Machine Learning: Science and Technology

Revving up ^{13}C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
23 September 2020

REVISED
9 January 2021

ACCEPTED FOR PUBLICATION
4 February 2021

PUBLISHED
12 May 2021

PAPER

Revving up ^{13}C NMR shielding predictions across chemical space: benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules

Amit Gupta, Sabyasachi Chakraborty and Raghunathan Ramakrishnan

Tata Institute of Fundamental Research, Centre for Interdisciplinary Sciences, Hyderabad 500107, India

E-mail: ramakrishnan@tifrh.res.in

Keywords: NMR machine learning, kernel ridge regression, drug compounds

Supplementary material for this article is available online

Abstract

The requirement for accelerated and quantitatively accurate screening of nuclear magnetic resonance spectra across the small molecules chemical compound space is two-fold: (1) a robust 'local' machine learning (ML) strategy capturing the effect of the neighborhood on an atom's 'near-sighted' property—chemical shielding; (2) an accurate reference dataset generated with a state-of-the-art first-principles method for training. Herein we report the QM9-NMR dataset comprising isotropic shielding of over 0.8 million C atoms in 134k molecules of the QM9 dataset in gas and five common solvent phases. Using these data for training, we present benchmark results for the prediction transferability of kernel-ridge regression models with popular local descriptors. Our best model, trained on 100k samples, accurately predicts isotropic shielding of 50k 'hold-out' atoms with a mean error of less than 1.9 ppm. For the rapid prediction of new query molecules, the models were trained on geometries from an inexpensive theory. Furthermore, by using a Δ-ML strategy, we quench the error below 1.4 ppm. Finally, we test the transferability on non-trivial benchmark sets that include benchmark molecules comprising 10–17 heavy atoms and drugs.

1. Introduction

Nuclear magnetic resonance (NMR) is an indispensable tool in chemistry, biochemistry and biophysics. It is fast, accurate, information-rich and non-destructive, making it the ideal technique for detecting or describing chemical bonding scenarios. As easy and trivial have most NMR experiments become, it is still a computationally expensive task to estimate NMR shielding tensors or coupling constants for large molecular datasets [1, 2]. While molecules with heavy atoms demand incorporation of relativistic corrections to achieve quantitative accuracy [3, 4], computational NMR spectroscopy without such subtle effects are routinely used in organic chemistry [5–11]. For a comprehensive review on computational NMR refer to [12]. Recently, Grimme *et al* [13] discussed the automated prediction of spin-spin coupled ^{1}H NMR in various solvents by accessing relevant conformers, to generate experimentally relevant NMR spectra, while Buevich *et al* [14] employed computer-assisted structure elucidation algorithms and predicted NMR results to analyze molecular geometries. Lauro *et al* [15] designed a protocol to identify stereoisomers using experimental and predicted NMR data.

Amongst many *ab initiosu* quantum chemistry frameworks [16–20], gauge-independent atomic orbital (GIAO) [21] is the most popular. Within the GIAO framework, Cartesian components of the NMR shielding tensor, σ_{ij}^q, of a nucleus q is calculated as the second-order magnetic response property [22, 23]

$$\sigma_{ij}^q = \frac{\partial^2 E}{\partial B_i \partial \mu_j^q}, \tag{1}$$

where E is the electronic energy of the molecule, B_i is a component of the external magnetic field, and μ_j^q is the jth component of the magnetic moment of the nucleus q. The isotropic shielding is defined as one-third of shielding tensor's trace, $\sigma_{\text{iso}} = (\sigma_{11} + \sigma_{22} + \sigma_{33})/3$. Comparison of predicted values of σ_{iso} with experimental results is done by calculating its 'shift' using a standard reference compound $\delta_{\text{iso}}^q = \sigma_{\text{iso}}^{\text{reference}} - \sigma_{\text{iso}}^q$ [24].

^{1}H and ^{13}C are amongst the most commonly studied NMR-active nuclei. Accurate *ab initio* computation of δ ^{13}C requires methods such as coupled-cluster singles doubles [25], or spin-component-scaled MP2 with a triple-zeta quality basis set to reach a mean error of < 1.5 ppm [1, 26], albeit incurring a cost which prohibits the method's applicability for high-throughput studies. Composite methods have been proposed—analogous to the G*n* thermochemistry methods [27]—that exploit the additivity in basis set and correlation corrections to reach a greater accuracy [25, 28]. Another method involves tailoring exchange-correlation functionals of Kohn–Sham density functional approximations such as WC04 and WP04 [29].

When relaxing the accuracy requirement—while retaining the generality—a density functional approach that has received wide attention, particularly for NMR calculations of both ^{1}H & ^{13}C nuclei, is mPW1PW91 [30]. This method has been shown to provide good results for acetals [31], pyramidalized alkenes [32], acetylenes, allenes, cumulenes [33, 34] and even natural products [5, 6]. The same approach has also been used to model the 2D-NMR spectrum of exo-2-norbornanecarbamic acid [35]. Further, a multi-reference standard approach [36] has shown consistent estimations of chemical shifts in solutions with a triple-zeta basis set [37]. Thus, even though Flaig *et al*'s [26] benchmark study ranked the B97-2 functional high, next only to the MP2 method, the consistency of mPW1PW91 has motivated several works including a recent effort by Gerrard *et al* [38], where the authors applied mPW1PW91 with the 6-311G(*d*,*p*) basis set to generate NMR chemical shielding and hetero-nuclear coupling constants of molecular components in experimentally characterized organic solids.

While direct application of DFT is feasible for any query molecule, the questions that arise in chemical compound space (CCS) explorations often concern property trends across large datasets, demanding realistically rapid evaluation of the desired property. To this end, machine learning (ML) based statistical inference, in combination with high-throughput *ab initio* computing, offers a viable alternative (see [39]). This approach has received such widespread attention that a recent competition on the world-wide web, Kaggle, for ML-aided prediction of NMR spectra [40] saw a participation of 2700 teams across the world. An earlier proof-of-concept study discussed the feasibility of exploiting the local behavior of NMR chemical shifts with ML to achieve transferability to systems that are larger than those used to train the model [41]. That work depended on a cut-off based local version of the Coulomb matrix (CM) descriptor [42]. Recently, Gao *et al* [43] explored deep neural networks (DNNs) in their 'DFT + ML' model and achieved mean-squared error of 2.10 ppm for ^{13}C chemical shifts compared to experimental values. DNN has also been employed for modeling electronic spectra [44–46]. Kernel-ridge regression (KRR) is another ML method offering accuracies comparable to that of DNN for spectroscopy applications [47, 48]. However, ML/deep learning may not be limited to single property applications when multiple properties can be explored [44, 49, 50].

