Non-covalent quantum machine learning corrections to density functionals

Pál D. Mezi and O. Anatole von Lilienfeld

Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Basel, Switzerland

(Dated: March 22, 2019)

We present non-covalent machine learning corrections to six physically motivated density functionals with systematic errors. We demonstrate that the missing physics can be recovered by the machine learning corrections. The models account for various types of non-covalent interactions, and enable accurate predictions of dissociation curves. The correction improves the description of molecular two- and three-body interactions crucial in large water clusters, and has atomic resolution. We show that given sufficient training instances the correction is more flexible than standard molecular mechanical dispersion corrections, and thus it can be applied for cases where many dispersion corrected density functionals fail, such as hydrogen bonding.

The efficient and accurate calculation of non-covalent interactions (exchange repulsion, electrostatics, induction, and dispersion [1]) is crucial in many practical problems in biochemistry, materials science, nanotechnology, and catalysis. However, the highly non-local nature of these interactions makes them difficult to capture by efficient computational methods such as many density functional (DF) approximations. [2-4] The leading terms in the long-range dispersion interaction energy can simply be described by the two-body London [5, 6] and three-body Axilrod-Teller [7, 8] potentials, and hence the missing dispersion interaction is often recovered by molecular mechanical dispersion corrections (e.g., TS [9, 10] or D3 [11, 12], or by nonlocal van der Waals DFs. [13-16] However, it is especially difficult to correct semi-local functionals simultaneously suffering from delocalization and dispersion errors at non-covalent overlapping electron density regions. [17-19] In such cases, one may consider more expensive fifth-rung functionals (on the Jacob’s ladder of DF approximations [20]) also containing information about the virtual orbitals. [21-27]

An alternative approach is to use efficient quantum machine learning (QML) models with representations uniquely describing the electronic structure problem to solve (for a given number of electrons without external fields). [28-30] Previously, such models were only applied for correcting approximate DF one- and two-body total energy errors on a set of water clusters and ice structures, or relative conformational enthalpy errors on a set of C7H10O2 isomers. [31, 32] However, no universal QML model has been presented yet for correcting simultaneously the inter- and intramolecular interaction energies of approximate DFs. Here, we propose non-covalent interaction (NCI) corrections from QML with high level of flexibility to correct the approximate DF description of intermolecular interactions, going beyond pure dispersion. We demonstrate that the missing physical effects can be learned by these NCI models. We analyze the errors of the resulting models from multiple aspects, and compare them to similarly efficient dispersion corrected functionals. Note that other machine learning (ML) models were also applied to improve upon approximately quantum chemical non-covalent interactions, [33, 34] but they generally lack uniqueness in the representation and thus can yield absurd results. [35]

In this paper, we use KRR (in the QML code [36]) with Gaussian kernel and the recently introduced representation of Faber et al. (FCHL18). [37] We perform a two-step procedure to decompose the correction into atomic contributions. In the first step, we determined the regression coefficient vector \( \alpha \),

\[
\alpha = (K + \lambda I)^{-1} y
\]

where \( K \) is a local symmetric kernel matrix with the matrix elements of \( \sum_{i,j} k(x_i^j, x_j^i) \) using the \( i \)th atomic elements of the \( i \)th training systems (molecular dimers and monomers) from the \( \mathbf{x} \) array of representation vectors, \( \lambda \) is the regularization strength (chosen to be \( \lambda = 10^{-8} \)) due to the very low level of noise in the well-converged CCSD(T) reference energies; for more information see the supplemental material), \( I \) is the identity matrix, and \( y \) is the training output vector. In the second step, we predict atomic contributions \( \tilde{y}_j^i \) for the \( i \)th atom of the \( j \)th test system (as well as their sum),

\[
\tilde{y}_j^i = \sum_j \alpha_j \sum_{j' \neq j} k(\tilde{x}_j^i, x_j^j)
\]

where the atomic kernel matrix elements \( k(\tilde{x}_j^i, x_j^j) = \exp(-\frac{\|\tilde{x}_j^i - x_j^j\|^2}{2\sigma^2}) \) (with an optimal hyperparameter \( \sigma = 2^8 \)) are calculated from the training \( \mathbf{x} \) and test \( \tilde{\mathbf{x}} \) arrays of atomic representation vectors.

