Title

Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods

Journal

Chem, 6(7)

ISSN

1925-6981

Authors

Haghighatlari, Mojtaba
Li, Jie
Heidar-Zadeh, Farnaz
et al.

Publication Date

2020-07-01

DOI

10.1016/j.chempr.2020.05.014

Peer reviewed

HHS Public Access

Author manuscript

Chem. Author manuscript; available in PMC 2021 July 09.

Published in final edited form as:

Chem. 2020 July 9; 6(7): 1527–1542. doi:10.1016/j.chempr.2020.05.014.

Learning to Make Chemical Predictions: the Interplay of Feature Representation, Data, and Machine Learning Methods

Mojtaba Haghighatlari[1,*], **Jie Li**[1,*], **Farnaz Heidar-Zadeh**[1,2,3,*], **Yuchen Liu**[1], **Xingyi Guan**[1], **Teresa Head-Gordon**[1,4,5]

[1]Kenneth S. Pitzer Theory Center and Department of Chemistry, University of California, Berkeley, CA, USA

[2]Center for Molecular Modeling (CMM), Ghent University, B-9052 Ghent, Belgium

[3]Department of Chemistry, Queen's University, Kingston, Ontario K7L 3N6, Canada

[4]Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[5]Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, CA, USA

SUMMARY.

Recently supervised machine learning has been ascending in providing new predictive approaches for chemical, biological and materials sciences applications. In this Perspective we focus on the interplay of machine learning method with the chemically motivated descriptors and the size and type of data sets needed for molecular property prediction. Using Nuclear Magnetic Resonance chemical shift prediction as an example, we demonstrate that success is predicated on the choice of feature extracted or real-space representations of chemical structures, whether the molecular property data is abundant and/or experimentally or computationally derived, and how these together will influence the correct choice of popular machine learning methods drawn from deep learning, random forests, or kernel methods.

Graphical Abstract

corresponding author: thg@berkeley.edu.

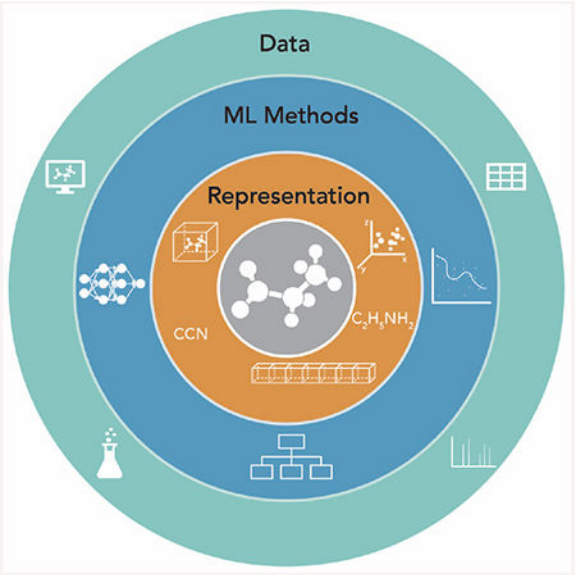
AUTHOR CONTRIBUTIONS

T.H-G., M.H., J.L, and F.H-Z conceived the scientific direction; T.H-G., M.H., J.L, and F.H-Z. wrote the manuscript; J.L. and Y.L. provided calculations; J.L., M.H., X.G., and Y.L. created the Figures. T.H-G., M.H., J.L., F.H-Z, and X.G. contributed insights and discussed and edited the manuscript.

*all authors contributed equally

DECLARATION OF INTERESTS

The authors declare no competing interests.



Successful machine learning in the chemical sciences relies on the interplay of three key components: chemical representation, machine learning method, and relevant data. Molecules can be represented as extracted descriptors (fingerprints, chemical identities) or direct representations (3D coordinates, electron densities), a choice that will depend on the machine learning approach such as deep learning, kernel methods, and statistical models, as well as on the type, quality, and abundance of training and testing data.

INTRODUCTION

The rise (again) of machine learning (ML) in the molecular sciences is a transformation of the traditional ways in which we perform computational chemistry. Unlike von Neumann machine algorithms, which articulate mathematical equations that can be solved in a logical progression, most machine *learning* is formulated as ”non-algorithmic” computing in applications where the complexity of the data or learning task makes the formulation of the sequence of symbolic functions impractical or impossible to define. In this case, machine learning is best applied when a symbolic algebra for chemical properties is difficult or impossible to solve, instead using (typically) supervised learning of well-curated data to map molecules to chemical properties. With appropriate strategies, ML has been successfully applied to quantum mechanically derived energy and force evaluation[1–3], molecular dynamics[4], three-dimensional structure prediction of small molecule crystals to large proteins[5–7], pathways for chemical reactivity and catalysis [8–10], and the rapid evaluation of spectroscopic and molecular properties[11–14].

ML has a long and storied history that builds on traditional mathematical programming, statistical and clustering models, and early meta-heuristic methods such as genetic algorithms and artificial neural networks (ANNs)[15]. Broadly speaking, the most popular machine learning approaches used in the chemical sciences today have evolved from these early efforts to now include non-parametric statistical learning such as decision trees and random forests, kernel-based models such as Gaussian Process regression (GPR) or Kernel

Ridge regression (KRR), and deep learning (DL) networks exemplified by convolutional neural networks (CNNs)[16].

Although machine learning methods were developed primarily by statisticians or computer scientists for other tasks such as image recognition[17], the chemical sciences domain has arguably advanced most effectively the development of novel feature representations, or *descriptors,* that informs the physical nature of the input-output mapping. These well designed descriptors offer many benefits including greater interpretability of the ML approach, to incorporate physical constraints on the learning parameters, or to better utilize a ML surrogate model for classification or regression.

But in order for ML methods and chemical descriptors to be effective requires the appropriate form and amount of the training data. If there is abundant training data which covers a wide scope of chemical space, it empowers DL networks with their (typically) huge number of parameters to discover complicated patterns in the data through successive transformations through their layers. For example, the popular CNNs have utilized widely available 3D representations in successful application to enzyme classification [18], molecular representations [19], and amino acid environment similarity analysis [20]. On the other hand, small datasets with a well formulated chemical representation can still be utilized by statistical or kernel models to make faithful predictions, such as predicting electronic structure correlation energies using sparse Hartreea Fock input [21]. Hence the choice of machine learning approach will be decided by whether the data stems from first-principles but limited in quantity due to expensive calculations from quantum mechanics (QM) or from abundant inexpensive calculations, or experimental data that may also be noisy, error prone, or difficult to interpret.

In this perspective, we first describe the three elements of successful prediction: ML methods, chemical feature representations, and dataset sizes and quality. We then illustrate their interplay for predicting nuclear magnetic resonance (NMR) chemical shifts, either through a combination of engineered features with random forest regression for protein NMR chemical shifts in solution[11] compared to shallow ANNs, while a deep learning CNN can improve performance over a KRR for chemical shift prediction in the solid state by exploiting physically motivated data augmentation[12]. Finally we conclude with an outlook for future directions of machine learning in the areas of feature representation development, data scarcity and sparsity, as well as physics-infused models and approaches to greater interpretability of machine learning.

THE COMPONENTS OF MACHINE LEARNING POPULAR ML METHODS IN THE CHEMICAL SCIENCES

Artificial neural network methods attempt to map the input-output relationship through a mathematical model which resembles the connections of neurons in a mammalian brain. In the chemical context, the input of a supervised machine learning model is a ”representation”, **x**, of a group of atoms that may form a drug molecule, a protein, a crystal structure, etc, and the output, **y**, is the chemical property of interest.

The most basic computing element of an ANN, the simple perceptron[22], is capable of performing linear or logistical regression and classification with appropriate activation functions (Figure 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.[23]

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 $\{\mathbf{x}\} \rightarrow f(y|\{x\})$ using a single layer rchitecture, nor what is the best feature representation of $\{x\}$ to ensure that it will perform well on previously unobserved target function data. Hence most of the recent excitement in machine learning is centered around 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. 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.