As for descriptors, successive improvements have been made by projecting the three-dimensional molecular chemical structure into multidimensional tensors [51], four-dimensional hyper-spherical harmonics [52], or a continuous representation such as the variant smooth overlap of atomic positions—SOAP [53, 54]. The latter with Gaussian process regression [55] predicted chemical shifts with root-mean-squared-error (RMSE) of 0.5/4.3 ppm for ^{1}H/^{13}C nuclei on 2k molecular solids, while with KRR it was successful in predicting ^{29}Si and ^{17}O NMR shifts in glassy aluminosilicates across a wide temperature range [56] comparable to fragment-based estimations [57].

The joint descriptor-kernel formalism of Faber, Christensen, Huang, and Lilienfeld (FCHL) uses an integrated Gaussian kernel function accounting for three-body interactions in atomic environment yielding highly accurate results for global molecular properties such as atomization energies [58]. Recently, FCHL-based KRR has been applied to model ^{1}H, ^{13}C shifts and J-coupling constants between these two nuclei for over 75k structures in the CSD [38]. For a test set, which was not part of training, the same study noted mean absolute errors (MAE) of 0.23 ppm/2.45 ppm/0.87 Hz (RMSE: 0.35 ppm/3.88 ppm/1.39 Hz) for δ ^{1}H/δ ^{13}C/$^1J_{\text{CH}}$, respectively.

Here, we present gas and (implicit) solvent phase mPW1PW91/6-311+G(2*d*,*p*)-level chemical shielding for all atoms in the QM9 dataset [59] comprising 130 831 stable, synthetically feasible small organic molecules with up to 9 heavy atoms C, N, O and F—henceforth denoted the QM9-NMR dataset. Initial structures of the QM9 molecules are based on the SMILES descriptors from GDB17 chemical Universe. We apply KRR-ML using training sets drawn from QM9-NMR benchmark control settings, and rationalize their influence on the performance of large ML models using up to 100k training examples. It has been recently

noted [60] that the Δ-ML approach [61] facilitates better ML accuracies. Thus, with converged settings, we provide benchmark learning curves for ML and Δ-ML methods based on three local descriptors—CM, SOAP and FCHL. Finally, we evaluate the transferability of the local ML models—trained only on QM9 molecules—to larger systems in non-trivial benchmark sets that include several drug molecules, a small subset of GDB17 [62] with molecules comprising 10–17 heavy atoms and linear polycyclic aromatic hydrocarbons (PAHs).

2. Methodology

Among popular ML frameworks, KRR is one of the most consistent and accurate [63] framework. In KRR formalism, for a query entity (molecule or atom), q, a generic property p from a reference (experiment or theory), is estimated as a linear combination of radial basis functions (RBFs a.k.a. kernel functions)—each centered at one training entity. Values of these RBFs are calculated at q, then the distances between q and N training molecules defined via their descriptors $\mathbf{d}$ is given as

$$p^{\text{est}}(\mathbf{d}_q) = \sum_{t=1}^{N} c_t k(|\mathbf{d}_q - \mathbf{d}_t|). \tag{2}$$

The coefficients, c_t, one per training datum, are obtained through ridge-regression by minimizing the least-squares prediction error:

$$\begin{aligned} \mathcal{L} &= \langle \mathbf{p}^{\text{ref}} - \mathbf{p}^{\text{est}} | \mathbf{p}^{\text{ref}} - \mathbf{p}^{\text{est}} \rangle + \lambda \langle \mathbf{c} | \mathbf{c} \rangle \\ &= \langle \mathbf{p}^{\text{ref}} - \mathbf{K}\mathbf{c} | \mathbf{p}^{\text{ref}} - \mathbf{K}\mathbf{c} \rangle + \lambda \langle \mathbf{c} | \mathbf{c} \rangle. \end{aligned} \tag{3}$$

The size of the kernel matrix is $N \times N$, each element defined in close analogy to the right side of equation (2), $K_{ij} = k(||\mathbf{d}_i - \mathbf{d}_j||)$, i and j going over N training elements, with $||\cdot||$ denoting a vector norm. Here, for the choice of CM and SOAP descriptors, we used the Laplacian kernel depending on an L_1 norm defined as $K_{ij} = \exp(-|\mathbf{d}_i - \mathbf{d}_j|/\omega)$, where ω defines the length scale of the exponential RBF. As shown in [64], optimal solution to equation (3) amounts to solving the linear system:

$$[\mathbf{K} + \lambda \mathbf{I}]\ \mathbf{c} = \mathbf{p}^{\text{reference}}. \tag{4}$$

The second term in the r.h.s. of equation (3) is apparent in equation (4); if the definition of the descriptor does not differentiate any two training entries, then $\mathbf{K}$ becomes singular and a unique solution to equation (4) can only be found with a non-zero value for the regularization strength, λ. Both ω and λ constitute hyperparameters in the model, that require a cross-validated optimization before out-of-sample predictions. Any non-zero value of λ determined by cross-validation is an indication of the presence of redundant training entries, either due to data-duplication or poor quality of the descriptor. As shown in [65], in the absence of redundant training entries, λ can be set to zero and the learning problem translates to solving $\mathbf{K}\mathbf{c} = \mathbf{p}^{\text{reference}}$. Alternatively, when linear dependencies may be anticipated—due to numerically similar descriptor differences—rendering an off-diagonal element of $\mathbf{K}$ to be ≈ 1, a finite $\lambda = \varepsilon$ may be used to shift the diagonal elements of $\mathbf{K}$ away from 1.0 and the lowest eigenvalue away from 0.0 thereby aiding Cholesky decomposition.

Prior regularization, $\mathbf{K}$ is a covariance or dispersion matrix with all of its off-diagonal elements bound strictly in the closed interval $[0, 1]$ with unit diagonal elements. As per [65], we can estimate ω independent of property by restricting K_{ij} corresponding to the largest descriptor difference, D_{ij}^{max}, to 0.5, as in

$$\omega_{\text{opt}}^{\text{max}} = D_{ij}^{\text{max}} / \log(2). \tag{5}$$

In the present study, we also explore the performances of the choices of ω based on D_{ij}^{mean} and D_{ij}^{median} that will differ from the value of ω_{opt} based on D_{ij}^{max} depending on the diversity of the training set descriptors:

$$\omega_{\text{opt}}^{\text{mean}} = D_{ij}^{\text{mean}} / \log(2); \omega_{\text{opt}}^{\text{median}} = D_{ij}^{\text{median}} / \log(2). \tag{6}$$

Later we show how these choices are in close agreement with ω_{opt} values found by a scan to minimize the error for a large hold out set. We also discuss how the kernel matrix constructed with $\omega_{\text{opt}}^{\text{median}}$ can be applied to model NMR shieldings from gas and different solvent phases.

