O.Z. wrote the code for the AutoDock-vina workflow. N.G., I.L. and M.H. contributed ideas and to discussions to the work. R.E.A. provided drug-design and biophysical guidance, coordinated with experimental/medicinal chemists, read/revised the manuscript.

5.8 References

[1] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):D1100–D1107, 2012.

[2] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. Pubchem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 2021.

[3] Teague Sterling and John J. Irwin. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

[4] Pavel G Polishchuk, Timur I Madzhidov, and Alexandre Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27(8):675–679, 2013.

[5] Ingo Reulecke, Gudrun Lange, Jürgen Albrecht, Robert Klein, and Matthias Rarey. Towards an integrated description of hydrogen bonding and dehydration: decreasing

false positives in virtual screening with the hyde scoring function. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(6):885–897, 2008.

[6] Bryan C Duffy, Lei Zhu, Hélène Decornez, and Douglas B Kitchen. Early phase drug discovery: cheminformatics and computational techniques in identifying lead series. *Bioorganic & medicinal chemistry*, 20(18):5324–5342, 2012.

[7] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[8] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954. PMLR, 2017.

[9] Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.

[10] Akshay Subramanian, Utkarsh Saha, Tejasvini Sharma, Naveen K Tailor, and Soumitra Satapathi. Inverse design of potential singlet fission molecules using a transfer learning based approach. *arXiv preprint arXiv:2003.07666*, 2020.

[11] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of cheminformatics*, 9(1):1–14, 2017.

[12] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

[13] Sai Krishna Gottipati, Boris Sattarov, Sufeng Niu, Yashaswi Pathak, Haoran Wei, Shengchao Liu, Simon Blackburn, Karam Thomas, Connor Coley, Jian Tang, et al. Learning to navigate the synthetically accessible chemical space using reinforcement learning. In *International Conference on Machine Learning*, pages 3668–3679. PMLR, 2020.

[14] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[15] Alex Zhavoronkov, Vladimir Aladinskiy, Alexander Zhebrak, Bogdan Zagribelnyy, Victor Terentiev, Dmitry S Bezrukov, Daniil Polykovskiy, Rim Shayakhmetov, Andrey Filimonov, Philipp Orekhov, et al. Potential 2019-ncov 3c-like protease inhibitors designed using generative deep learning approaches. 2020.

[16] Navneet Bung, Sowmya R Krishnan, Gopalakrishnan Bulusu, and Arijit Roy. De novo design of new chemical entities for sars-cov-2 using artificial intelligence. *Future medicinal chemistry*, 13(06):575–585, 2021.

[17] Jannis Born, Matteo Manica, Joris Cadow, Greta Markert, Nil Adell Mill, Modestas Filipavicius, Nikita Janakarajan, Antonio Cardinale, Teodoro Laino, and María Rodríguez Martínez. Data-driven molecular design for discovery and synthesis of novel ligands: a case study on sars-cov-2. *Machine Learning: Science and Technology*, 2(2):025024, 2021.

[18] Woosung Jeon and Dongsup Kim. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports*, 10(1):1–11, 2020.

[19] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[20] S. Merity, N. S. Keskar, and R. Socher. Regularizing and optimizing lstm language models, 2018.

[21] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[22] Richard A. Friesner, Robert B. Murphy, Matthew P. Repasky, Leah L. Frye, Jeremy R. Greenwood, Thomas A. Halgren, Paul C. Sanschagrin, and Daniel T. Mainz. Extra precision glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *Journal of Medicinal Chemistry*, 49(21):6177–6196, 2006.

[23] Jayme L Dahlin, J Willem M Nissink, Jessica M Strasser, Subhashree Francis, LeeAnn Higgins, Hui Zhou, Zhiguo Zhang, and Michael A Walters. Pains in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging hts. *Journal of medicinal chemistry*, 58(5):2091–2113, 2015.

[24] Wen Cui, Kailin Yang, and Haitao Yang. Recent progress in the drug development targeting sars-cov-2 main protease as treatment for covid-19. *Frontiers in Molecular Biosciences*, 7:398, 2020.

[25] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, et al. Structure of m pro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, 2020.

[26] Hagit Achdout, Anthony Aimon, Elad Bar-David, Haim Barr, Amir Ben-Shmuel, James Bennett, Melissa L Bobby, Juliane Brun, Sarma BVNBS, Mark Calmiano, et al. Covid

moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *BioRxiv*, 2020.

[27] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[28] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.

[29] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

[30] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

[31] Ruth Brenk, Alessandro Schipani, Daniel James, Agata Krasowski, Ian Hugh Gilbert, Julie Frearson, and Paul Graham Wyatt. Lessons learnt from assembling screening libraries for drug discovery for neglected diseases. *ChemMedChem*, 3(3):435, 2008.

[32] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[33] Tom Everitt and Marcus Hutter. Avoiding wireheading with value reinforcement learning. In *International Conference on Artificial General Intelligence*, pages 12–22. Springer, 2016.