The most classical example of a DL architecture are the CNNs that were originally introduced and popularized by LeCun for handwriting and other image recognition tasks[17]. CNNs are neural networks that use convolution operations 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. Figure 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”[24]. 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.

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 better learning heuristics that are better established relative to early ANNs[16]: regularization through appropriate loss functions and back-propagation, 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[25] and Adam[26]; 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[16], 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.[27]

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[28]. More recently, KRR has realized excellent performance on regression prediction for molecular properties such as NMR chemical shifts for small molecules either in solution[29,30] or in the solid state[31]. In this case the physical understanding of a chemical system helped in the creation of a reasonable kernel function. Specifically the SOAP kernel[32] 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[32,33].

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[34].

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.[16] 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[35]. 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”[16]. 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[36], and are starting to replace the use of SVMs in classification tasks more broadly.

FEATURE REPRESENTATION

Similar to all modeling tasks, a representation or descriptor, is a mathematical abstraction of the inherent nature of the input, **x**, such as its chemical structure. Therefore, it is subject to the limitations of omitted features that may be influential for the property of interest. Thus, it is common practice to add more physical details into the representation such that they then correlate better/easier with target properties, **y**. In fact, research topics like quantitative structure property/activity relationships (QSPR/QSAR) have been popular and effective in the feature domain before modern machine learning has become more widespread. For ML, feature representations, when matched with the capabilities of the learning algorithms, are our most effective means to learn a chemical pattern/trend in data[37].

There are key criteria that we should consider for the construction of new descriptors:

- **uniqueness.**

  The representation should be unique with respect to the relative spatial arrangement of atoms. Often we need to develop descriptors that are invariant to the symmetries of the system (e.g., translation, rotation, atomic permutation, etc.), but are also distinctive for asymmetries (e.g., stereochemical chirality of molecules). Hence we prefer a one-to-one mapping not only for the easier training of ML models but also for a better generalizability and prediction performance.

- **universality.**

  The representation should be easily extendable to any system. If a descriptor is more representative to the fundamental chemical nature of the system, it also exhibits better transferability to new and future datasets. This is a key point for the accelerated exploration of molecular space, for example by means of virtual high-throughput screening[38].

- **efficiency.**

  The representation should be computationally efficient. The key advantage of any ML model to its computational or experimental alternatives is the efficiency. However, for some type of descriptors the cost of feature representation is narrowly comparable to the generation of reference (computational) data. For example, this is specifically the case for higher-order many-body interactions [39].

  Fulfilling all these criteria for the development of a desirable descriptor is a challenging task that necessitates expert knowledge of chemistry and computer science. In addition, the comparison of descriptors in terms of performance and efficiency is a nontrivial task, as it strongly depends on the data type and molecular diversity. Thus, for a given data set and choice of ML method, a fair comparison of feature representations also requires the same training setup in terms of training set size and sampling. The main reason is that if a data set is sparse and less representative of the entire molecular space, their feature representation is also limited to the available molecular makeup. Thereby, the resulting prediction performance is also restricted to the applicability domain of model that is imposed by training data.

  Considering a broad spectrum of representations used to build ML models [40], the required chemical information to encode molecular descriptors varies based on their availability and necessity for a given task. For example, inspired by QM we might consider atomic numbers, Z, and their chemical bonding sufficient to differentiate chemical systems from each other (2D descriptors). Moreover, if we aim to ultimately sidestep expensive QM calculations, we hope for the availability of atomic coordinates in order to correlate with rigorous electronic properties of the system (3D descriptors)[41]. Basic inputs with topological features of chemical structures such as type and size of ring or walk and path counts are also useful.

  The computational cost of obtaining the chemical information affects the overall efficiency of feature representation, and should be considered for their usage. For instance, the choice of 3D descriptors for training on QM computational data may require almost equally expensive geometry optimization for data generation. Thus, for future predictions the cost of preparing ML model inputs will be comparable to the reference QM calculations. However, if the atomic coordinates are available in advance, e.g., from experimental characterizations, or if the reference data is more demanding than geometry optimization (e.g., experimental data that is not easy to simulate such as melting point or solubility) the computational cost is often justifiable.

  In addition, physicochemical properties such as electronegativity, polarizability, and ionization potential has been commonly used in the drug discovery community. These types of data can be obtained using first principles or data mining, and has its roots in the

bioinformatics and cheminformatics domains. Descriptors based on such processed information are commonly referred to as hand-crafted descriptors or ”engineered” features[42].

The employed techniques for determining feature representation rely on different factors, including data type, ML approach, and of course the developer’s creativity. For example, one may consider a molecule as a weighted graph with features assigned to its nodes and edges, i.e., atomic features and bond features, and their consecutive interactions of atomic and bond features of their nearest neighbours. Thus, the overall representation is built using local atomic environments that rely on 2D chemical information. In 2015, Duvenaud and coworkers applied this idea in the form of graph convolutional networks (GCNs) to generalize the well established fingerprint algorithms that describe molecular makeups[43]. The hierarchical complexity of GCNs helped to extract from the topological combination of atomic and bond features an accurate explanation of a variety of chemical properties. Since then, a large number of published studies have reported successful improvements by tuning types of atomic/bond features and their interactions[13,44]. Several recent studies also consider non- bonded interactions (i.e., disconnected nodes) by accounting for interatomic distances as pairwise features[45,46] (see Figure 2). Alternatively, one may consider many-body interactions beyond only pairs of atoms and assign a unique functional form, e.g. symmetry functions, to represent the histogram of available interactions up to a certain degree[32,33.47]. Thus, similar to composing molecular descriptors from atomic and pairwise features, they decompose many-body interactions and build a descriptor that relies on all terms individually and simultaneously.

More recent attention has focused on the provision of 3D structures with minimum information loss. The idea is to represent molecules to the ML model in the same way that they are visualized [12,18], e.g., using a set of atomic densities. This type of representation has similarities with both QM, i.e., by providing electron density distribution of atoms, and computer vision, i.e., by replicating human vision using the complete configuration of the elements of a system [48].

Due to the flexibility in design and hierarchical manipulation of latent feature space, neural networks have become the cornerstone of creative ideas to integrate chemical information with the ML workflow. The results of such efforts has created a new branch of feature representation that is commonly known as learned features. Later in this paper, we present notable examples from our lab of employing engineered and learned features in the course of molecular property prediction.

TYPES OF DATA AND THEIR ABUNDANCE

The quality of labelled chemical datasets, composed of (**x**, **y**) pairs, is one of the key components in developing an accurate and predictive ML model. Even though generating systematic and exhaustive datasets which samples the chemical space computationally has arrived recently[49], experimental datasets are indispensable because some properties are either difficult or impossible to compute. In developing ML models, one needs to be aware

of the inherent differences between computational and experimental data, and take them into account when designing suitable representation for a given target property.

The reliability, accuracy, and reproducibility of computational data is improved by applying a concrete computational protocol across the dataset and carefully choosing and reporting its parameters, like level of theory, basis set, convergence criteria, and number of grid points. Even though similar standards can be applied in generating experimental data, the nature of experimental protocol or experimental conditions (e.g., solvent, temperature, pH) is most likely different as the data is commonly compiled from various sources. This leads to an inherent inconsistency in data compounded by different measurement errors in different experiments. For example, Nuclear magnetic resonance (NMR) chemical shift prediction utilizes X-ray crystal structures and solution NMR measurements to define the (**x**, **y**) labelled pairs, although their correspondence is not one-to-one.