The prediction error of an ML model can be unconditionally quenched with increasing training set size for a good choice of the descriptor; however, the exponential nature of the learning rate often necessitates an increase in the model's size by orders of magnitude. While the resulting surge in the computational cost

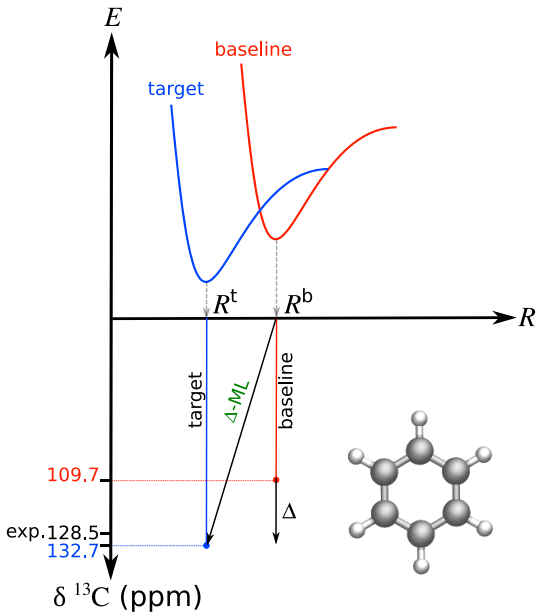

Figure 1. Δ-ML of NMR chemical shifts exemplified by benzene; the model is trained to predict property from a target-level theory using as inputs atomic coordinates and the property from a baseline theory. The targetline and baseline shifts of benzene were computed at mPW1PW91/6-311+G(2*d*,*p*)@B3LYP/6-31G(2*df*,*p*) and B3LYP/STO-3G@PM7 levels, respectively.

associated with the ML model's execution speed is seldom prohibitive, training with examples of the order of 10^6 places too severe hardware restrictions. When such hardware limit is reached for training, further drop in an ML model's error can be attained by training on the deviation of the property from inexpensive, yet qualitatively accurate baseline values in a Δ-ML fashion [61]:

$$\Delta p(\mathbf{d}_q^{\text{bas.}}) = p^{\text{tar.}}(\mathbf{d}_q^{\text{tar.}}) - p^{\text{bas.}}(\mathbf{d}_q^{\text{bas.}}). \tag{7}$$

The ML problem now involves solving for $\mathbf{Kc} = \boldsymbol{\Delta p}$. For any new prediction, Δ is augmented with the baseline:

$$p^{\text{est.}}(\mathbf{d}_q^{\text{tar.}}) = p^{\text{bas.}}(\mathbf{d}_q^{\text{bas.}}) + \sum_{t=1}^{N} c_t' k(|\mathbf{d}_q^{\text{bas.}} - \mathbf{d}_t^{\text{bas.}}|) \tag{8}$$

where c_t' is now obtained through ridge-regression by minimizing the property differences between target and baseline. Figure 1 illustrates Δ-ML for the modeling of NMR shifts with an example molecule. Often, for any given molecule, the determination of minimum energy geometry at the target level incurs a greater computational requirement than that is needed for the estimation of NMR shielding. In the Δ-ML framework, this problem can be alleviated by using atomic coordinates calculated at the same or a different baseline level for the construction of descriptors. Hence, new predictions can be rapidly made using structural information calculated at the baseline-level. In the present study, we use NMR properties calculated with an inexpensive theory (B3LYP/STO-3G@PM7) as a baseline. The Δ-ML model is trained on the differences between the target (mPW1PW91/6-311+G(2*d*,*p*)@B3LYP/6-31G(2*df*,*p*)) and the baseline (figure 1). For a new query, the model predicts the 'Δ' to which calculated baseline value is added.

The formal requirements for a chemical descriptor have been discussed by others [53, 66–71]. Design of structure-based molecular descriptors drew inspiration from the success of generic coordinates such as atom-centered symmetry functions [72, 73]. Here, we explore CM [42], SOAP [53, 54, 74] and FCHL (without alchemical correction) [58]. For modeling local properties such as NMR shielding, CM can be truncated by a cutoff radius, r_{cut} [41], but there is always the possibility of failing to establish injective mapping between three dimensional molecular structure and the query property [75]. Hence, more robust approaches include row-norm sorted CM or 'bag-of-bonds', [76–78]. Models based on all 3 descriptors show similar prediction times. For a detailed account of solver and prediction times on a single run, see table S2.

3. Computational details

For training data in ML, we collected B3LYP/6-31G(2*df*,*p*)-level minimum energy geometries of 134k molecules in the QM9 dataset from [59]. Those structures that have been reported to fragment during the

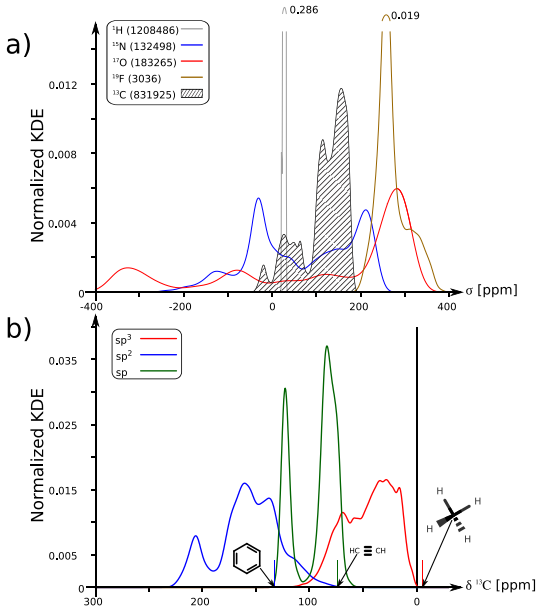

Figure 2. Range of NMR shielding properties in the QM9-NMR dataset. (a) Element-wise distribution of the shielding (σ in ppm) of all the nuclei in 131k QM9 molecules. (b) ^{13}C chemical shifts (δ in ppm) for the QM9 molecules classified according to hybridization. KDE stands for kernel density estimation.

geometry relaxation (3054 in total) were excluded in this study. NMR shielding tensors of selected stable nuclei were calculated at the mPW1PW91/6-311+G(2*d*,*p*)-level in a single-point fashion within the GIAO framework [79–81] using Gaussian-16 suite of programs [82]. In all DFT calculations, integration grid was set to `Ultrafine` with a `VeryTight` SCF threshold. To use as a baseline property in Δ-ML, we used NMR shielding calculated at the B3LYP/STO-3G level with geometries optimized at the PM7 level, the latter done with MOPAC [83]. ^{13}C isotropic shielding tensors, σ, were converted to ^{13}C chemical shifts, δ, using a reference value for σ corresponding to that of tetramethylsilane (TMS), which was calculated in gas phase to be 186.97 ppm. We have also computed shielding tensors for the entire 131k set with implicit modeling of the solvents—carbon tetrachloride (CCl_4), tetrahydrofuran (THF), acetone, dimethyl sulfoxide (DMSO), and methanol—with the polarizable continuum model (PCM) [84]. This was achieved by invoking `SCRF` in Gaussian-16 and specifying the solvent name and retaining default settings.