As baseline methods, we selected from each of the second-, third-, and fourth-run usually more repulsive (BLYP, [38, 39] TPSS, [40] B3LYP [41]) and a usually less repulsive (PBE, [42] SCAN, [43, 44] PBE0 [45]) functional with physically motivated forms and thus expectedly with systematic errors. We also expect better description of the midrange correlation from the less repulsive functionals, and reduced self-interaction error from...
functionals on higher rungs. In order to be able to generalize our correction to also improve intramolecular interactions, we trained the machine learning model on atomization energies, and derived the interaction energy correction from the atomization energy corrections for monomers and dimer/cluster.

FIG. 1. a) Learning curves (from 100 random cross validation runs adequate for varying training/test proportion) for the direct ML and NCI models on S66x8 using N and (66-N) potential curves in training and test set, respectively. Overall mean absolute errors (weighted by the number of equilibrium dimers/clusters in the datasets) on a broader dataset training the NCI models only on S66x8 (small) and on the broader dataset (large).

First, we present learning curves (FIG. 1a) for training the NCI models on the biochemically relevant S66x8 set. As expected,[32] the learning curves of the NCI models run more or less parallel shifted downwards by a factor depending on the inherent accuracy of the baseline methods compared to the learning curve corresponding to the direct ML model. The learning onset is reached after including 5 potential curves into the training set. Despite the already promising overall accuracy of the NCI models trained on S66x8, using only S66x8 in the training set leads to less transferable models biased towards small dimers with H, C, N, and O atoms and few functional groups.

To achieve a more transferable model, we extended our training set by further potential curves and water clusters (FIG. 1b, for more details see supplementary material). We will use this extended training set hereafter.

In a first step, we compare the accuracy of the SCAN+NCI interaction energies to the accuracy of the corresponding D3 dispersion corrected interaction energies [12] with Becke-Johnson (BJ) damping and three-body terms (TABLE I, more details in the supplemental material). The NCI correction usually works better than the D3 correction for small dimers and even for large water clusters, but it is worse for large host-guest complexes and molecular crystals. The success of the NCI corrected schemes on water clusters suggests that the NCI models can correct errors also other than dispersion. For these neutral water clusters, our NCI correction is on par with the most accurate dispersion corrected DF[6](MAE: 0.94 kcal mol$^{-1}$ with PW6B95-D3(0)) [11] for the literature.

The NCI correction performs well for various types of interactions (FIG. 2). The SCAN+NCI interaction energies are more accurate than the SCAN+D3 ones for hydrogen-bonding and dispersion but less accurate for dipole-dipole interaction and induction. Note that hydrogen bonds can be recognized by three atoms, while the dipole moment is a global property of molecules that is poorly described by local representations, as already discussed in refs.[37, 65] In the other cases, the two correction schemes work similarly.

In a second step, we have examined how the NCI model works at non-equilibrium intermolecular distances (FIG. 3). The NCI correction works well at various relative intermolecular distances (defined as the actual ($r$) over equilibrium ($r_e$) distance between the two monomer centers of mass). While the D3 correction has difficulties at shorter relative intermolecular distances (especially for hydrogen-bonding), where density functionals often overestimate the magnitude of the interaction because of their delocalization error.