[34] Douglas R Houston and Malcolm D Walkinshaw. Consensus docking: improving the reliability of docking in a virtual screening context. *Journal of chemical information and modeling*, 53(2):384–390, 2013.

[35] Antoine Daina, Olivier Michielin, and Vincent Zoete. Swissadme: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific reports*, 7(1):1–13, 2017.

[36] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[37] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[38] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[39] DASSAULT SYSTÈMES. Biovia discovery studio, 2016.

[40] Jamshaid Ahmad, Saima Ikram, Fawad Ahmad, Irshad Ur Rehman, and Maryam Mushtaq. Sars-cov-2 rna dependent rna polymerase (rdrp)–a drug repurposing study. *Heliyon*, 6(7):e04502, 2020.

[41] Theresa Klemm, Gregor Ebert, Dale J Calleja, Cody C Allison, Lachlan W Richardson, Jonathan P Bernardini, Bernadine GC Lu, Nathan W Kuchel, Christoph Grohmann, Yuri Shibata, et al. Mechanism and inhibition of the papain-like protease, plpro, of sars-cov-2. *The EMBO journal*, 39(18):e106275, 2020.

[42] Nicholas H Moeller, Ke Shi, Özlem Demir, Surajit Banerjee, Lulu Yin, Christopher Belica, Cameron Durfee, Rommie E Amaro, and Hideki Aihara. Structure and dynamics of sars-cov-2 proofreading exoribonuclease exon. *bioRxiv*, 2021.

[43] C Lee Ventola. The antibiotic resistance crisis: part 1: causes and threats. *P & T : a peer-reviewed journal for formulary management*, 40(4):277–83, apr 2015.

[44] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

[45] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pages 1058–1066. PMLR, 2013.

[46] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[47] Jeremy Howard and Sylvain Gugger. Fastai: a layered api for deep learning. *Information*, 11(2):108, 2020.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[49] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[50] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.

[51] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[52] Terra Sztain, Rommie Amaro, and J. Andrew McCammon. Elucidation of cryptic and allosteric pockets within the sars-cov-2 main protease. *Journal of Chemical Information and Modeling*, 61(7):3495–3501, 2021. PMID: 33939913.

[53] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, Yinkai Duan, Jing Yu, Lin Wang, Kailin Yang, Fengjiang Liu, Rendi Jiang, Xinglou Yang, Tian You, Xiaoce Liu, Xiuna Yang, Fang Bai, Hong Liu, Xiang Liu, Luke W Guddat, Wenqing Xu, Gengfu Xiao, Chengfeng Qin, Zhengli Shi, Hualiang Jiang, Zihe Rao, and Haitao Yang. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, 2020.

[54] Rommie E Amaro and Adrian J Mulholland. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *Journal of Chemical Information and Modeling*, 60(6):2653–2656, jun 2020.

[55] Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, dec 2009.

[56] Oleg Trott and Arthur J Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, jan 2010.

[57] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, David E Shaw, Perry Francis, and Peter S Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, mar 2004.

[58] Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *Journal of Medicinal Chemistry*, 47(7):1750–1759, mar 2004.

[59] Schrödinger Release 2021-3. Glide, 2021.

[60] G Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, 2013.

[61] Schrödinger Release 2021-3. Protein Preparation Wizard, 2021.

[62] Chresten R. Søndergaard, Mats H. M. Olsson, MichaÅł Rostkowski, and Jan H. Jensen. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pka values. *Journal of Chemical Theory and Computation*, 7(7):2284–2295, 2011. PMID: 26606496.

[63] Mats H. M. Olsson, Chresten R. Søndergaard, Michał Rostkowski, and Jan H. Jensen. Propka3: Consistent treatment of internal and surface residues in empirical pka predictions. *Journal of Chemical Theory and Computation*, 7(2):525–537, 2011. PMID: 26596171.

[64] Chao Lu, Chuanjie Wu, Delaram Ghoreishi, Wei Chen, Lingle Wang, Wolfgang Damm, Gregory A Ross, Markus K Dahlgren, Ellery Russell, Christopher D Von Bargen, Robert Abel, Richard A Friesner, and Edward D Harder. OPLS4: Improving Force Field Accuracy on Challenging Regimes of Chemical Space. *Journal of Chemical Theory and Computation*, 17(7):4291–4300, jul 2021.

[65] William L Jorgensen, David S Maxwell, and Julian Tirado-Rives. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society*, 118(45):11225–11236, nov 1996.

[66] The Open Babel Package, 2016.

[67] Edward Harder, Wolfgang Damm, Jon Maple, Chuanjie Wu, Mark Reboul, Jin Yu Xiang, Lingle Wang, Dmitry Lupyan, Markus K Dahlgren, Jennifer L Knight, Joseph W Kaus, David S Cerutti, Goran Krilov, William L Jorgensen, Robert Abel, and Richard A Friesner. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. *Journal of Chemical Theory and Computation*, 12(1):281–296, jan 2016.