We have retained the same settings—mPW1PW91/6-311+G(2*d*,*p*)@B3LYP/6-31G(2*df*,*p*)—to calculate the NMR shielding of benchmark molecules that we selected for validating QM9-based ML models. Initial unrelaxed structures of the linear PAHs studied here have been taken from [85]. From the GDB17 dataset, we have randomly selected eight subsets of molecules, each with 25 molecules comprising 10–17 heavy atoms (200 in total). Further, we collected drug molecules present in the GDB17 dataset identified in [62, 86–88]. In addition, we also collected 12 somewhat larger drug molecules from [89]. The corresponding SMILES strings of all these 'validation' molecules, when available, were converted to initial Cartesian coordinates using the program Openbabel [90]. Initial coordinates of the 12 large drug molecules were created using the Avogadro [91] program. All molecules have been subjected to preliminary geometry relaxation performed with the force field MMFF94 [92]. We used the default settings in DScribe [93] and QML [94] to calculate the SOAP descriptor and the FCHL kernel matrix, respectively. All ML calculations have been performed using codes written in Fortran90 with interfaces to the SCALAPACK [95] numerical library.

4. Results and discussions

4.1. QM9-NMR dataset

For a systematic exploration of NMR properties across the QM9 CCS, QM9-NMR dataset was created as per the procedures outlined in section 3. This dataset consists of data for stable 130,831 molecules amounting to 1 208 486 (1.3 M), 831 925 (832 k), 132 498 (132 k), 183 265 (183 k), 3 036 (3 k), NMR values for H, C, N, O, and F nuclei, respectively. DFT-level NMR shielding of these elements (figure 2(a)) demonstrate the expected range of values. In case of H, the most deshielded nucleus corresponds to the one from the cationic ammonium ion (encountered in zwitterionic molecules), while the most shielded proton belongs to a highly-strained secondary amine bonded to N. Methane offers the most shielded environment for ^{13}C in QM9, whereas the most deshielded C features in a highly strained multiply-fused-ring molecule. Similarly,

for N, the most shielded nuclei comes from a strained tertiary amine, whereas for O it features in a strained ring. Most deshielded N and O nuclei belong to a zwitterionic molecule.

Besides extrema, figure 2(a) also highlights the chemical diversity of the QM9 dataset. For C and H atoms, majority of the NMR shielding parameters come from C(sp^3)-H bonds, indicating QM9 to largely comprise saturated organic molecules. Unsaturated molecules form a relatively smaller fraction of QM9 as can be seen in its shielding distribution function (between 0 and 75 ppm for C, figure 2(b). N atom distribution shows two sharp distribution peaks at about 200 and −35 ppm belonging to primary amine and cyano groups, respectively. Most frequent O atoms consist of ether linkages, while F atoms show a characteristic broad distribution around 250 ppm.

NMR shielding values for the entire dataset have also been calculated with continuum models of five commonly used polar and non-polar organic solvents: acetone, CCl_4, DMSO, methanol, and THF. Formally, a polar solvent will result in a more deshielded environment. However, the influence of the solvent is non-uniform across various C atoms in a molecule depending on the local environment of an atom in the molecule. Other effects such as hydrogen-bonding, halogen bonding may further influence the chemistry of the molecule resulting in unexpected chemical shifts. Thus, it is necessary to build a database comprising NMR shielding tensors calculated at various solvent media. Here as a first step, we computed the shielding values of the molecules in different solvents under a PCM framework. For a better description of the solvent environment it is essential to go beyond continuum modeling by using micro-solvation models that account for explicit solute-solvent interactions. The solvents were chosen to represent diverse environments: non-polar, polar aprotic and polar protic. For any given ^{13}C nucleus, the spread in the shielding values due to the choice of the medium is at the most ±4 ppm (figure S2) with minima and mean at 0 and 0.56 ppm, respectively, suggesting most nuclei to be minimally influenced by the implicit solvent environment. Hence, for ML predictions to differentiate the results from various phases, it is necessary that the models' prediction accuracy is much less than ±4 ppm.

QM9-NMR dataset also contains B3LYP/STO-3G NMR shielding constants for all the 130 831 molecules. Although the current ML study concerns itself with NMR shielding of the C-atom, the QM9-NMR dataset can be used to model other nuclei as well. To facilitate such and other *ab initio* benchmark efforts, the entire QM9-NMR dataset, comprising gas and solvent phase results, is now a part of the openly accessible MolDis big data analytics platform [96], https://moldis.tifrh.res.in/data/QM9NMR.

4.2. Atoms-in-molecule ML modeling of NMR shielding

Following the generation of the QM9-NMR dataset with 812k ^{13}C nuclei, we have selected a random set of 100k entries for training the ML models. Further, a separate subset of 50k nuclei—not overlapping with the 100k training entries—was kept for validating the ML models. Additionally, we have compared the distribution of the training and validation subsets with the total set, and found the normalized density distributions to be similar (figure S1 is available online at stacks.iop.org/MLST/2/035010/mmedia). Therefore, we believe that the ML models based on large training sets presented here do not suffer from a selection bias.

For CM and SOAP descriptors, we used $\omega_{\text{opt}}^{\text{mean}}$ of 422.78 and 18.85, respectively which also agreed to values obtained via cross-validation (see table S1 and figure S4). Since FCHL implementation in QML does not provide D_{ij} values, a grid search showed the best ω to be 0.3 (figure S5). In all ML calculations, we used $\lambda = 10^{-3}$ determined through a logarithmic grid search. Since NMR shifts are a local property, each training entity is an atom with its environment determined by a distance cutoff. For CM, SOAP, and FCHL we found optimal cutoffs to be 2.3, 2.0, and 4.0 Å, respectively (see figure S3). Figure S7 features the distribution of the kernel matrix elements based on 10k training examples for all three descriptors. While the distributions for CM and SOAP are rather univariate, FCHL's K_{ij} values show a multivariate distribution, implying the latter model to be sensitive to the choice of the kernel width. To save computational time, alchemical corrections in FCHL were switched off. At the limit of large training set size, alchemical corrections made negligible improvements (see figure S8).