The SCAN+NCI seeming molecular two-body $C_6$ dispersion coefficients (fitting a $-C_6/r^6$ function to the three longest intermolecular distances, the actual dispersion coefficients may be smaller) of the London dispersion dominated S66x8 hetero- and homodimers are in good agreement with the reference (FIG. 4). Even though, the NCI correction does not contain explicitly information neither about the long-range behavior of dispersion nor about the values of the dispersion coefficients, it can

TABLE I. Mean absolute errors (kcal mol$^{-1}$) in the interaction energies for the different test sets with our most robust NCI approach or D3 corrections.

| Dataset                  | SCAN+NCI | D3 |
|--------------------------|----------|----|
| Biochemically relevant   | 0.70     | 0.14 |
| Non-covalent blind test  | 0.40     | 0.25 |
| Halogen-containing       | 0.49     | 0.24 |
| Water clusters           | 4.85     | 0.94 |
| Molecular crystals       | 4.69     | 2.38 |
| Host-guest complexes     | 9.98     | 3.66 |

$^a$ Not represented in the training set.
FIG. 2. Mean absolute errors (kcal mol$^{-1}$) the pure, NCI, and D3 corrected SCAN methods on various types of interactions in biochemically relevant and halogen-containing (X: halogen atom) dimers.\cite{47, 51}

FIG. 3. Mean absolute errors (kcal mol$^{-1}$) with respect to the relative intermolecular distance on the S66x8 database for the pure, NCI, and D3 corrected SCAN methods. The potential energy curves for a water-peptide dimer can be compared to the CCSD(T) reference in the inset. (The arrows represent the physical effects missing from the D3 but described by the NCI correction.)

lead to a similarly accurate and meaningful description of long-range dispersion as the D3 correction.

In a third step, the molecular many-body decomposition of the NCI corrected interaction energies on the water icosamers. The results in FIG. 5 show that the largest correction appears in the two-body term. The NCI model goes beyond the atom-pairwise picture, which provides more flexibility to properly account for the accurate description of the interaction energy in water clusters. The correction does not affect the four-body and higher-order terms, because the FCHL representation contains information only up to atomic three-body. Hence the good overall accuracy relies on an error cancellation between

FIG. 4. Predicted versus reference seeming molecular two-body $C_6$ dispersion coefficients (kcal mol$^{-1}$) of S66x8 dimers with London dispersion for the pure, NCI, and D3 corrected SCAN methods. The potential energy curves for the cyclopentane-neopentane dimer can be compared to the CCSD(T) reference (black dashed line) in the inset.

FIG. 5. Predicted versus reference CCSD(T) molecular two-body interaction energy terms (kcal mol$^{-1}$) of the water icosamers for the pure, NCI, and D3 corrected SCAN methods. To provide context, the molecular one-, three-, and higher many-body interaction energy terms are shown in the inset.
Finally, we have also extracted atomic level information from the NCI models by exploiting the atomic kernel matrix in the prediction. The atomic contributions (between an atom in monomer A and all the atoms in monomer B) to the NCI interaction energy correction (without monomer relaxation) in large host-guest complexes (FIG. 6) correlate well with the pairwise atomic contributions to the D3 correction ($\Delta E^I = \frac{1}{2} \sum_{J \in j} E_{IJ}^6$) for smaller interaction energy contributions. The randomness in the scatter plot of the NCI versus D3 atomic contributions is higher, however, for larger interaction energy contributions. (Note that the atomic contributions to the two corrections do not have to agree necessarily since the atomic decomposition of the interaction energy is ambiguous. Also note that the host-guest complexes in FIG. 6 were not represented in the training set.) In general, the NCI model places the attractive correction on the more polarizable (non-hydrogen) atoms, or on atoms involved in $\pi$ delocalization, and the repulsive (or less attractive) correction on atoms of different monomers close to each other, or on atoms involved in hydrogen bonding.

We applied QML to correct the non-covalent interactions in various van der Waals complexes calculated by efficient DF approximations. Learning curves demonstrate that the NCI corrections can complement well approximate DFs for various types of interactions and at various relative intermolecular distances. The correction captures physical effects missing from the DF methods, has an effect on the first three many-body interaction energy terms, and distributes the interaction energy terms reasonably on the atoms of monomers. Among the examined NCI models, the most robust is based on the SCAN functional. The NCI correction is more flexible than standard molecular mechanical dispersion corrections, and it can capture more of the missing physical effects than the generally used D3 correction. As evinced, the NCI correction is more accurate than conventional dispersion corrections for example when hydrogen bonding dominates the interaction. Furthermore, its error will decrease systematically if more training data were provided thanks to the generalization power of QML models. To facilitate applications, we make our machine learning models available in the supplementary material. Future improvements of the NCI corrections will be possible as soon as more high-level reference non-covalent interaction energies become available. Another possible extension of this work could be its extension within the recently introduced multi-level combination technique.