The comprehensiveness of computational data is systematically improvable by continued enumeration of chemical compounds and their properties. In contrast, experimental measurements are timeconsuming and resource-intensive, and adding additional data points to an experimental set is difficult, thus they sample chemical space more sparsely. This has led to combining experimental and computational data in some cases[50]. To further capture the inherent complexity of experimental data, their feature representation can be augmented with environmental conditions (like temperature, pH, and solvent). For example, hydrogen-bonding environments from crystal waters in the X-ray structure were also included in the prediction of a chemical shift of atoms in proteins to account for solvent effects[11]. Data augmentation from computation can be designed to incorporate ensemble averaging of experimental structures, such as introducing backbone flexibility commensurate with X-ray diffraction [51] and/or side chain repacking that reproduces NMR J-couplings[52] for proteins. Alternatively, one can include multiple input representations to the same property value which also increases the size of the dataset. Typically these augmentation approaches seek the sweet spot of low computational cost and high chemical/structural diversity to achieve the desirable experimental prediction accuracy.

INTERPLAY OF REPRESENTATION, DATA, AND MACHINE LEARNING ALGORITHM TO PREDICT CHEMICAL PROPERTIES

NMR spectroscopy is one of the most important molecular probes of chemical composition, structure and dynamics of small molecules through to large proteins. The least invasive techniques of NMR are the chemical shifts and spin-spin splittings which can be measured to very high accuracy. Because they are sensitive to their functional groups, detailed geometries, and chemical environments, they allow for prediction of solution phase protein structures or to identify or verify the structure of chemical compounds in the crystalline phase.[53]

The connection between NMR chemical shifts to structural or dynamical properties, while true in principle, is nevertheless sometimes difficult to reveal in practice through direct assignment of the spectrum. One solution to this problem is to rely on expensive QM methods that often can accurately predict spectral observables from structure of small molecular fragments[54]. While chemical physics approaches have achieved considerable

success in spectral assignment and structure determination, here we consider two recent examples of supervised learning approaches where the interplay of chemical descriptors, data size and augmentation strategies, and choice of ML algorithm has significantly improved the accuracy of chemical shift predictions and their connections to complex structure in aqueous solution and in the solid state.

ENGINEERED FEATURES AND RANDOM FOREST REGRESSION TO PREDICT CHEMICAL SHIFTS FOR AQUEOUS PROTEINS

Given the expense of QM calculation for magnetic properties, heuristic NMR ”calculators” have been developed for efficient chemical shift evaluations for aqueous proteins. In particular, the single-layer feed-forward network developed and packaged as SPARTA+[55] remains among the most popular of chemical shift prediction methods. Better predictive power can also be gained by exploiting sequence homology as that used in SHIFTX2[56], as the expectation is that as more sequence and spectroscopic data is deposited in public repositories, it will allow interpolation to replace extrapolation for a variety of NMR observables.

Even with these successes, these algorithms are still open to change as modern day ML approaches march forward alongside accumulating biological data. Furthermore, engineered features are ideal for predicting experimental chemical shifts for proteins in solution because they are not overly sensitive to different instantaneous conformations in the thermalized ensemble while still differentiating between atomic environments of aqueous proteins that exhibit different chemical shifts. Specifically, classification features, like whether an atom is involved in a hydrogen bond or a residues secondary structure category, are relatively stable for different conformations in the ensemble relative to the coordinates of the atoms in 3D space, while still being distinct enough for different residues, secondary structure, or proteins being predicted.

Recently engineered features extracted from protein X-ray crystal structures has been utilized together with random forest regression to formulate the UCBShift chemical shift predictor for aqueous proteins.[11] All backbone atoms and the side chain β-carbon chemical shifts of a residue are mapped from numerical and non-numerical features built from the geometries and biophysical properties of a tripeptide centered at the target residue. The features were designed with uniqueness, universality and efficiency in mind, which include backbone and side chain torsion angles, BLOSUM numbers identifying the likelihood of residue substitution, secondary structure, hydrogen bond geometries, ring currents, half surface exposure, accessible surface area, and non-linear transformations of distance features which have physical relevance from QM. All of these features are formulated as internal properties of the protein which naturally exhibit translational and rotational invariance, hence being unique to the structure itself while universal for the global frame. All these features could be efficiently calculated from a given protein structure within seconds.

However, the universality of the representation is limited to proteins without functional group modifications or bonding with ions, ligands or other hydrogen-bonding motifs with water. To increase the applicability of our ML model, we have also included extraction of

crystal water positions in the evaluation of features such as hydrogen bonding, and alignment scores that characterize sequence and structural homology to other proteins with recorded chemical shifts, aiding the chemical shift prediction through learned direct transfer if the similarity is faithful enough to the query protein.[11]

The UCBShift algorithm utilizes two successive decision tree ensemble models (Figure 3a), one which differentiates the various atomic environments in a protein utilizing engineered features, and a second that make predictions based on the most similar sequence and/or structural alignments in the training dataset. As a result, UCBShift has significantly lower root-mean-square-error (RMSE) when applied to an independently generated test dataset when compared to SPARTA+ and SHIFTX2 on all the relevant protein atom types (Figure 3b). Further analysis of the total number of decisions made in each tree, which is visualized in Figure 3c, reveals that the QM-inspired transformations of the features account for more than 20% of the feature importance.[11] Even though some features like half surface exposure seem to play a more limited role in prediction, their existence extends the model's capacity in recognizing some atomic environments which might be differentiated by this feature, therefore making the representation more unique.

SPARTA+[55] and SHIFTX2[56], which are based on simpler machine learning models, as well as our own attempts with deep recurrent neural networks with residual connections, have not performed as well as the random forest model presented here. This is because simple MLs do not have sufficient capacity to recognize the complexity of the mapping from engineered features to chemical shifts, and the limited number of well-formulated structure-chemical shift pair in the dataset prevents those more complicated deep neural networks to effectively train. This consequence once again reinforces that choice of the ML method, together with the appropriate representation, need to be regulated by the size and intrinsic structure of the dataset in order to achieve excellent predictive power for solution- phase NMR properties. Future improvements are still possible once more data becomes available so that features could be learned directly from a deep learning setup.

CHEMICAL SHIFT PREDICTION IN THE SOLID STATE FROM LEARNED FEATURES USING DL AND DATA AUGMENTATION

Crystal structures of small molecules can be identified by comparing the experimental measurements of solid-state NMR chemical shifts with the calculated results using DFT, typically using the Gauge- Including Projector-Augmented Waves (GIPAW) method [54]. However, because of the cubic scaling with the size of the atomic basis sets used in the DFT calculation, ML methods have been investigated to approximate the QM physics. For example, a shallow ANN using engineered features was used to predict chemical shifts (and quadrupolar couplings) in silica materials using symmetry functions operating on the Cartesian coordinates to respect rotational invariance of the chemical shift value to applied magnetic field[57]. Paluzzo et al. devised a ML approach using 3D structures, while also directly incorporating rotational symmetry using KRR and the SOAP kernel[32], yielding very good results for chemical shift prediction for small molecule crystal systems[31]. Even though a significant acceleration factor was achieved over QM using these ML approaches, the training data generation using DFT is itself a bottleneck, thereby making a shallow ANN

necessary, while the quadratic-to-cubic complexity for calculating and inverting the kernel matrix makes it also impractical for KRR to treat larger datasets, although a number of strategies including parallelization can mitigate the cost[34].

The question we set out to address was whether a deep learning approach was tenable for the prediction of DFT chemical shifts for hydrogen (^{1}H), carbon (^{13}C), nitrogen (^{15}N) and oxygen (^{17}O) of organic molecules in molecular crystals. The input representation was comprised of the 3D coordinates of atoms in the unit cell taken from the Cambridge Structural Database (CSD), ”imagery” that was ideally suited to a multi-resolution CNN based on a DenseNet approach as shown in Figure 4a. This way of presenting molecules is similar to the sum of Gaussians representation of Bartok et al. [58] or the use of atom-centered wavelets as used by Eickenberg and co-workers[59].