After determining the most appropriate hyperparameters for various choices of descriptors, we collected 10 training sets of sizes: 100, 200, 500, 1k, 2k, 5k, 10k, 20k, 50k, and 100k. We ensured that each smaller dataset is a subset of a larger one making the learning monotonous. We solved the linear equations of ML (equation (4)) using Cholesky decomposition, and the trained machine was used to predict NMR shifts of 50k out-of-sample validation set. Mean absolute error (in ppm) for these 50k predictions was admitted as the sole performance metric of the training accuracy (figure 3). It also shows the performances of Δ-ML carried out using B3LYP/STO-3G NMR parameters at PM7 structures.

Overall, one notes from figure 3 that for all descriptors Δ-ML models converging by more than an order of magnitude faster (in terms of training set size) than direct ML ones. It is also evident that among all three

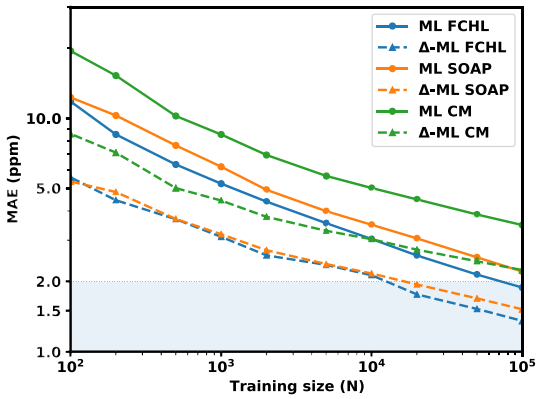

Figure 3. ML and Δ-ML out-of-sample prediction errors for CM, SOAP and FCHL descriptors. For increasing training set size, *N*, mean absolute error (MAE) in the prediction of NMR shielding of 50k hold-out ^{13}C atoms are shown.

Table 1. Prediction errors of FCHL-based ML and Δ-ML models, with 100k training examples, in different media. Mean absolute errors in the prediction of the NMR shielding of 50k hold-out ^{13}C atoms are reported in ppm.

Medium (ε)	ML	Δ-ML
Gas	1.88	1.36
CCl_4 (2.228)	1.91	1.38
THF (7.426)	1.99	1.48
Acetone (20.493)	1.93	1.42
Methanol (32.613)	1.94	1.42
DMSO (46.826)	1.99	1.49

descriptors, FCHL delivers the best performance with an average prediction error of <2 ppm; the error drops below 1.4 ppm for Δ-ML modeling. However, it may be noted that for training set sizes $\leqslant$10k, both FCHL and SOAP-based Δ-ML models yielded identical predictions, with SOAP showing an exponential learning rate, while FCHL showing a slightly faster rate going from 10k to 20k training examples—both Δ-ML models delivering ≈ 2 ppm accuracy already for 20k training. The similarity in performance between SOAP and FCHL can be observed in figure S10 both yielding similar correlation-coefficient across the training set in both ML and Δ-ML. For all case-studies, we used FCHL-100k ML and Δ-ML machines.

The origin of accuracy limiting factors in ML was further investigated by categorizing errors based on the NMR shielding range and the representation of the categorized region in the training dataset (figure S9). We found that the errors were not uniformly distributed and for certain shielding constant ranges, mean and variance of predictions were more erratic. The anomalous error in the -25 to 25 ppm region can be explained by (a) under-representation of the aromatic or unsaturated systems in QM9 and (b) larger chemical diversity in the shifts of the unsaturated regions. In the 150–175 ppm region, where we found majority of C shielding values of the QM9-NMR dataset to lie, the prediction errors were rather low and less spread out. We note that as more data is added in the erroneous regions (such as the -25 to 25 ppm region), the accuracy of the NMR machine improves.

We also probed if the baseline ^{13}C shielding values computed in gas phase can be utilized also for modeling DFT-level values in various solvents. While it is possible to simultaneously model on multiple property vectors by feeding in a rectangular matrix—row of column vectors—to the Cholesky procedure, the cost of training can be slightly minimized by inverting the kernel matrix once and multiplied with any arbitrary property vector to get new training coefficients [65]. Table 1 demonstrates the versatility of this approach. Inverted FCHL-100k kernel instantly yielded trained machines for all solvents, with sub 2 ppm accuracy. From gas phase to DMSO ($\varepsilon = 46.8$), we note a modest deterioration in performance.

4.3. Application of FCHL-based ML and Δ-ML models

The magnitude of NMR chemical shift/shielding of a ^{13}C nucleus in a molecule is governed by its local environment. The inherent locality of this property implicitly suggests the shielding effect to drop with increasing distance. Subsequently, the information gained from a local moiety of a small test molecule can be reasonably transferred to the same local environment in a large molecule, provided the moiety is not perturbed by chemical interactions alien to the training molecule.

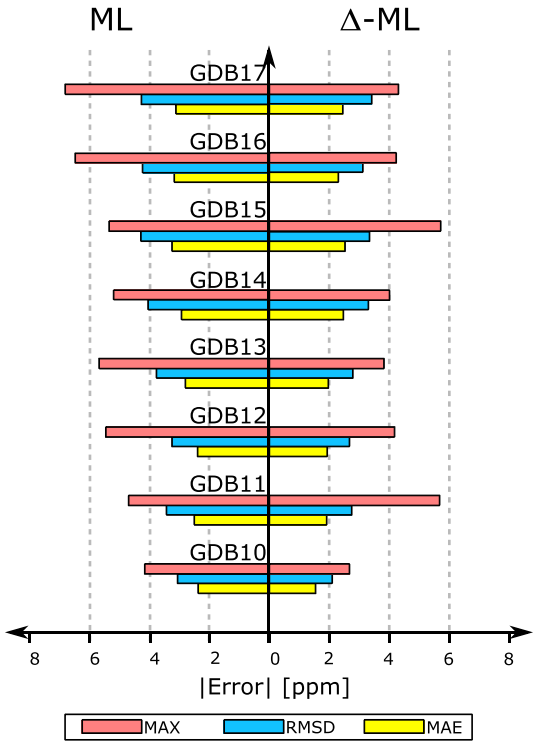

Figure 4. Error metrics for ML and Δ-ML predictions of mPW1PW91/6-311+G(2*d*,*p*)-level ^{13}C chemical shifts of randomly chosen 25 molecules with *n* heavy atoms from the GDB*n* subsets of GDB17; $n = 10, \ldots, 17$. MAE, RMSD and MAX correspond to mean absolute error, root-mean-squared-deviation and maximal error after averaging over all C atoms in a molecule, respectively.