Appendix

5.A Methodology Details

Tokens in the generative model. Here we provide a complete list of tokens used in the generative model:

- Standard SELFIES tokens: [’#C’, ’#N’, ’#O’, ’#S’, ’=B’, ’=C’, ’=I’, ’=N’, ’=O’, ’=P’, ’=S’, ’=Se’, ’=Si’, ’B’, ’Br’, ’Br+2’, ’Branch1_1’, ’Branch1_2’, ’Branch1_3’, ’Branch2_1’, ’Branch2_2’, ’Branch2_3’, ’C’, ’Cl’, ’Cl+2’, ’Cl+3’, ’Expl=Ring1’, ’Expl=Ring2’, ’F’, ’I’, ’I+2’, ’I+3’, ’N’, ’O’, ’P’, ’Ring1’, ’Ring2’, ’S’, ’Se’, ’Si’]
- Modifier tokens: [“H+expl”, “H2+expl”,“H3+expl”,“+expl”,“Hexpl”, “H2expl”,“H-expl”, “H2-expl”,“H3-expl”,“-expl”,“expl”]
- Functional tokens: [“Break”]

When sampling molecules represented as SELFIES strings, the first token was always selected as the “Break” token. Then each token was sampled with probability distribution predicted by the generative model. Once the “Break” token was selected again, or the total number of tokens exceeded 500, a single molecule sampling process was considered complete. For each modifier token in the sampled string, it was combined into the previous token and was connected by the “ˆ” symbol. For example, ...[C][Hexpl]... would be converted to ...[CˆHexpl] to satisfy SELFIES syntax. If the token before a modifier token could not be modified, the sampled string would be considered invalid and was discarded.

Docking Preparation and Procedures. Stzain et al.[52] simulated SARS-CoV-2 Mpro (PDB ID 6LU7[53]) with Gaussian Accelerated Molecular Dynamics to characterize active site and dimer interface dynamics, as well as elucidate the presence of cryptic binding pockets. In total, Sztain et al. produced 6 microseconds of enhanced-sampled Mpro conformations.[52] These extensive simulations represent an invaluable resource for SARS-CoV-2 antiviral design, and as such Sztain et al. shared their trajectories publicly (https://amarolab.ucsd.edu/covid19.php) in accordance with the data sharing philosophy put forth by Amaro and Mulholland.[54] To ensure we were selecting biologically relevant Mpro conformations for use in our molecule generation workflow, we selected receptor structures from Sztain et al.’s simulations. Selection of each receptor structure and subsequent protein preparation steps are described below.

Mpro Active Site Receptor Selection and Preparation: To generate molecules targeting the Mpro active site, we selected the representative structure from the most populated cluster identified in Sztain et al.'s enhanced sampling trajectories of Mpro dimer,[52] simulated with a covalently bound inhibitor called N3. From Sztain et al.'s freely available files, the filename of the selected protease structure was "5.0_2.0_147.0_147.0_295.0_c0.pdb". We deleted the covalently bound N3 from this structure, taking care not to delete the catalytic Cysteine atoms (resids 145 and 451). We then modified the C145 and C451 atom names so that they reflected canonical Cysteine atom names. The Mpro structure was then prepared with AutoDockTools[55] and Schrödinger's Protein Preparation Wizard for docking in AutoDock Vina and Glide Ligand Docking, respectively (see Protein Preparation section below for more details). The cartesian coordinates for the active site center were found by calculating the center of mass of the C145 bound N3I covalent inhibitor before the inhibitor was deleted ([atomselect top "resname N3I and resid 145"]). This center of mass (x=54.58, y=45.92, z=75.06) was used to define the center of the active site during receptor grid generation steps in AutoDock Vina and Glide docking.

Scoring Generated Molecules with AutoDock Vina. AutoDockTools[55] was used to convert Mpro .pdb files to AutoDock Vina[56] compatible .pdbqt files. Additionally, AutoDockTools[55] was used to convert generated molecule structure files to AutoDock Vina compatible .pdbqt files. Gastieger charges were used for all AutoDock Vina structures. A cubic receptor grid of 30Å x 30Å x 30Å was centered around binding site's central coordinate (listed above), with a grid spacing of 1.0Å.