We utilized the chemical environment for each atom whose chemical shift is predicted is represented on a 3D grid with a calculated Gaussian density at each atom center. This input representation describes local bonding characteristics that arrange atoms into 3D shapes with more global spatial organization. CNNs are ideally suited to the 3D structural data and electron density representation. This is because the network architecture of a CNN was originally formulated to operate on data that has temporal organization, i.e. 2D images arranged in a time series, but for which the time axis can be replaced by a 3rd spatial dimension to represent the electron density distribution. Hence we benefited from the open access to the original DFT chemical shifts calculated on 2000 organic molecules containing ~30 – 40 atoms to create the labelled data.[31]

Furthermore, better data representations and data quantity proved crucial to the success of our DL approach. First, we showed that the chemical environment for each atom type could be represented by multiple resolutions (MR), thereby incorporating the atomic densities of the other atoms over different grid sizes of d (4Å, 6Å, 8Å, 10Å, and 14Å) with 16×16×16 voxels, and representing each resolution with its own dedicated channel. Under each resolution, we divided the density based on the atom types into 4 different channels for ^{1}H, ^{13}C, ^{15}N, ^{17}O, respectively, similar to RGB channels used in image recognition. Second, given the limited number of examples in the training dataset, and the prohibitive expense of creating an order of magnitude more data, we recognized that a cheap data augmentation method was obviously available. Instead of enforcing chemical shift invariance through explicit rotational symmetry operations, we instead just augmented the data by rotating the Cartesian coordinates of atoms randomly with the Euler angles uniformly distributed between $[-\frac{\pi}{2},\frac{\pi}{2}]$ along each of x, y and z axis. During the training phase, both the original data and augmented data are included in the training dataset, while during the testing phase we average the prediction results over the different rotation configurations, thereby manifesting ensemble learning.[16]

Figure 4b summarizes the results for the MR-3D-DenseNet workflow in terms of the root mean square error (RMSE) for chemical shift prediction of the DFT results for each atom type. Using the greater capacity of the MR-3D-DenseNet deep network, we obtain nearly 15% improvement for ^{17}O and close to 25% for ^{13}C, ^{15}N, and ^{1}H chemical shifts over

KRR, for which hydrogen chemical shifts are similar in error between *ab initio* calculations and experimental measurements.

CONCLUSIONS AND OUTLOOK

We have shown that three key components of a machine learning workflow - algorithm, features, and data - are inexorably intertwined for achieving predictive success for biological, chemical, and materials applications. Inconsistent decisions about choice of ML methods and feature representation with respect to size and source of data can lead to inaccurate deployment of techniques in molecular property prediction. We illustrated that proper execution of the ML triad can lead to successful prediction of NMR chemical shifts of molecules in solution or for crystalline states. When a molecular property is symmetry-invariant we can enforce symmetry operations in the ML method or through data augmentation that permits one to exploit a deep learning strategy. Data augmentation was important in the solid state NMR chemical shift deep learning model, since generating new QM data is computationally expensive. By contrast while chemical shifts are very sensitive to 3D representations, the predicted solution NMR data is noisy and highly averaged and feature extracted data can be beneficial. We showed this on the example of aqueous protein chemical shift prediction, in which the choice of expertly-crafted features can facilitate the learning task in a direct and concise form to avoid redundancy and curse of dimensionality.

Not surprisingly, these three key components are also at the heart of current developments in ML, and many open questions and challenges need to be addressed to push the boundaries toward new applications. In terms of feature representation, systematic development of new descriptors, standardization of their evaluation, and easier accessibility via user interfaces (e.g., Python libraries) are necessary to establish their long-term development [60]. A transparent and sustained study of feature representation would involve researchers with variety of domain knowledge and expertise to accelerate future developments.

In addition, technical challenges involved with scarce and sparse data sets need to be vigorously discussed, as it is often a prevalent case for applying ML to real-world chemical applications when expensive calculations or difficult experiments are the bottleneck. The good news is that the entire ML community has been giving more attention to this issue, and as a result, techniques that can deal with limited data have been growing:

- methods that are intrinsically suitable for small size data, such as kernel methods or low-variance models with feature reduction capabilities,
- methods that leverage small size data in the learning task, for example by transfer learning [61] from pre-trained or high-fidelity models, or in some cases by multitask learning,
- methods for data curation, such as learning to impute missing data[10] or representation goals can be achieved with data augmentation when it is not practical to incorporate symmetries explicitly in the feature representation[12]
- decreasing the number of data generation trials via sampling methods using an active learning (AL) approach [60,62]; AL methods can discover the uncertainty of

trained models in the high-dimensional data distribution and query more informative training data that improves the model most

- future work would be to investigate imbalanced data and underrepresented regions of the solution space studied in the form of unsupervised ML techniques, such as clustering methods.

Moreover, future research needs to examine more closely the *interpretability* of chemical ML models. In this regard, gaining chemical insights and understanding direct relationships between molecular observables and their properties is the ultimate goal. The issue with model interpretability is that ML methods are designed to learn patterns (or mappings) in high-dimensional data that is otherwise not obvious to ourselves. Thus, predictive models are generally perceived as a so-called black-box model with limited transparency[63]. However, we argue that this would not be the case if we work with hand-crafted features and the simplest possible ML model. Hence part of the current concern regarding interpretability of ML models stems from highly parameterized models with arbitrary choice of hidden states. Thus, a general practice for the future work is to simplify state-of-the-art models and evaluate model shrinkage as part of a cost-benefit analysis. For instance, a learning curve based on the number of trainable parameters (or number and type of layers) should become a trend in applications of deep learning models.

On the other hand, ML complexity is still sometimes needed, and the interpretability of models requires greater contributions from expertise equipped with domain knowledge. For example relevance propagation techniques[64] can help interpret a trained ML model. Recent efforts on developing visualization tools can also help to monitor the change/gain by adding extra layers to deep learning models.

Often there is a fear that a new approach such as ML could supplant existing computational chemistry techniques or suppress models designed for physical insight. This will never be the case. Given the considerable progress in algorithms, molecular feature representation, and data accessibility, we expect interest in applying ML to almost any vein of chemical and materials sciences will continue to grow and ultimately settle in as a long term player in the general computational chemistry landscape.

ACKNOWLEDGMENTS

The authors thank the National Institutes of Health for support under Grant No 5U01GM121667. FHZ thanks the Research Foundation-Flanders (FWO) Postdoctoral Fellowship for support of this work. 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.

REFERENCES

1. Chmiela Stefan, Sauceda Huziel E., Poltavsky Igor, Muller Klaus Robert, and Tkatchenko Alexandre. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. Comput. Phys. Commun, 240:38–45, 2019.
2. Smith Justin S., Nebgen Benjamin T., Zubatyuk Roman, Lubbers Nicholas, Devereux Christian, Barros Kipton, Tretiak Sergei, Isayev Olexandr, and Roitberg Adrian E.. Approaching coupled

cluster accuracy with a general-purpose neural network potential through transfer learning. Nat. Commun, 10(1): 1–8, 2019. [PubMed: 30602773]