Using the 100k ML and Δ-ML machines, we investigate how well these properties can be estimated for larger molecules. The graph-based design of the GDB datasets allows one to explore CCS in an unbiased fashion. In figure 4, we explore our ML and Δ-ML models' performances across GDB*n* datasets (where $n = 10, 11, \ldots, 17$) using 25 randomly chosen molecules per *n*. Each of these 200 molecules were relaxed at the B3LYP/6-31G(2*df*,*p*) level with reference NMR shielding tensors calculated at the mPW1PW91/6-311+G(2*d*,*p*) level (section 3).

As expected, Δ-ML generally improves upon ML consistently yielding lower MAE and RMSD. Further, we note maximum average error per molecule (MAX) to overall improve with Δ-ML except for GDB15 and GDB11 possibly due to systems with interactions alien to QM9. In figure 4, ML provides an MAE of <4 ppm across the datasets while it is usually below 3 ppm for Δ-ML. Arguably, 25 random molecules is not an accurate representation of the entire dataset in question and hence trends intrinsic to these subsets are not transferable across sets. Still, a general observation can be made: increasing number of heavy atoms introduces long-range influences on moieties rendering our machines somewhat inefficient—MAE of both ML and Δ-ML generally increases as we explore larger systems. Thus, apart from small fluctuation in the error trend across GDB10-GDB17—possibly due to sampling bias—the local models trained on QM9 provide quantitative prediction for molecules larger than those used in training.

GDB*n* Universe contains many drug molecules [62, 86–88]. Figure S11 displays 40 such molecules. We tested our models' efficiency in predicting ^{13}C NMR shielding values, and present their error metrics across direct and Δ-ML machines for each molecule. As Spearman coefficients are sensitive to numerical precision, we utilized a modified version by mapping the stick spectrum of the NMR shielding by a step function of height 1 and width of 1 ppm, when needed. The largest Δ-ML error encountered in this set is for desflurane due to the presence of di- and tri-fluoro methyl groups that are under-represented in the training set. The second largest Δ-ML error is for diethylcarbamazine stemming from deficiencies in baseline data. We note a total of 25 and five systems to show MAE higher than 3.0 ppm in ML and Δ-ML modeling, respectively. Barring six systems, Δ-ML improves upon direct ML's MAE, a trend previously noted in figure 3 and figure 4; not only does it improve MAE, but for 14 systems it also improves ρ. Evidently, Δ-ML modeling helps to reach semi-quantitative predictions due to the accuracy of the baseline. In figure S12, we present 12 extra-GDB17 drugs with their error metrics and DFT chemical shifts. As expected, Δ-ML outperforms direct-ML consistently across all 12 molecules. The highest deviation is noted for Morphine with Δ-ML presenting an MAE >3.0 ppm possibly because of moieties under-represented in QM9. However, for others the inherent locality of NMR shifts aid prediction. Drugs with extended delocalization present errors, since

conjugation is inadequately captured in our ML-models. This deficiency is further noted in figure S13 where predicted chemical shifts of interstitial atoms of PAHs show the maximum deviation.

5. Conclusion

We present the QM9-NMR dataset that augments the QM9 set [59]—containing DFT-level structures and properties of 134k organic molecules—with NMR shielding values computed at the mPW1PW91/6-311+G(2*d*,*p*)-level for about 2.4 million atoms constituting the molecules in this dataset. It may be further extended by including *J*-coupling between ^{1}H and other nuclei so that the diverse array of nuclei and properties present in QM9-NMR may aid seamless data-mining or ML studies. The impressive size of the dataset compelled us to explore solvent-phase values using an implicit solvation model, which however may not be adequate to describe effects due to the solute-solvent explicit interactions as addressed in [97]. We focus on predicting the isotropic shielding values of ^{13}C nuclei in QM9 entries through KRR-ML models with Laplacian kernels. Upon benchmarking the performance of ML models across three descriptors: CM, SOAP and FCHL, we note a monotonous improvement in learning with increasing training set size up to 100k, with respect to predictions for a 50k hold-out set, where an FCHL-based (without alchemical corrections) ML-model showed the least MAE of 1.88 ppm. Δ-ML, using PM7 geometries and B3LYP/STO-3G baseline values, improves upon this accuracy to yield an MAE of 1.36 ppm. This is an improvement over the current record in out-of-sample prediction error in data-driven ^{13}C nuclei NMR shielding modeling [38]. SOAP-based ML model's under-performance could be speculated to the use of Laplacian kernel-based KRR when Gaussian process regression is more effective. The performance drops with increasing diversity of validation molecules but the target being of local nature benefits from our models and aids in the prediction of ^{13}C shielding in molecules much larger than those in training set. Such a trend has been noted during the validation of ^{13}C shielding for a random subset of 25 molecules collected from GDB10 to GDB17 sets. Although, the prediction accuracy decreased with increasing molecular sizes, the MAE reported across datasets remained within 4.0 ppm for ML and 3.0 ppm for Δ-ML. When predicting ^{13}C shielding of drug molecules—one containing 40 drug molecules from GDB17 Universe (figure S11), and the other containing 12 drugs with 17 or more heavy atoms (figure S12)—Δ-ML improves upon ML's performance with the MAE decreasing from 3.7/4.2 ppm to 2.3/2.6 ppm for 40-drug/12-drug datasets, respectively. However, delocalization in linear PAHs (figure S13) proves challenging because of the small cutoff values decided from cross validation on molecules lacking such effects.

While the deficiency in our models should not fade with other local descriptors [98], augmenting the training set with systems displaying extended conjugation such as PAHs, fullerenes, etc, or improving upon the current baseline for Δ-ML should lead to better accuracies. This opens exciting possibilities of ML-guided analysis into nucleus independent chemical shifts complimenting the latest tight-binding model for PAHs [99]. Although our 100k training set is an adequate representation of the QM9 dataset, adaptive sampling method employed in [38] might be useful when using smaller training sets. Given the locality of the shielding property, it may be helpful to employ different machines [47] trained on *sp*, sp^2 and sp^3 C—to account for systematic deviations in each groups. Finally, one can always improve the QM9-NMR dataset by estimating the effects from geometries obtained at ωB97XD with triple-zeta quality basis set.

Data availability statement

The data that support the findings of this study are openly available in the MolDis repository, https://moldis.tifrh.res.in/data/QM9NMR. For further details see supplementary information. The same information may also be obtained from the authors through an email request.