Re-scoring Generated Molecules with Glide Ligand Docking. As Schrödinger's Grid-Based Ligand Docking and Energetics (Glide) protocol[57, 58, 59] is one of the most well-trusted docking protocols available, we re-scored all our generated molecules in each Mpro binding site with Glide Standard Precision docking.[57, 58] To do so, we prepared each Mpro protein/receptor structure and all generated molecule structures for Glide docking. Schrödinger's Protein Preparation Wizard[60, 61] was used to prepare the Mpro receptor structures selected from Sztain et al.'s trajectories for Glide docking according to the following settings: Bond orders were calculated, missing hydrogens were added, and disulfide bonds were created all according to default options. Protein protonation states were assigned with PropKa around pH=7.0.[62, 63] A restrained minimization of all hydrogen atoms was then conducted according to the OPLS4 force field.[64]

Schrödinger's Receptor Grid Generation tool was used to prepare the Mpro structure for Glide docking according to the following settings: The center of the binding site was defined according to the center calculated above. The outer grid box size was set to 30Å x 30Å x 30Å, inner grid box size was set to 10Å x 10Å x 10Å. Grid points were placed every 1.0 Å. Receptor atom van der Waal radii were not scaled (i.e., scaled by a factor of 1.00) and the charge cut off for polarity was set to 0.25. Atom types were assigned according to OPLS 2005 atom types.[65]

To ensure we were utilizing identical molecules for comparison between AutoDock Vina results and Glide SP results, i.e. with respect to stereochemistry, we took output structures from AutoDock Vina (in .pdbqt format) and converted (with Open Babel[66]) first to .pdb files and then (with Open Babel) to .sdf files (SDF files being compatible for Schrödinger's LigPrep). Schrödinger's LigPrep module was then used to prepare all AutoDock Vina output structures for docking with Glide according to the following settings: Max allowed number of atoms per molecule was set to the default of 500. To again ensure that we docked structures identical to those docked with AutoDock, ionization states were not generated, tautomers were not generated, and chiral centers were not varied. Molecules were minimized according to the OPLS3 force field[67] and structures were written to .mae format for docking with Glide. Finally, all resulting structures were converted back to SMILES using openbabel [66] and then compared to the iMiner-generated SMILES strings to ensure consistency between structures proposed by the ML model and the actual structures docked by Glide. Any inconsistent structure was discarded in subsequent analysis.

Schrödinger's Glide Ligand Docking[57, 58, 22] module was used to re-score all generated molecules according to the Glide SP scoring function. The following settings were used during Glide SP ligand docking: Ligands were docked into each respective receptor according to a flexible ligand/rigid receptor docking protocol in which ligand bonds, angles and dihedral degrees of freedom were explored during docking. The top binding mode per molecule was saved and a Standard Precision Glide score was reported in kcal/mol for each molecule. The OPLS4 force field[64] was used for energetic evaluations and scoring. Glide SP scores were then compared, for each molecule, to AutoDock Vina scores.

All docking input files and protein structures will be shared in conjunction the data sharing philosophy put forth by Rommie E. Amaro and Adrian Mulholland.[54]

5.B Suporting Tables

Table 5.B.1: The 51 molecules from the final set

Index	Canonical SMILES	Vina score	Glide gscore	SA score
1	O=C(Nc1cccc(CN2CCN(c3cccc(O)c3)CC2)c1)c1cccc2ccccc12	-9.10	-8.07	2.85
2	O=C(NCc1ccc(S(=O)(=O)c2ccc(OC(F)(F)F)cc2)cc1)c1ccc2cn[nH]c2c1	-9.40	-8.14	2.72
3	O=C(Nc1ccnc2c(=O)[nH]ccc12)c1cc(Oc2ccc3ccccc3c2)ccc1F	-9.40	-8.02	2.88
4	O=C(Nc1cccc(C(=O)c2c[nH]c(-c3cccc4ccccc34)n2)c1)c1ccccc1	-9.10	-8.17	2.97
5	O=S(=O)(c1ccc(Cn2nc3ccccc3c2O)cc1)c1nc2ccccc2[nH]1	-9.10	-8.09	2.85
6	O=S(=O)(c1cccc(C(F)(F)F)c1)c1cc(-c2ccncc2)cc2cccnc12	-9.20	-8.20	2.99
7	C[C@H](Nc1ncnc2c(C(N)=O)cccc12)c1cccc(NC(=O)c2ccc(C#N)cc2)c1	-9.10	-8.47	3.34
8	O=C(c1cc(Cc2nnc(O)c3ccccc23)ccc1F)N1CCN(C(=O)C(F)(F)F)CC1	-9.30	-8.32	3.11
9	O=C(Cn1cc(-c2ccccc2)nn1)Nc1nc(-c2ccc(-n3cncn3)cc2)cs1	-9.10	-8.33	3.44
10	N#Cc1ccc(C(=O)N2CCN(C(=O)COC(=O)c3ccc(O)c(-c4ccccc4)c3)CC2)cc1	-9.40	-8.71	3.29
11	O=S(=O)(Nc1cccc(OCc2ccc3ccccc3n2)c1)c1ccc(C(F)(F)F)cc1	-9.10	-8.01	3.07
12	O=C(NCc1ccc(S(=O)(=O)c2cccc(N3CCOCC3)c2)cc1)c1ccc2ccccc2c1	-9.20	-8.24	3.21
13	O=C(N/N=C\c1ccc(OCc2ccc(-c3ccccc3)cc2)cc1)c1cccs1	-9.10	-8.86	3.34
14	O=C(c1cncc(-c2ccc(F)cc2)c1)N1CC(Oc2ccc3[nH]c4ccncc4c3c2)C1	-9.50	-8.04	3.31
15	Cc1cc(Nc2ncc(F)c(-c3cnc4ccccc4c3)n2)cc(-c2nnc[nH]2)c1F	-10.00	-8.09	3.12
16	O=C(Nc1nonc1NC(=O)N1CCN(c2ccc(F)cc2)CC1)c1ccc(F)cc1	-9.10	-8.10	3.33
17	O=S(=O)(Nc1cccc(-c2ccc(Nc3ccncc3)nn2)c1)c1cccc(C(F)(F)F)c1	-9.30	-8.25	3.31
18	O=S(=O)(Nc1ccc2[nH]nc(-c3ccc(F)cc3)c2c1)c1cccc(-c2ccc[nH]2)c1			