PDM thanks Jan Gerit Brandenburg for providing data and information about the calculations on the S66x8 molecular dimers and the X23 molecular crystals. This research was supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No 772834.

[1] B. Jeziorski, R. Moszynski, and K. Szalewicz, Chem. Rev. 94, 1887 (1994).
[2] S. Kristyany and P. Pulay, Chem. Phys. Lett. 229, 175 (1994).
[3] P. Hobza, T. Reschel, et al., J. Comput. Chem. 16, 1315 (1995).
[4] J. Pérez-Jordá and A. D. Becke, Chem. Phys. Lett. 233, 134 (1995).
[5] F. London, Trans. Faraday Soc. 33, 8b (1937).
[6] H. B. G. Casimir and D. Polder, Phys. Rev. 73, 360 (1948).
[7] B. Axilrod and E. Teller, J. Chem. Phys. 11, 299 (1943).
[8] M. Aub, Proc. R. Soc. London, Ser. A 257, 464 (1960).
[9] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. 102, 073005 (2009).
[10] O. A. von Lilienfeld and A. Tkatchenko, J. Chem. Phys. 132, 234109 (2010).
[11] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. 132, 154104 (2010).
[12] S. Grimme, S. Ehrlich, and L. Goerigk, J. Chem. Phys. 32, 1456 (2011).
[13] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. 92, 246401 (2004).
[14] K. Lee, E. D. Murray, L. Kong, B. I. Lundqvist, and D. C. Langreth, Phys. Rev. B 82, 081101 (2010).
[15] O. A. Vydrov and T. Van Voorhis, Phys. Rev. Lett. 103, 073005 (2009).
K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, and M. Wang, “Molpro, version 2012.1, a package of ab initio programs,” (2012), see.

[70] A. Bergner, M. Dolg, W. Küchle, H. Stoll, and H. Preuß, Mol. Phys. 80, 1431 (1993).

[71] H. Stoll, B. Metz, and M. Dolg, J. Comput. Chem. 23, 767 (2002).

[72] M. Kallay, Z. Rolik, J. Csontos, P. Nagy, G. Samu, D. Mester, J. Csoka, B. Szabo, I. Ladjanszki, L. Szegedy, B. Ladoczki, K. Petrov, M. Farlas, P. D. Mezei, and B. Hegely, “Mrcc, a quantum chemical program suite,” See also Z. Rolik, L. Szegedy, I. Ladjanszki, B. Ladoczki, and M. Kallay, J. Chem. Phys. 139, 094105 (2013), as well as: www.mrcc.hu.

[73] M. A. Marques, M. J. Oliveira, and T. Burnus, Comput. Phys. Commun. 183, 2272 (2012).

[74] F. Weigend and R. Ahlrichs, Phys. Chem. Chem. Phys. 7, 3297 (2005).

[75] J. G. Brandenburg, J. E. Bates, J. Sun, and J. P. Perdew, Phys. Rev. B 94, 115144 (2016).

[76] T. H. Dunning, J. Chem. Phys. 90, 1007 (1989).

[77] J. Kresse, G.; Furthmüller, Phys. Rev. B 54, 11169 (1996).