Acknowledgments

The authors thank Vipin Agarwal, Kaustubh R Mote and O Anatole von Lilienfeld for fruitful discussions. We acknowledge support of the Department of Atomic Energy, Government of India, under Project Identification No. RTI 4007. All calculations have been performed using the Helios computer cluster, which is an integral part of the MolDis Big Data facility, TIFR Hyderabad (https://moldis.tifrh.res.in/).

ORCID iD

Raghunathan Ramakrishnan https://orcid.org/0000-0002-7288-9238

References

[1] Helgaker T, Jaszuński M and Ruud K 1999 *Chem. Rev.* **99** 293
[2] Mulder F A and Filatov M 2010 *Chem. Soc. Rev.* **39** 578
[3] Bagno A, Rastrelli F and Saielli G 2003 *J. Phys. Chem.* A **107** 9964
[4] Novotny J, Sojka M, Komorovsky S, Necas M and Marek R 2016 *J. Am. Chem. Soc.* **138** 8432
[5] Bifulco G, Gomez-Paloma L and Riccio R 2003 *Tetrahedron Lett.* **44** 7137
[6] Cimino P, Gomez-Paloma L, Duca D, Riccio R and Bifulco G 2004 *Magn. Reson. Chem.* **42** S26
[7] Seymour I D, Middlemiss D S, Halat D M, Trease N M, Pell A J and Grey C P 2016 *J. Am. Chem. Soc.* **138** 9405
[8] Bamine T *et al* 2017 *J. Phys. Chem.* C **121** 3219
[9] Molchanov S, Rowicki T, Gryff-Keller A and Koźmiński W 2018 *J. Phys. Chem.* A **122** 7832
[10] Guzzo R N, Rezende M J C, Kartnaller V, Carneiro J W d M, Stoyanov S R and da Costa L M 2018 *J. Mol. Struct.* **1157** 97
[11] Sarotti A M 2018 *Org. Biomol. Chem.* **16** 944
[12] Lodewyk M W, Siebert M R and Tantillo D J 2012 *Chem. Rev.* **112** 1839
[13] Grimme S, Bannwarth C, Dohm S, Hansen A, Pisarek J, Pracht P, Seibert J and Neese F 2017 *Angew. Chem, Int. Ed.* **56** 14763
[14] Buevich A V and Elyashberg M E 2018 *Magn. Reson. Chem.* **56** 493
[15] Lauro G, Das P, Riccio R, Reddy D S and Bifulco G 2020 *J. Org. Chem.* **85** 3297
[16] Keith T A and Bader R F 1992 *Chem. Phys. Lett.* **194** 1
[17] Keith T A and Bader R F 1993 *Chem. Phys. Lett.* **210** 223
[18] Mauri F, Pfrommer B G and Louie S G 1996 *Phys. Rev. Lett.* **77** 5300
[19] Gregor T, Mauri F and Car R 1999 *J. Chem. Phys.* **111** 1815
[20] Kutzelnigg W, Fleischer U and Schindler M 1990 *Deuterium and Shift Calculation* (Berlin: Springer) pp 165–262
[21] Ditchfield R 1972 *J. Chem. Phys.* **56** 5688
[22] Hinchliffe A 1987 *Ab Initio Determination of Molecular Properties* (Bristol: Adam Hilge)
[23] Gauss J 2000 *Modern Methods and Algorithms of Quantum Chemistry* **3** 541 (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.141.1291&rep=rep1&type=pdf)
[24] Mehring M 2012 *High Resolution NMR Spectroscopy in Solids, NMR Basic Principles and Progress* (Berlin: Springer)
[25] Price D R and Stanton J F 2002 *Org. Lett.* **4** 2809
[26] Flaig D, Maurer M, Hanni M, Braunger K, Kick L, Thubauville M and Ochsenfeld C 2014 *J. Chem. Theory Comput.* **10** 572
[27] Curtiss L A, Redfern P C and Raghavachari K 2011 *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1** 810
[28] Semenov V A, Samultsev D O and Krivdin L B 2019 *J. Phys. Chem.* A **123** 8417
[29] Wiitala K W, Hoye T R and Cramer C J 2006 *J. Chem. Theory Comput.* **2** 1085
[30] Adamo C and Barone V 1998 *J. Chem. Phys.* **108** 664
[31] Migda W and Rys B 2004 *Magn. Reson. Chem.* **42** 459
[32] Vázquez S 2002 *J. Chem. Soc. Perkin Trans.* **2** 2100
[33] Wiberg K B, Hammer J D, Zilm K W and Cheeseman J R 1999 *J. Org. Chem.* **64** 6394
[34] Wiberg K B, Hammer J D, Zilm K W, Keith T A, Cheeseman J R and Duchamp J C 2004 *J. Org. Chem.* **69** 1086
[35] Bassarello C, Cimino P, Gomez-Paloma L, Riccio R and Bifulco G 2003 *Tetrahedron* **59** 9555
[36] Sarotti A M and Pellegrinet S C 2009 *J. Org. Chem.* **74** 7254
[37] Sarotti A M and Pellegrinet S C 2012 *J. Org. Chem.* **77** 6059
[38] Gerrard W, Bratholm L A, Packer M J, Mulholland A J, Glowacki D R and Butts C P 2020 *Chem. Sci.* **11** 508
[39] Cobas C 2020 *Magn. Reson. Chem.* **58** 512–9
[40] Bratholm L A 2020 *et al* arXiv:2008.05994
[41] Rupp M, Ramakrishnan R and von Lilienfeld O A 2015 *J. Phys. Chem. Lett.* **6** 3309
[42] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 *Phys. Rev. Lett.* **108** 058301
[43] Gao P, Zhang J, Peng Q, Zhang J and Glezakou V-A 2020 *J. Chem. Inf. Model.* **60** 3746
[44] Ghosh K, Stuke A, Todorović M, Jooorgensen P B, Schmidt M N, Vehtari A and Rinke P 2019 *Adv. Sci.* **6** 1801367
[45] Westermayr J and Marquetand P 2020 *J. Chem. Phys.* **153** 154112
[46] Rankine C D, Madkhali M M and Penfold T J 2020 *J. Phys. Chem.* A **124** 4263–70
[47] Ramakrishnan R, Hartmann M, Tapavicza E and von Lilienfeld O A 2015 *J. Chem. Phys.* **143** 084111
[48] Xue B-X, Barbatti M and Dral P O 2020 *J. Phys. Chem.* A **124** 7199
[49] Westermayr J, Faber F A, Christensen A S, von Lilienfeld O A and Marquetand P 2020 *Mach. Learn.: Sci. Technol.* **1** 025009
[50] Pronobis W, Schütt K T, Tkatchenko A and Müller K-R 2018 *Eur. Phys. J.* B **91** 178
[51] Huo H and Rupp M 2017 (arXiv: 1704.06439) p 13754 (available at: www.researchgate.net/profile/Matthias_Rupp/publication/316429137_Unified_Representation_for_Machine_Learning_of_Molecules_and_Crystals/links/597ba6ff0f7e9b880293f6bf/Unified-Representation-for-Machine-Learning-of-Molecules-and-Crystals.pdf)
[52] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104** 136403
[53] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev.* B **87** 184115
[54] De S, Bartók A P, Csányi G and Ceriotti M 2016 *Phys. Chem. Chem. Phys.* **18** 13754
[55] Paruzzo F M, Hofstetter A, Musil F, De S, Ceriotti M and Emsley L 2018 *Nat. Commun.* **9** 1
[56] Chaker Z, Salanne M, Delaye J-M and Charpentier T 2019 *Phys. Chem. Chem. Phys.* **21** 21709
[57] Hartman J D, Kudla R A, Day G M, Mueller L J and Beran G J 2016 *Phys. Chem. Chem. Phys.* **18** 21686
[58] Faber F A, Christensen A S, Huang B and von Lilienfeld O A 2018 *J. Chem. Phys.* **148** 241717
[59] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2014 *Sci. Data* **1** 140022
[60] Dral P O, Owens A, Dral A and Csányi G 2020 *J. Chem. Phys.* **152** 204110
[61] Ramakrishnan R, Dral P O, Rupp M and von Lilienfeld O A 2015 *J. Chem. Theory Comput.* **11** 2087
[62] Ruddigkeit L, Van Deursen R, Blum L C and Reymond J-L 2012 *J. Chem. Inf. Model.* **52** 2864
[63] Faber F A *et al* 2017 *J. Chem. Theory Comput.* **13** 5255
[64] Ramakrishnan R and von Lilienfeld O A 2017 *Rev. Comput. Chem.* **30** 225
[65] Ramakrishnan R and von Lilienfeld O A 2015 *CHIMIA* **69** 182
[66] Mauri A, Consonni V and Todeschini R 2017 Molecular descriptors *Handbook of Computational Chemistry* ed J Leszczynski, A Kaczmarek-Kedziera, T Puzyn, M G Papadopoulos, H Reis and M K Shukla (Cham: Springer Int. Publishing) pp 2065–93
[67] Randić M 1996 *J. Math. Chem.* **19** 375