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
19	O=c1c2ccccc2n2c3c(cccc13)C(NS(=O)(=O)/C=C/c1ccccc1)=[SH]2	-9.30	-8.03	3.25
20	O=S(=O)(Nc1cccc(-c2ccc(Nc3ccccc3)nc2)c1)c1cccc(C(F)(F)F)c1	-9.20	-8.20	3.44
21	O=S(=O)(Nc1ccc(-c2cccc(-c3nnn[nH]3)c2)cc1)c1cccc(-c2cccc(O)c2)c1	-9.40	-8.54	3.35
22	O=C1Cc2ccc(S(=O)(=O)c3cn(-c4ccccc4)nc3-c3ccc(F)cc3F)cc2CN1	-9.60	-8.50	3.38
23	O=C1Cc2ccc(-c3cccc4cc(NS(=O)(=O)c5cccc(C(F)(F)F)c5)ccc34)cc2N1	-9.50	-8.29	3.45
24	C[C@H](Nc1ncnc2c(C(N)=O)cccc12)c1cccc(NC(=O)c2ccc(F)cc2)c1	-9.60	-8.02	3.03
25	O=S(=O)(Nc1ccc(F)c(-c2cnc3cc(F)ccn23)c1)c1cccc(-c2ccccc2)c1	-9.10	-8.18	3.22
26	O=S(=O)(Nc1cc(-c2ccc3nc[nH]c3c2)c2ccccn12)c1cccc2ccccc12	-9.80	-8.23	3.40
27	O=S(=O)(Nc1ncn[nH]1)c1ccc(Oc2cccc(-c3cc4ccccc4[nH]3)c2)cc1	-9.30	-8.05	3.25
28	O=S(=O)(Cc1ccccc1F)Nc1ccc2[nH]nc(-c3nc4ccccc4[nH]3)c2c1	-9.20	-8.46	3.29
29	O=S(=O)(Nc1ccc(S(=O)(=O)c2ccc(F)cc2)cc1)c1cc(-c2ccccc2)ncn1	-9.20	-8.43	3.06
30	O=S(=O)(c1ccc2[nH]c(Nc3ccccc3F)nc2c1)c1cccc2cccnc12	-9.60	-8.32	3.26
31	O=S(=O)(c1cccc(-c2cc3cc(O)ccc3o2)c1)c1nc2c(F)cccc2s1	-9.20	-8.02	3.10
32	O=S(=O)(c1ccccc1)N(Cc1ccc(Oc2ccc(-c3ncc[nH]3)cc2)cc1)Cc1cccnc1	-9.40	-8.46	3.49
33	O=S(=O)(c1cccc(/C=N\c2nc3ccccc3[nH]2)c1)c1cccc(-c2ccncc2)c1	-9.20	-8.36	3.32
34	O=S(=O)(Nc1ccc(S(=O)(=O)c2ccc(F)cc2)cc1)c1cc(-c2ccccc2)ncn1	-9.50	-8.07	3.33
35	O=S(=O)(Nc1cccc(/N=C/c2cccc(O)c2)c1)c1cccc(-c2ccccc2)c1	-9.60	-8.32	3.26
36	O=S(=O)(c1ccc(-c2ccccn2)cc1)N1CCc2cnc(Nc3ccccc3)nc2C1	-9.20	-8.16	3.38
37	O=S(=O)(Nc1cccc(-c2cc3ccccc3cn2)c1)c1cccc(-c2ccccc2)c1	-9.70	-8.16	3.44
38	O=S(=O)(c1cccc(-c2ccccc2)c1)c1ccc2[nH]c(Nc3cccc(O)c3)nc2c1	-9.90	-8.06	3.25

Table S1 (continued)