Supplemental material for “Non-covalent quantum machine learning corrections to density functionals”

Pál D. Mezei and O. Anatole von Lilienfeld

Here, we provide more details about the accuracy of the training and test energies. Our extended training set contains the potential curves from the biologically relevant S66x8 database \[47\] with 0.05 kcal mol\(^{-1}\) uncertainty, the blind test (BT) of Taylor et al. \[48\] with 0.04 kcal mol\(^{-1}\) uncertainty, and the halogen-containing X40x10 database \[51\] with 0.04 kcal mol\(^{-1}\) uncertainty, as well as the neutral water clusters from the BEGDB \[52, 67\] and GMTKN30 \[53, 64\] databases with less than 0.01 kcal mol\(^{-1}\) uncertainty up to the decamers, and with 0.4 kcal mol\(^{-1}\) uncertainty for the iicosamers due to the basis set superposition error. \[55\]

We also tested our models on the per monomer lattice energies X23 molecular crystals \[56\] with 0.7 kcal mol\(^{-1}\) uncertainty, and on the interaction energies of five very large neutral host-guest (HG) complexes (TCNQ@tweezer: tetracyanoquinone-tweezer; DCB@tweezer: 1,4-dicyanobenzene-tweezer; C\(_60\)@catcher: buckyball-catcher; GLH@mcycle: glycine anhydride-macrocycle; ADOH@CB7: 1-hydroxyadamantane-cucurbit[7]uril) from the S12L set \[61\] with 1.1 kcal mol\(^{-1}\) uncertainty. \[68\]

We took the reference interaction energies from the literature and calculated (in Molpro \[69\]) the monomer atomization energies using the method introduced by Brauer et al. \[47\] (with the ECP28MWB and ECP46MDF effective core potentials \[70, 71\] for the bromine and iodine atoms) converged to the CCSD(T) basis set limit.

For baseline, we computed converged DF interaction energies. In the molecular calculations (in MRCC with LibXC \[72, 73\]), we applied the aug-def2-QZVP(-g) basis set \[74\] for almost all the systems (with def2-ECP-28 \[74\] for iodine atoms). The effect of diffuse functions is around 0.1 kcal mol\(^{-1}\) for the S66x8 set comparing our results to the ones published by Brandenburg et al. \[75\]. While the neglect of the g functions caused only 0.06 kcal mol\(^{-1}\) uncertainty.

For water clusters, we expect larger basis set effects, thus we used the aug-def2-QZVP basis set for interactions weaker than 100 kcal mol\(^{-1}\) and the aug-cc-pV5Z \[76\] basis set for interactions stronger than 100 kcal mol\(^{-1}\). For the large host-guest complexes, we accelerated the hybrid DF calculations of the large host-guest complexes applying the def2-QZVP(-g) basis set. \[74\] because the effect of diffuse functions on the more compact densities is less pronounced.

In the periodic calculations (using the VASP quantum chemistry software \[77\]), we obtained converged interaction energies from the procedure described in ref \[56\] on the PBE-TS geometries of Reilly and Tkatchenko. \[57\] The average uncertainty caused by the geometry is about 0.4 kcal mol\(^{-1}\) comparing our results to the ones pub-
lished by Brandenburg et al. [75].

Finally, we also present a comparison of all the examined pure, NCI, and D3 corrected DF methods in supplementary TABLE II and a more detailed comparison of the NCI models in supplementary TABLE III.

TABLE II. Mean absolute errors (kcal mol$^{-1}$) in the interaction energies for the different test sets with various baseline methods and with NCI or D3 corrections.