[68] Randić M 1997 *J. Chem. Inform. Comput. Sci.* **37** 672
[69] Pozdnyakov S N, Willatt M J, Bartók A P, Ortner C, Csányi G and Ceriotti M 2020 (arXiv: 2001.11696)
[70] von Lilienfeld O A, Ramakrishnan R, Rupp M and Knoll A 2015 *Int. J. Quantum Chem.* **115** 1084
[71] Todeschini R and Consonni V 2000 Methods and principles in medicinal chemistry *Handbook of Molecular Descriptors* vol 11, ed R Mannhold, H Kubinyi and H Timmerman (New York: Wiley-VCH) p 688
[72] Behler J and Parrinello M 2007 *Phys. Rev. Lett.* **98** 146401
[73] Behler J 2011 *J. Chem. Phys.* **134** 074106
[74] Engel E A, Anelli A, Hofstetter A, Paruzzo F, Emsley L and Ceriotti M 2019 *Phys. Chem. Chem. Phys.* **21** 23385
[75] Moussa J E 2012 *Phys. Rev. Lett.* **109** 059801
[76] Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O A, Müllerr K-R and Tkatchenko A 2015 *J. Phys. Chem. Lett.* **6** 2326
[77] Huang B and von Lilienfeld O A 2016 *J. Chem. Phys.* **145** 161102
[78] Pronobis W, Tkatchenko A and Müllerr K-R 2018 *J. Chem. Theory Comput.* **14** 2991
[79] Ditchfield R 1974 *Mol. Phys.* **27** 789
[80] Wolinski K, Hinton J F and Pulay P 1990 *J. Am. Chem. Soc.* **112** 8251
[81] Cheeseman J R, Trucks G W, Keith T A and Frisch M J 1996 *J. Chem. Phys.* **104** 5497
[82] Frisch M J *et al* Gaussian 16, Revision C.01 2016
[83] Stewart J J 2016 'Mopac2016' Stewart computational chemistry(Colorado Springs, CO) (available at: http://OpenMOPAC.net)
[84] Tomasi J, Mennucci B and Cammi R 2005 *Chem. Rev.* **105** 2999
[85] Chakraborty S, Kayastha P and Ramakrishnan R 2019 *J. Chem. Phys.* **150** 114106
[86] Fink T and Reymond J-L 2007 *J. Chem. Inf. Model.* **47** 342
[87] Blum L C and Reymond J-L 2009 *J. Am. Chem. Soc.* **131** 8732
[88] Blum L C, van Deursen R and Reymond J-L 2011 *J. Comput. Aided Mol. Des.* **25** 637
[89] Corey E J, Czakó B and Kürti L 2007 *Molecules and Medicine* (New York: Wiley)
[90] O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T and Hutchison G R 2011 *J. Cheminformatics* **3** 33
[91] Hanwell M D, Curtis D E, Lonie D C, Vandermeersch T, Zurek E and Hutchison G R 2012 *J. Cheminformatics* **4** 17
[92] Halgren T A 1996 *J. Comput. Chem.* **17** 490
[93] Himanen L, Jäger M O J, Morooka E V, Federici Canova F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 *Comput. Phys. Commun.* **247** 106949
[94] Christensen A S, Faber F A, Huang B, Bratholm L A, Tkatchenko A, Müller K-R, and von Lilienfeld O A 2017 QML: a python toolkit for quantum machine learning (available at: https://github.com/qmlcode/qml)
[95] Blackford L S *et al* 1997 *Scalapack Users' Guide* (Philadelphia, PA: Society for Industrial and Applied Mathematics)
[96] Krishnan S, Ghosh A, Gupta M, Kayastha P, Senthil S, Das S K, Kandpal S C, Chakraborty S, Gupta A, and Ramakrishnan R 2020 MolDis: a big data analytics platform for molecular discovery (available at: https://moldis.tifrh.res.in/)
[97] Molchanov S and Gryff-Keller A 2017 *J. Phys. Chem.* A **121** 9645
[98] Langer M F, Goeßmann A and Rupp M 2020 (arXiv: 2003.12081)
[99] Kilymis D, Bartók A P, Pickard C J, Forse A C and Merlet C 2020 *Phys. Chem. Chem. Phys.* **22** 13746–55