Index	Canonical SMILES	Vina score	Glide gscore	SA score
		-9.70	-8.16	3.21
39	O=S(=O)(c1ccc(-c2ccccc2)cc1)c1cccc(-c2cnc3cnccn23)c1			
		-9.20	-8.13	3.29
40	O=S(=O)(c1cccc(-n2cnc3cc(F)cnc32)c1)c1nc2ccc(-c3ccccn3)cc2s1			
		-9.30	-8.31	3.47
41	O=S(=O)(Nc1cccc2cc(-n3ccnc3O)ccc12)c1cccc(-c2ccccc2)c1			
		-9.60	-8.18	3.19
42	O=S(=O)(Oc1cccc(-c2cnc3ccccn23)c1)Oc1cc(-c2ccccc2)[nH]n1			
		-9.20	-8.16	3.46
43	O=S(=O)(Nc1cccc(-c2coc(-c3ccccc3)n2)c1)c1ccc2ccccc2c1			
		-9.70	-8.46	3.48
44	O=S(=O)(c1ccc(Nc2cc3ccncc3cn2)cc1)c1cccc(-c2ccccc2)c1			
		-9.20	-8.21	3.24
45	NS(=O)(=O)c1ccc2ccc(NS(=O)(=O)c3cccc(-c4ccccc4)c3)cc2c1			
		-9.10	-8.03	3.08
46	O=S(=O)(c1ccc(Cc2cnc3nc(F)ccc3c2O)cc1)c1ccc2ccccc2n1			
		-9.40	-8.03	3.09
47	O=S(=O)(Nc1cc2cncn(Cc3nnc[nH]3)c-2c1)c1cccc(-c2ccccc2)c1			
		-9.20	-8.22	3.23
48	O=S(=O)(Oc1cc(-c2cccc(N3CCNCC3)c2)ccc1F)c1ccc2cc[nH]c2c1			
		-9.30	-8.17	3.45
49	O=S(=O)(c1cccc(C2=CC(O)=NN(O)N2)c1)c1cccc(-c2cncnc2)c1			
		-9.10	-8.71	3.47
50	O=S(=O)(Nc1cc(-c2nc3ncccc3[nH]2)ccc1F)c1cccc(-c2cccnc2)c1			
		-9.60	-8.01	3.21
51	O=S(=O)(c1cccc(-c2ccccc2)c1)c1cccc(-c2cc(-c3ccncc3)cnc2O)c1			
		-10.00	-8.30	3.44

5.C Supporting Figures

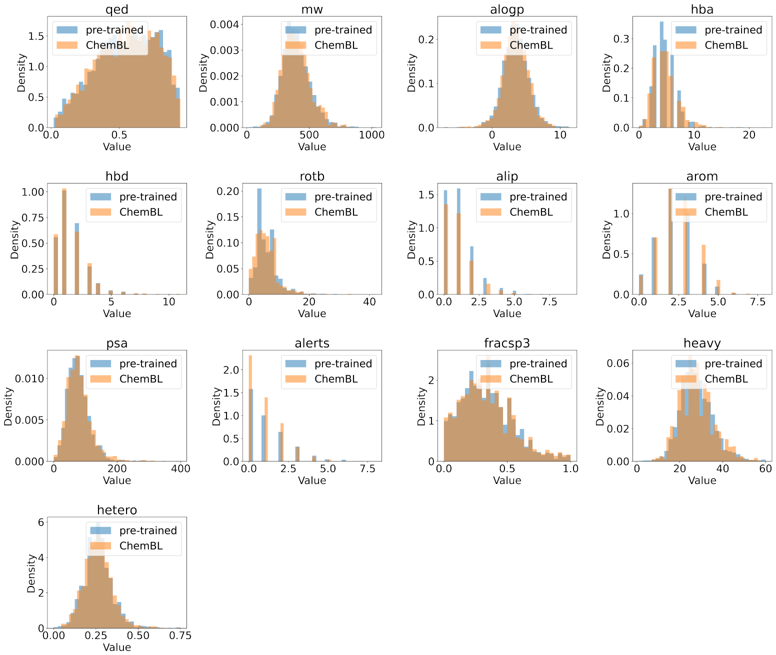


Figure 5.C.1: *Distribution comparisons for 13 different properties of the generated molecules from the pretrained model with molecules from the training dataset (ChEMBL).* The molecular properties considered are well-recognized chemical features related to the drug-likeliness of a molecule which can be obtained through 2D topological connectivity of the molecule: fraction of *sp*^{3} hybridized carbons(fracsp3), number of heavy atoms(heavy), fraction of non-carbon atoms in all heavy atoms(hetero), number of hydrogen bond donors(hbd) and acceptors(hba), number of rotatable bonds(rotb), number of aliphatic(alip) and aromatic rings(arom), molecular weight(mw), quantitative estimate of drug-likelihood (QED) value[29], approximate log partition coefficient between octanol and water (alogP)[30], polarizable surface area (PSA), and the number of structural alerts(alerts).[31]