3. Amabilino Silvia, Bratholm Lars A., Bennie Simon J., Vaucher Alain C., Reiher Markus, and Glowacki David R.. Training Neural Nets to Learn Reactive Potential Energy Surfaces Using Interactive Quantum Chemistry in Virtual Reality. J. Phys. Chem. A, 123(20):4486–4499, 2019. [PubMed: 30892040]
4. Y Wang R J M Lamim Ribeiro, and P Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. Curr. Opin. Struct. Biol, 61:139–145, 2020. [PubMed: 31972477]
5. Sanchez-Lengeling Benjamin and Aspuru-Guzik Alan. Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400):360–365, 2018. [PubMed: 30049875]
6. Senior Andrew W, Evans Richard, Jumper John, Kirkpatrick James, Sifre Laurent, Green Tim, Qin Chongli, Zidek Augustin, Nelson Alexander W R, Bridgland Alex, and Others. Improved protein structure prediction using potentials from deep learning. Nature, pages 1–5, 2020.
7. AlQuraishi Mohammed. End-to-end differentiable learning of protein structure. Cell Syst., 8(4):292–301, 2019. [PubMed: 31005579]
8. Brickel Sebastian, Das Akshaya K, Unke Oliver T, Turan Haydar T, and Meuwly Markus. Reactive molecular dynamics for the [C1Ã¢Â Â CH 3 Ã¢Â Â Br] Ã¢Â Â reaction in the gas phase and in solution: a comparative study using empirical and neural network force fields . Electron. Struct, 1(2):24002, 2019.
9. Shakouri Khosrow, Behler Jörg, Meyer Jörg, and Kroes Geert Jan. Accurate Neural Network Description of Surface Phonons in Reactive Gas-Surface Dynamics: N2 + Ru(0001). J. Phys. Chem. Lett, 8(10):2131–2136, 2017. [PubMed: 28441867]
10. Schwaller Philippe, Laino Teodoro, Gaudin Theophile, Bolgar Peter, Hunter Christopher A, Bekas Costas, and Lee Alpha A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. ACS Cent. Sci, 5(9): 1572–1583, 2019. [PubMed: 31572784]
11. Li Jie, Bennett Kochise C, Liu Yuchen, Martin Michael V, and Head-Gordon Teresa. Accurate prediction of chemical shifts for aqueous protein structure on Ã¢Â Â ÄúReal WorldÃ¢Â Â Äù data. Chem. Sci, 11(12):3180–3191, 2020.
12. Liu Shuai, Li Jie, Bennett Kochise C., Ganoe Brad, Stauch Tim, Head-Gordon Martin, Hexemer Alexander, Ushizima Daniela, and Head-Gordon Teresa. Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. J. Phys. Chem. Lett, 10:4558–4565, 2019. [PubMed: 31305081]
13. Yang Kevin, Swanson Kyle, Jin Wengong, Coley Connor, Eiden Philipp, Gao Hua, Guzman- Perez Angel, Hopper Timothy, Kelley Brian, Mathea Miriam, Palmer Andrew, Settels Volker, Jaakkola Tommi, Jensen Klavs, and Barzilay Regina. Analyzing Learned Molecular Representations for Property Prediction. J. Chem. Inf. Model, 59(8):3370–3388, 2019. [PubMed: 31361484]
14. Haghighatlari Mojtaba, Vishwakarma Gaurav, Afzal Mohammad Atif Faiz, and Hachmann Johannes. A Physics-Infused Deep Learning Model for the Prediction of Refractive Indices and Its Use for the Large-Scale Screening of Organic Compound Space. ChemRxiv, pages 1–9, 2019a. doi: 10.26434/chemrxiv.8796950.v1.
15. Russell Stuart and Norvig Peter. Artificial Intelligence: A Modern Approach. Prentice Hall, 3 edition, 2010.
16. Goodfellow Ian, Bengio Yoshua, and Courville Aaron. Deep Learning. MIT Press, 2016.
17. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, and Jackel LD. Backpropagation Applied to Handwritten Zip Code Recognition. Neural Comput., 1(4):541–551, 12 1989.
18. Amidi Afshine, Amidi Shervine, Vlachakis Dimitrios, Megalooikonomou Vasileios, Paragios Nikos, and Zacharaki Evangelia I. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. PeerJ, 6:e4750, 2018. [PubMed: 29740518]
19. Kuzminykh Denis, Polykovskiy Daniil, Kadurin Artur, Zhebrak Alexander, Baskov Ivan, Nikolenko Sergey, Shayakhmetov Rim, and Zhavoronkov Alex. 3D Molecular Representations

Based on the Wave Transform for Convolutional Neural Networks.Mol. Pharm, 15(10):4378–4385, 2018. [PubMed: 29473756]

20. Torng Wen and Altman Russ B. 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics, 18(1):302, 2017. [PubMed: 28615003]
21. Welborn Matthew, Cheng Lixue, and Miller Thomas F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. J. Chem. Theory Comput, 14(9):4772–4779, 2018. [PubMed: 30040892]
22. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev, 65(6):386–408, 1958. [PubMed: 13602029]
23. Taskinen Jyrki and Yliruusi Jouko. Prediction of physicochemical properties based on neural network modelling. Adv. Drug Deliv. Rev, 55(9): 1163–1183, 2003. [PubMed: 12954197]
24. Huang G, Liu Z, Pleiss G, Van Der Maaten L, and Weinberger K. Convolutional Networks with Dense Connectivity. IEEE Trans. Pattern Anal. Mach. Intell, page 1, 2019. doi: 10.1109/TPAMI.2019.2918284.
25. Tieleman Tijmen and Hinton Geoffrey. 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.
26. Kingma Diederik P and Ba Jimmy. Adam: A method for stochastic optimization. arXiv, 2014. doi: arXiv1412.6980.
27. Belkin Mikhail, Hsu Daniel, Ma Siyuan, and Mandal Soumik. Reconciling modern machine learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci, 116(32):15849–15854, 2019. [PubMed: 31341078]
28. Zhao CY, Zhang HX, Zhang XY, Liu MC, Hu ZD, and Fan BT. Application of support vector machine (SVM) for prediction toxic activity of different data sets. Toxicology, 217(2-3): 105–119, 2006. [PubMed: 16213080]
29. Rupp Matthias, Ramakrishnan Raghunathan, and Lilienfeld O Anatole Von. Machine learning for quantum mechanical properties of atoms in molecules. J. Phys. Chem. Lett, 6(16):3309–3313, 2015.
30. Gerrard Will, Bratholm Lars A, Packer Martin J, Mulholland Adrian J, Glowacki David R, and Butts Craig P. IMPRESSION—prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. Chem. Sci, 2020.
31. Paruzzo Federico M, Hofstetter Albert, Musil Felix, De Sandip, Ceriotti Michele, and Emsley Lyndon. Chemical shifts in molecular solids by machine learning. Nat. Commun, 9(1):4501, 2018. [PubMed: 30374021]
32. Bartok Albert P, Kondor Risi, and Csanyi Gabor. On representing chemical environments. Phys. Rev. B, 87(18):184115, 2013.
33. Behler Jorg and Parrinello Michele. Generalized Neural-Network Representation of High Dimensional Potential-Energy Surfaces. Phys. Rev. Lett, 98(14):146401, 4 2007. [PubMed: 17501293]
34. You Yang, Demmel James, Hsieh Cho-Jui, and Vuduc Richard. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems. Proc. 2018 Int. Conf Supercomput, pages 307–317, 2018. doi: 10.1145/3205289.3205290.
35. Breiman Leo. Classification and regression trees. Routledge, 2017.
36. Svetnik Vladimir, Liaw Andy, Tong Christopher, Culberson J Christopher, Sheridan Robert P, and Feuston Bradley P. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci, 43(6):1947–1958, 2003. [PubMed: 14632445]
37. Haghighatlari Mojtaba and Hachmann Johannes. Advances of machine learning in molecular modeling and simulation. Curr. Opin. Chem. Eng, 23:51–57, 1 2019.
38. Hachmann J, Afzal MAF, Haghighatlari M, and Pal Y. Building and deploying a cyberinfrastructure for the data-driven design of chemical systems and the exploration of chemical space. Mol. Simul, 2018.
39. Pozdnyakov Sergey N, Willatt Michael J, Bartok Albert P, Ortner Christoph, Csanyi Gabor, and Ceriotti Michele. ArXiv.
40. Faber Felix A, Hutchison Luke, Huang Bing, Gilmer Justin, Schoenholz Samuel S, Dahl George E, Vinyals Oriol, Kearnes Steven, Riley Patrick F, and Lilienfeld O Anatole Von. Prediction errors of

molecular machine learning models lower than hybrid DFT error. J. Chem. Theory Comput, 13(11):5255–5264, 2017. [PubMed: 28926232]