| Dataset | S66x8 | BT X40x10 (H$_2$O)$_n$ | X23 | HG |
|---------|-------|------------------------|-----|----|
| BLYP    | 3.10  | 2.03                   | 2.40| 9.12|17.96|39.59 |
| +NCI    | 0.13  | 0.24                   | 0.29| 1.19|8.75 |3.73  |
| +D3     | 0.19  | 0.33                   | 0.38| 3.43|3.73 |6.07  |
| PBE     | 1.62  | 1.13                   | 1.04| 1.67|11.64|25.51 |
| +NCI    | 0.11  | 0.24                   | 0.21| 1.05|3.58 |6.43  |
| +D3     | 0.25  | 0.30                   | 0.37| 7.07|1.31 |0.94  |
| TPSS    | 2.24  | 1.45                   | 1.50| 4.14|15.12|31.33 |
| +NCI    | 0.11  | 0.20                   | 0.24| 0.80|3.99 |8.24  |
| +D3     | 0.24  | 0.31                   | 0.33| 4.19|1.30 |3.35  |
| SCAN    | 0.70  | 0.40                   | 0.49| 4.85|4.69 |9.98  |
| +NCI    | 0.14  | 0.25                   | 0.24| 0.94|2.38 |3.66  |
| +D3     | 0.25  | 0.16                   | 0.36| 6.89|2.20 |1.35  |
| B3LYP   | 2.48  | 1.59                   | 1.93| 6.25| -   |33.55 |
| +NCI    | 0.11  | 0.19                   | 0.24| 1.03| -   |4.93  |
| +D3     | 0.15  | 0.22                   | 0.29| 3.69| -   |4.93  |
| PBE0    | 1.55  | 1.02                   | 1.08| 1.69| -   |23.38 |
| +NCI    | 0.10  | 0.19                   | 0.18| 0.95| -   |8.14  |
| +D3     | 0.23  | 0.26                   | 0.30| 4.38| -   |2.58  |
TABLE III. Mean absolute errors (kcal mol$^{-1}$) in the interaction energies on various test sets for the direct ML and NCI models using different baseline methods.

| Test set                  | Direct | BLYP | PBE | TPSS | SCAN | B3LYP | PBE0 |
|---------------------------|--------|------|-----|------|------|-------|------|
| S66x8 all                 | 0.84   | 0.13 | 0.11| 0.11 | 0.14 | 0.11  | 0.10 |
| S66x8 hydrogen bonding    | 0.99   | 0.12 | 0.10| 0.10 | 0.14 | 0.09  | 0.10 |
| S66x8 $\pi$-stacking      | 0.86   | 0.16 | 0.10| 0.12 | 0.12 | 0.13  | 0.10 |
| S66x8 London dispersion   | 0.96   | 0.16 | 0.13| 0.11 | 0.17 | 0.12  | 0.10 |
| S66x8 mixed               | 0.56   | 0.11 | 0.10| 0.09 | 0.12 | 0.09  | 0.10 |
| Blind test all            | 1.25   | 0.24 | 0.24| 0.20 | 0.25 | 0.19  | 0.19 |
| X40x10 all                | 1.14   | 0.29 | 0.21| 0.24 | 0.24 | 0.24  | 0.18 |
| X40x10 London dispersion  | 0.58   | 0.20 | 0.17| 0.16 | 0.20 | 0.14  | 0.13 |
| X40x10 induction          | 1.60   | 0.24 | 0.24| 0.23 | 0.13 | 0.13  | 0.13 |
| X40x10 dipole-dipole      | 1.83   | 0.55 | 0.20| 0.29 | 0.24 | 0.45  | 0.45 |
| X40x10 $\pi$-stacking     | 1.93   | 0.17 | 0.19| 0.28 | 0.26 | 0.21  | 0.21 |
| X40x10 halogen bonding    | 0.76   | 0.32 | 0.19| 0.24 | 0.29 | 0.27  | 0.19 |
| X40x10 halogen-$\pi$      | 0.53   | 0.31 | 0.27| 0.27 | 0.21 | 0.28  | 0.27 |
| X40x10 hydrogen bonding   | 1.64   | 0.28 | 0.22| 0.25 | 0.24 | 0.23  | 0.19 |
| X23 all                   | 36.58  | 8.75 | 3.58| 3.99 | 2.38 | -     | -    |
| X23 subset\(^{a}\)        | 16.69  | 4.64 | 1.47| 1.42 | 2.03 | -     | -    |
| X23 dispersion            | 56.19  | 16.48| 5.79| 7.35 | 2.67 | -     | -    |
| X23 hydrogen bonding      | 21.49  | 2.80 | 1.89| 1.40 | 2.15 | -     | -    |
| Host-guest all            | 34.61  | 3.73 | 6.43| 8.24 | 3.66 | 4.93  | 8.14 |

\(^{a}\) defined by Brandenburg et al. [75]