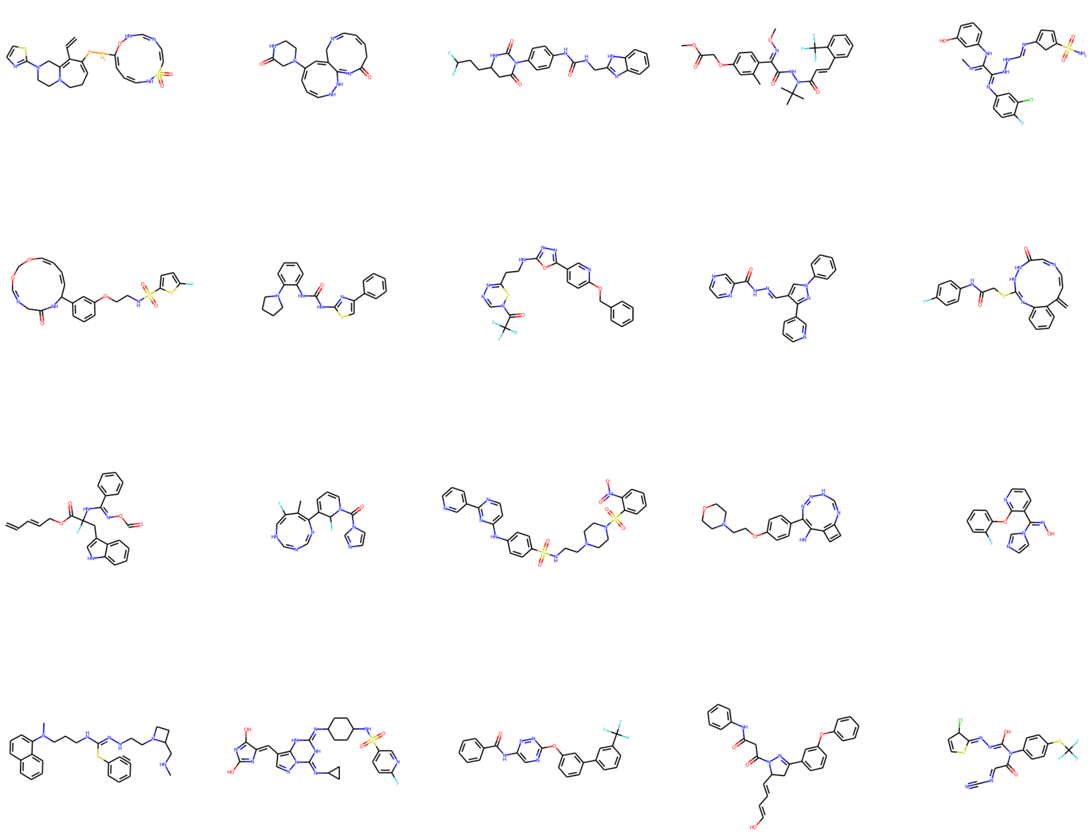


Figure 5.C.2: Example molecules generated using reinforcement learning without utilizing the drug-likeliness metric as additional reward. Many of these molecules are not drug-like, i.e. having large rings, or having a high proportion of hetero atoms.

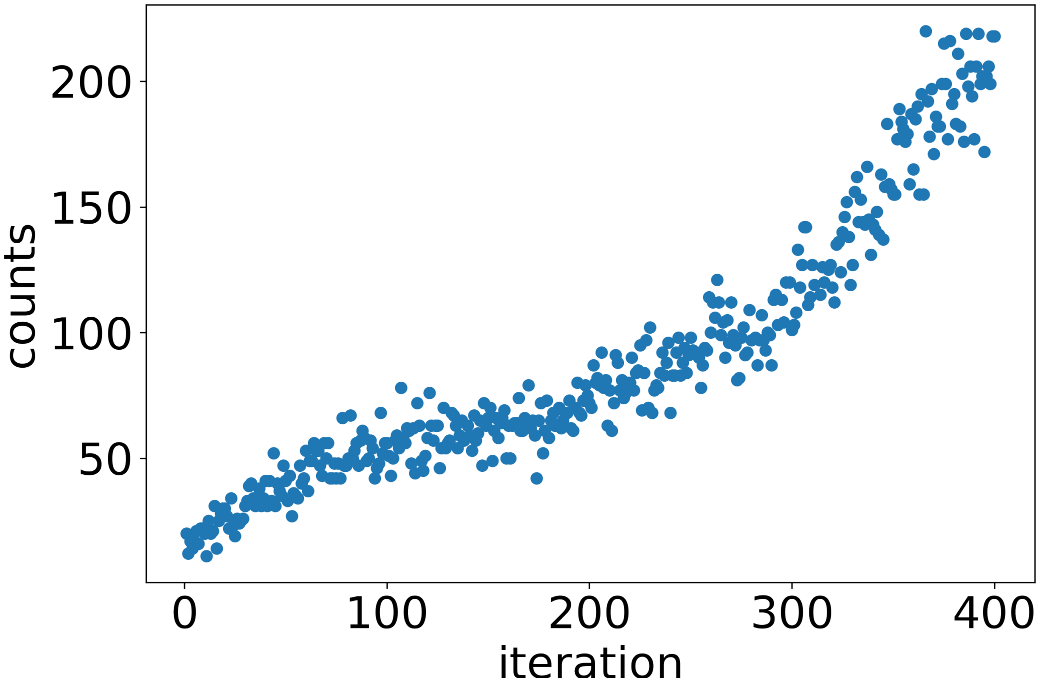


Figure 5.C.3: Number of molecules selected into the vina-selected set from each RL training iteration