41. Butler Keith T, Davies Daniel W, Cartwright Hugh, Isayev Olexandr, and Walsh Aron. Machine learning for molecular and materials science. Nature, 559(7715):547–555, 2018. [PubMed: 30046072]
42. Rajan Krishna. Informatics for materials science and engineering: data- driven discovery for accelerated experimentation and application. Amsterdam: Butterworth-Heinemann, 2013.
43. Duvenaud David, Maclaurin Dougal, Aguilera-Iparraguirre Jorge, Gomez-Bombarelli Rafael, Hirzel Timothy, Aspuru-Guzik Alan, and Adams Ryan P.. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Adv. Neural Inf. Process. Syst, pages 2224–2232, 2015. doi: 10.1021/acs.jcim.5b00572.
44. Kearnes Steven and Riley Patrick. Molecular graph convolutions: moving beyond fingerprints. J. Comput. Aided. Mol. Des, 2016.
45. Schutt Kristof T., Sauceda Huziel E., Kindermans Pieter-Jan, Tkatchenko Alexandre, and Muller Klaus- Robert. SchNet : A deep learning architecture for molecules and materials. J. Chem. Phys, 148(24):241722, 2018. [PubMed: 29960322]
46. Chen Chi, Ye Weike, Zuo Yunxing, Zheng Chen, and Ong Shyue Ping. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. Chem. Mater, 31(9):3564–3572, 2019.
47. Christensen Anders S, Bratholm Lars A, Faber Felix A, and Lilienfeld O Anatole von. FCHL revisited: Faster and more accurate quantum machine learning. J. Chem. Phys, 152(4):44107, 2020.
48. Forsyth David A and Ponce Jean. Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference, 2002 ISBN 0130851981.
49. Ramakrishnan Raghunathan, Dral Pavlo O, Rupp Matthias, and Lilienfeld O Anatole von. Quantum chemistry structures and properties of 134 kilo molecules. Sci. Data, 1:140022, 2014. [PubMed: 25977779]
50. Jha Dipendra, Choudhary Kamal, Tavazza Francesca, Liao Wei-keng, Choudhary Alok, Campbell Carelyn, and Agrawal Ankit. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. Nat. Commun, 10(1):5316, 2019. [PubMed: 31757948]
51. Friedland Gregory D and Kortemme Tanja. Designing ensembles in conformational and sequence space to characterize and engineer proteins. Curr. Opin. Struct. Bio, 20:377–384, 2010. [PubMed: 20303740]
52. Bhowmick A and Head-Gordon T. A Monte Carlo Method for Generating Side Chain Structural Ensembles. Structure, 23(1):44–55, 2015. [PubMed: 25482539]
53. Cui Jinlei, Olmsted David L, Mehta Anil K, Asta Mark, and Hayes Sophia E. NMR Crystallography: Evaluation of Hydrogen Positions in Hydromagnesite by 13C { 1H} REDOR Solid- State NMR and Density Functional Theory Calculation of Chemical Shielding Tensors. Angew. Chemie Int. Ed, 58(13):4210–4216, 2019.
54. Pickard Chris J and Mauri Francesco. All-electron magnetic response with pseudopotentials: NMR chemical shifts. Phys. Rev. B, 63(24):245101, 2001.
55. Shen Yang and Bax Ad. 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. [PubMed: 20628786]
56. Han Beomsoo, Liu Yifeng, Ginzinger Simon W, and Wishart David S. SHIFTX2: significantly improved protein chemical shift prediction. J. Biomol. NMR, 50(1):43, 2011. [PubMed: 21448735]
57. Cuny Jerome, Xie Yu, Pickard Chris J, and Hassanali Ali A. Ab initio quality NMR parameters in solid-state materials using a high-dimensional neural-network representation. J. Chem. Theory Comput, 12(2):765–773, 2016. [PubMed: 26730889]
58. Bartok Albert P, Payne Mike C, Kondor Risi, and Csanyi Gabor. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. Phys. Rev. Lett, 104(13):136403, 4 2010. [PubMed: 20481899]

59. Eickenberg Michael, Exarchakis Georgios, Him Matthew, Mallat StÃ¢Â Â ©phane, and Thiry Louis. Solid harmonic wavelet scattering for predictions of molecule properties. J. Chem. Phys, 148(24):241732, 2018. [PubMed: 29960365]

60. Haghighatlari Mojtaba, Vishwakarma Gaurav, Altarawy Doaa, Subramanian Ramachandran, Kota Bhargava Urala, Sonpal Aditya, Setlur Srirangaraj, and Hachmann Johannes. ChemML : A Machine Learning and Informatics Program Package for the Analysis , Mining , and Modeling of Chemical and Materials Data. Wiley Interdiscip. Rev. Mol. Sci, (n/a):e1458, 2019b. doi: 10.1002/wcms.1458.

61. Yamada Hironao, Liu Chang, Wu Stephen, Koyama Yukinori, Ju Shenghong, Shiomi Junichiro, Morikawa Junko, and Yoshida Ryo. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. ACS Cent. Sci, 5(10): 1717–1730, oct 2019. [PubMed: 31660440]

62. Smith Justin S, Nebgen Ben, Lubbers Nicholas, Isayev Olexandr, and Roitberg Adrian E. Less is more: Sampling chemical space with active learning. J. Chem. Phys, 148(24):241733, 2018. [PubMed: 29960353]

63. Roscher Ribana, Bohn Bastian, Duarte Marco F, and Garcke Jochen. Explainable Machine Learning for Scientific Insights and Discoveries. Arxiv:1905.08883, 5 2019.

64. Bach Sebastian, Binder Alexander, Montavon Gregoire, Klauschen Frederick, Muller Klaus Robert, and Samek Wojciech. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One, 10(7), 2015. doi: 10.1371/journal.pone.0130140.

Opportunities/Challenges:

- *Scarce and sparse chemical data sets.* The number of unique small molecules is practically infinite, with number estimates of possible synthesizable small molecules ranging from 10^{24}-10^{60}. Machine learning could expand coverage of chemical space – for instance in design, synthesis, and development stages of drugs – that traditionally is a resource-intensive task.
- *Machine learning chemical reactions* is a far more difficult task than using *ab initio* data to train non-reactive potential energy surfaces. It is the next frontier of machine learning in the molecular sciences– to generate a predictive map of chemical reactivity space which can chart all reaction pathways in complex environments.
- *Physics-informed machine learning:* Machine learning is designed to determine patterns in high-dimensional data that is otherwise not obvious and thus are perceived to have limited transparency. A worthy goal is to develop physics-inspired features and creation of data sets that model experiments for better understanding of machine learning outcomes.

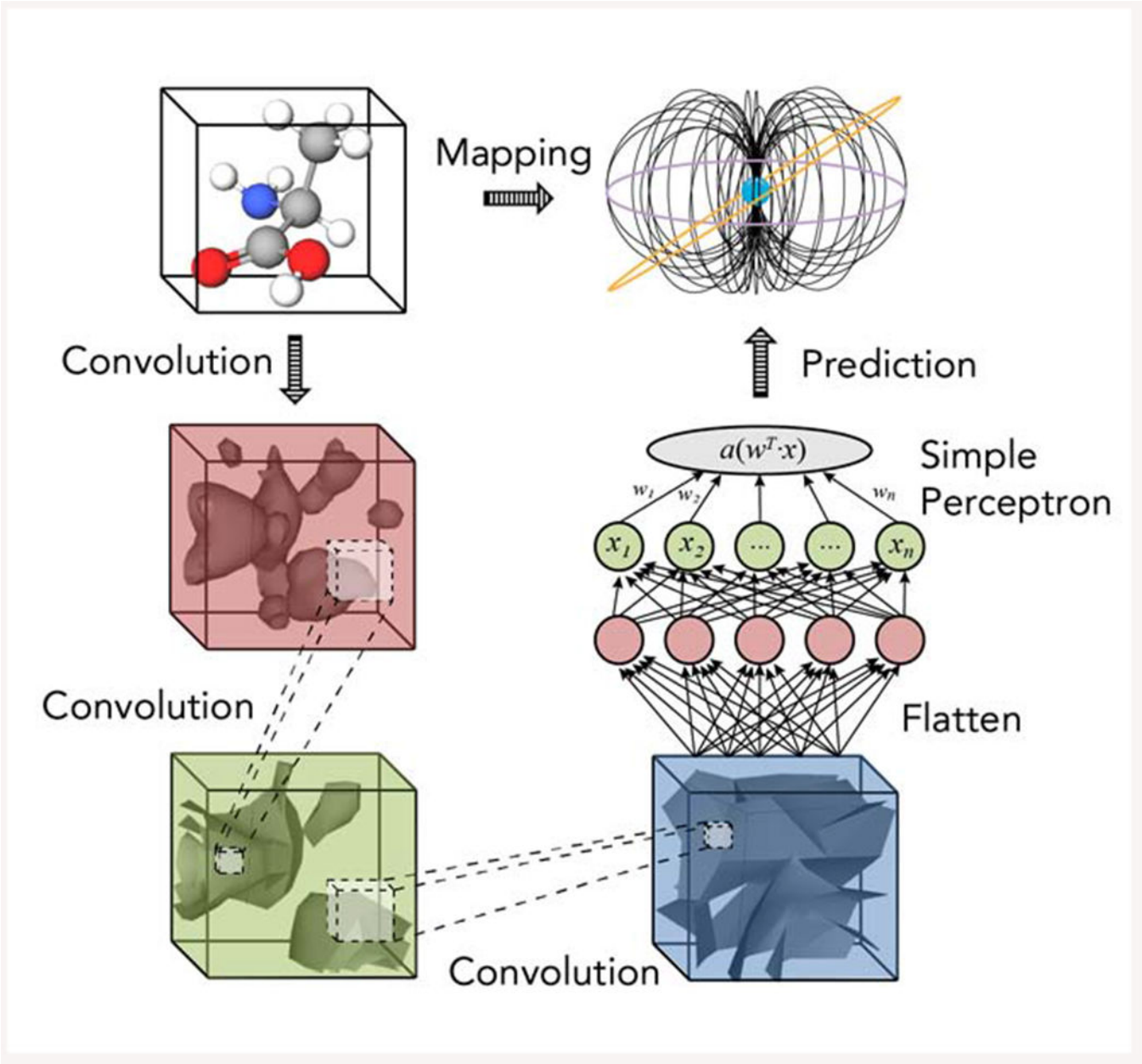


Figure 1.

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 or scalar coupling value.

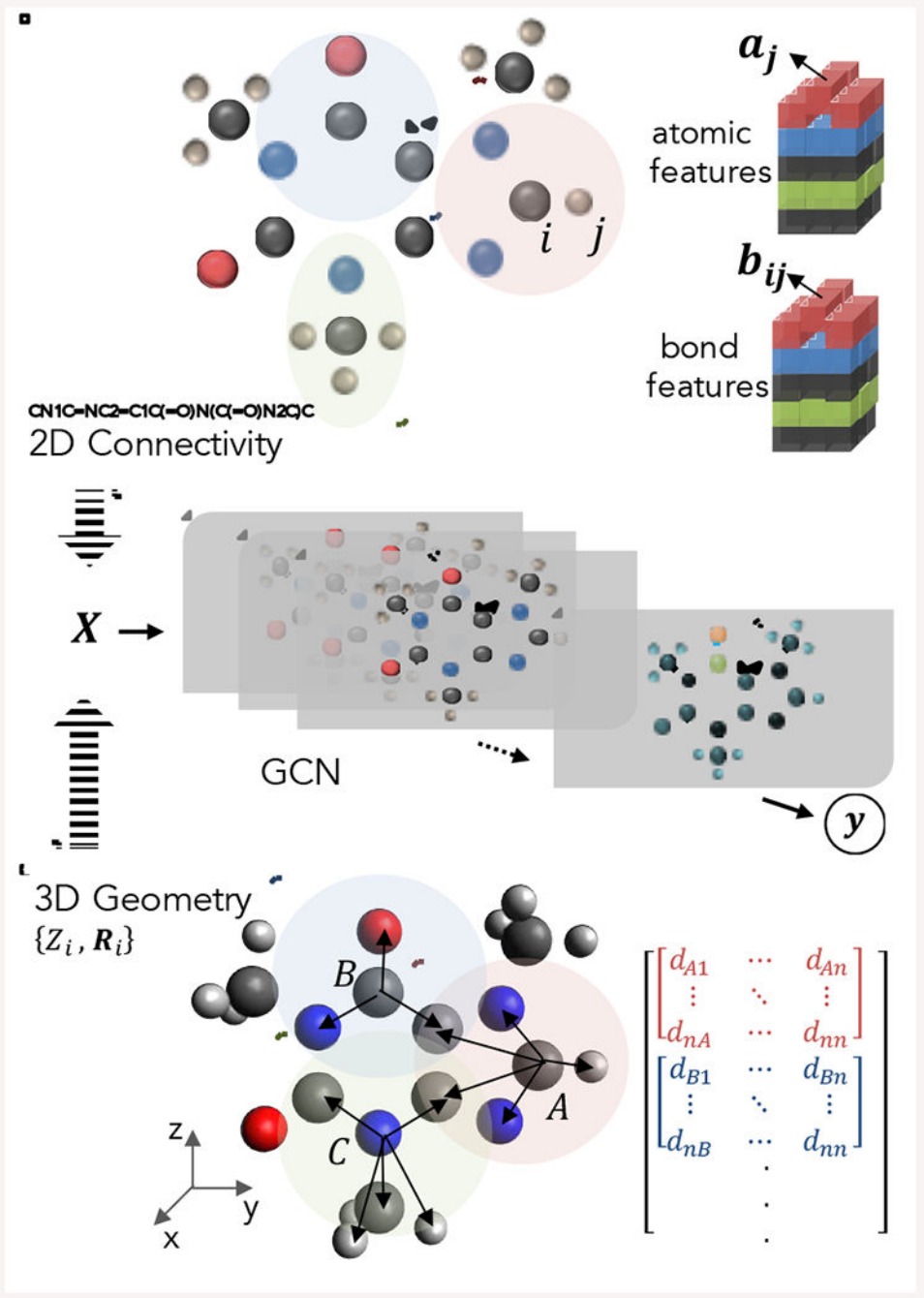
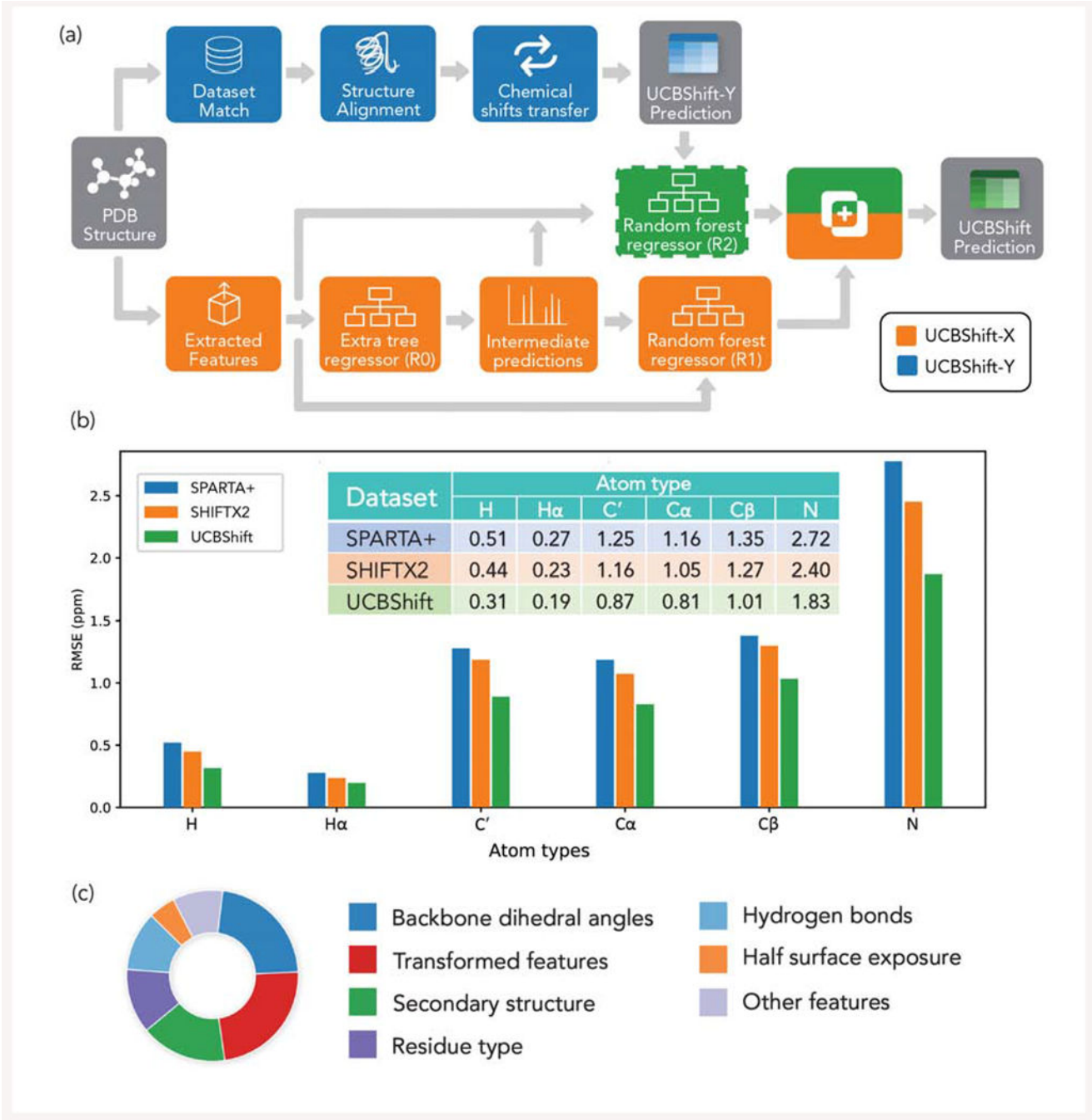