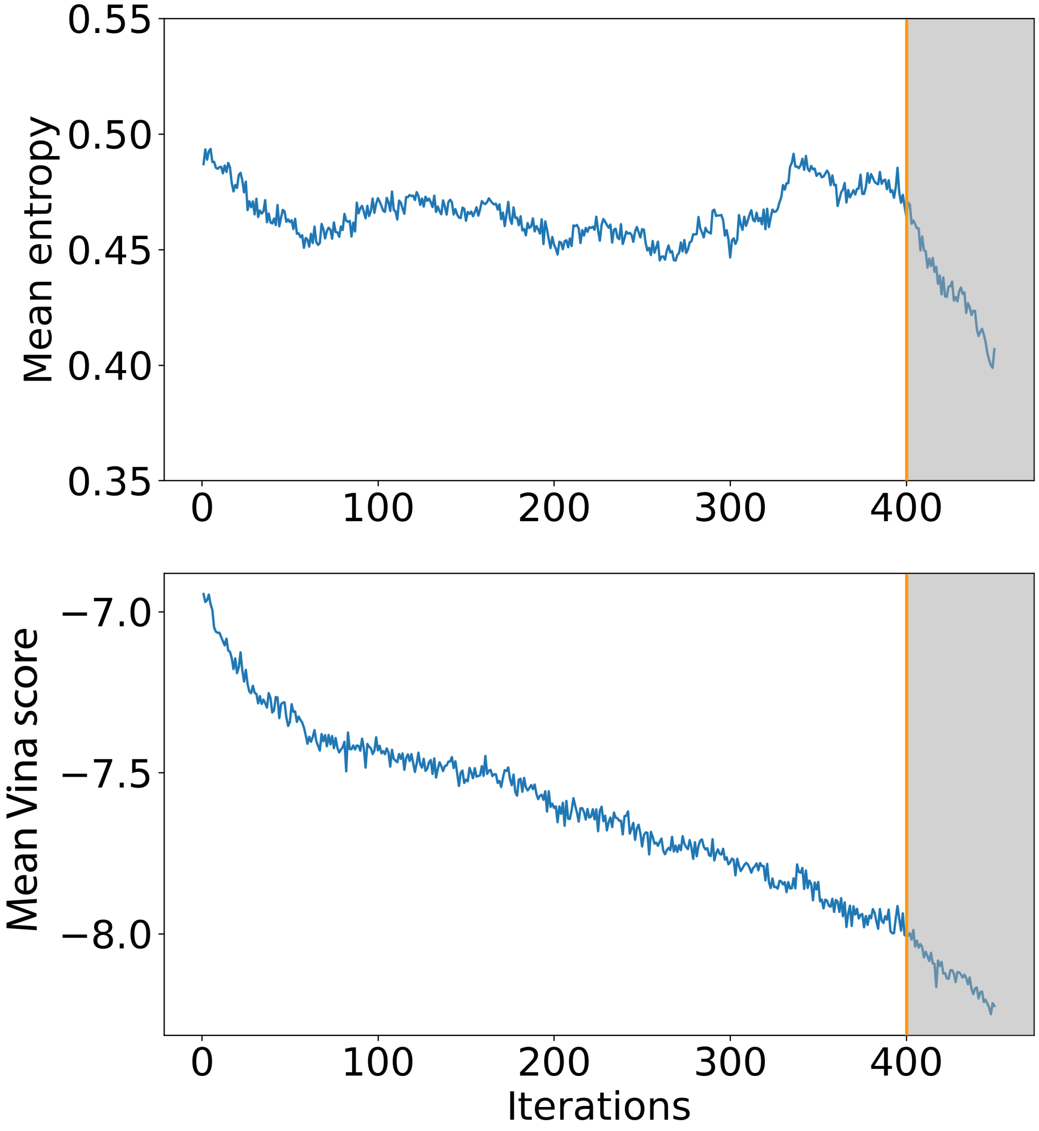


Figure 5.C.4: Change of mean entropy of model-predicted token probabilities and mean Vina scores of generated molecules during the training process. The model after RL is the model at iteration 400, and any molecules generated after iteration 400 are not considered for subsequent analysis.