Figure 2.

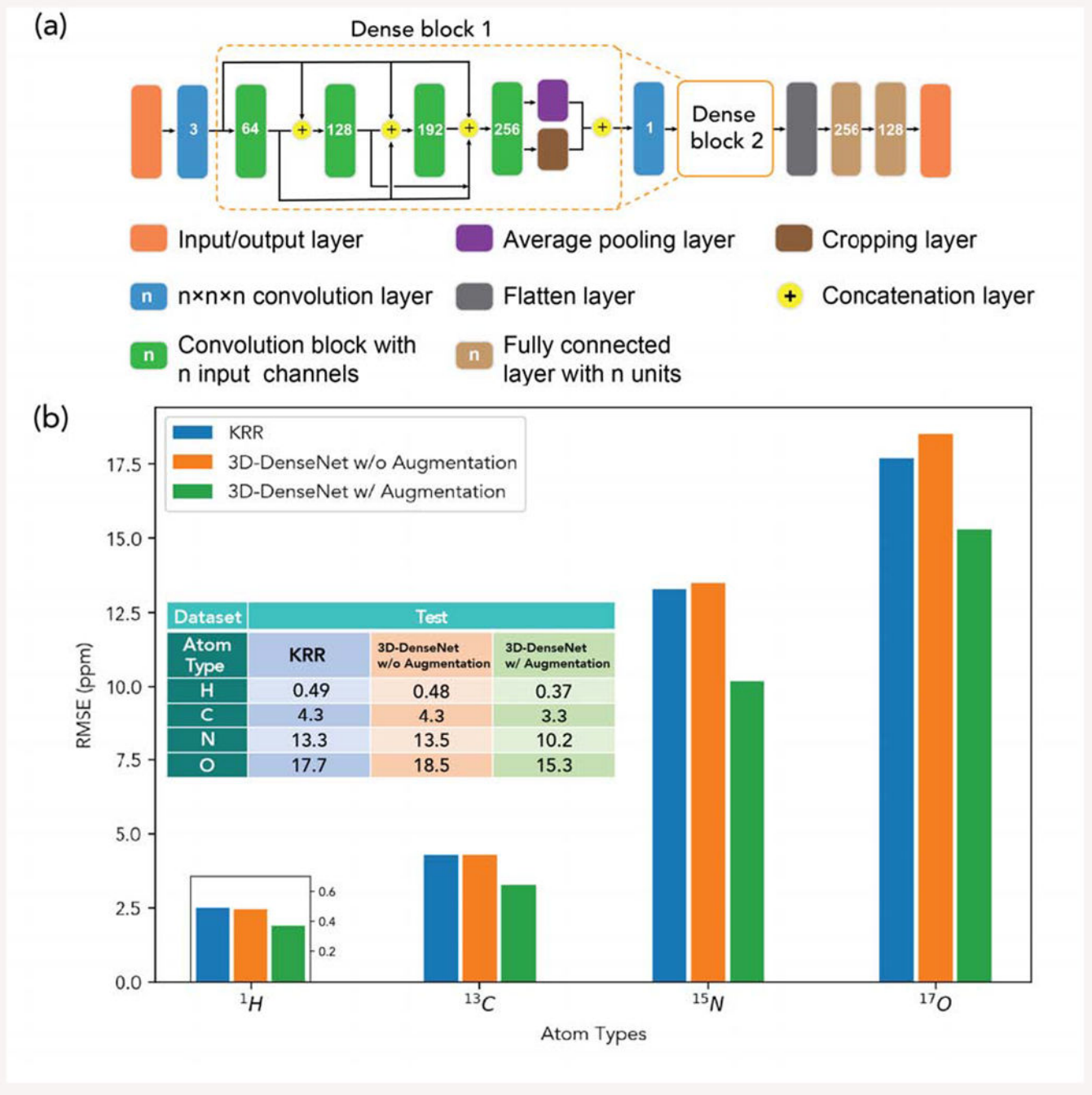
The illustration of graph convolutional networks (GCN) with different representations of the caffeine molecule as input. Molecular information can be represented as atomic and bond feature tensors extracted from connectivity based 2D information, or as distance matrices obtained from 3D coordinates, or any other form of sensible chemical representations.



Dataset	Atom type					
	H	Hα	C′	Cα	Cβ	N
SPARTA+	0.51	0.27	1.25	1.16	1.35	2.72
SHIFTX2	0.44	0.23	1.16	1.05	1.27	2.40
UCBShift	0.31	0.19	0.87	0.81	1.01	1.83

Figure 3.

(a) Illustration of the UCBShift algorithm[11] (b) Testing RMSEs (ppm) for each atom type from SPARTA+, SHIFTX2 and UCBShift, when evaluated on an independently generated test dataset (c) Relative importance of all the input features analyzed from UCBShift model.



Dataset	Test		
Atom Type	KRR	3D-DenseNet w/o Augmentation	3D-DenseNet w/ Augmentation
H	0.49	0.48	0.37
C	4.3	4.3	3.3
N	13.3	13.5	10.2
O	17.7	18.5	15.3

Figure 4.

(a) Illustration of MR-3D-DenseNet architecture (b) Testing RMSEs (ppm) for each atom type from KRR, 3D-DenseNet without data augmentation and 3D-DenseNet with data augmentation.