PM6-ML: The Synergy of Semiempirical Quantum Chemistry and Machine Learning Transformed into a Practical Computational Method

Martin Nováček and Jan Řezáč*

Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, 160 00 Prague, Czech Republic

***e-mail: rezac@uochb.cas.cz**

August 9, 2024

Abstract

Machine learning (ML) methods offer a promising route to the construction of universal molecular potentials with high accuracy and low computational cost. It is becoming evident that integrating physical principles into these models, or utilizing them in a Δ-ML scheme, significantly enhances their robustness and transferability. This paper introduces PM6-ML, a Δ-ML method that synergizes the semiempirical quantum-mechanical (SQM) method PM6 with a state-of-the-art ML potential applied as a universal correction. The method demonstrates superior performance over standalone SQM and ML approaches and covers a broader chemical space than its predecessors. It is scalable to systems with thousands of atoms, which makes it applicable to large biomolecular systems. Extensive benchmarking confirms PM6-ML's accuracy and robustness. Its practical application is facilitated by a direct interface to MOPAC. The code and parameters are available at `https://github.com/Honza-R/mopac-ml`.

1 Introduction

Machine learning (ML) methods represent a promising avenue for the construction of universal molecular potentials with high accuracy and favorable computational cost. However, it is becoming increasingly evident that incorporating some physics into the model confers advantages, rendering the model more robust and reducing the amount of data required for training. This approach can be taken to the extreme, whereby machine learning can be applied on top of a complete but approximate computational method and trained to reproduce more accurate calculations. This approach, designated delta-machine learning (Δ-ML), is particularly well-suited for enhancing the precision of computationally inexpensive yet broadly applicable computational chemistry methods.

The same topic can be viewed from an alternative perspective. Semiempirical quantum-mechanical (SQM) methods[1,2] offer a distinctive combination of universal applicability, rooted in a solid physical background, and of favorable computational efficiency. However, this efficiency is achieved by introducing approximations that limit the accuracy of the method. We, along with other researchers, have been engaged in efforts to address some of these limitations, particularly those related to the description of non-covalent interactions (NCIs),[3–8] with the goal of developing methodology applicable to large systems such as biomolecules.[9–11] These corrections, along with novel methods incorporating them, have led to a notable advancement over the previous state of the art. However, there remain challenges that cannot be resolved through this approach.

This work builds on our experience with the development of corrections for SQM methods. Instead of using additional corrections addressing specific phenomena, we have constructed a Δ-ML model combining PM6,[12] an universal SQM method, with state-of-the-art ML potential applied as correction and trained to reproduce high-quality DFT calculations. The resulting method, named PM6-ML, is shown to outperform both standalone SQM and ML methods. In comparison to earlier SQM-based Δ-ML methods, this new approach offers greater accuracy and, for the first time, covers a wider chemical space, what renders it applicable to real-world chemical problems.

The selection of PM6 as the baseline method was driven by its inherent properties and

the availability of a linear-scaling implementation,[13] which allows it to be applied to very large molecular systems. PM6 is a classical SQM method based on the NDDO (neglect of diatomic differential overlap) approximation,[14] which is universally applicable to a broad chemical space.[12] It served as the main platform for the development of our earlier corrections for non-covalent interactions, and with these, as PM6-D3H4X,[4,5] it is one of the most accurate SQM methods in our primary area of applications, namely biomolecules. These findings were validated by extensive benchmarking, which simultaneously revealed the method's most significant deficiencies. In particular, the PM6-D3H4X SQM method exhibited poor description of non-covalent interactions at very short distances[15,16] and significant errors in relative energies of conformers,[17] which are limitations shared by other SQM methods. Despite our best efforts, we were unable to identify a satisfactory solution to these issues through additional corrections or reparametrization of PM6 itself. This ultimately led us to pursue the ML correction as a potential solution.

Since our goal is to reproduce high quality reference data, the ML potential used as a correction must be able to achieve high accuracy, but it must also be data efficient, since the reference computations themselves are quite demanding. Both requirements are met by the Equivariant Transformer (ET) models, which represent the current state of the art in ML potentials. Among the few implementations of ETs available at the time we started this project, we had chosen the TorchMD-NET/ET potential[18] because it had already demonstrated its applicability as a stand-alone ML potential covering the same chemical space we were targeting (note that the potential is labeled "TorchMD-NET/ET" for the purpose of distinguishing it from the TorchMD-NET framework, which also implements other models).

The primary data utilized for the training of the correction is the SPICE database,[19] which offers comprehensive coverage of biomolecules, organic compounds, and ions comprising 15 elements (H, C, N, O, P, S, F–I, Li–K, Mg, Ca). This represents a significant advantage over the preceding SQM-based Δ-ML approaches,[20,21] which are applicable to only four elements (H, C, N, O). Another advantage of the SPICE database is that it has been computed at a very high level using one of the top-performing DFT function-

als, ωB97M-D3BJ,[22] in a large def2-TZVPPD basis set.[23] We complemented the SPICE dataset by additional systems covering non-covalent interactions taken from the NCIAtlas database.[16,24–27]

The development of the method is supported by extensive benchmarking, including comparisons with previous Δ-ML approaches, a range of SQM methods, and multiple standalone ML potentials. The methods were evaluated using a diverse collection of datasets that encompassed the intended applications to organic and biomolecular systems. Furthermore, selected methods were evaluated in a realistic scenario derived from previous research on protein-ligand interactions.

The results presented here demonstrate that the PM6-ML method outperforms both the components it is constructed from – namely, SQM calculations or machine learning alone. The machine learning is able to correct errors in the SQM calculations with unprecedented accuracy. Conversely, the solid physical basis lends the resulting method robustness, which is difficult to achieve with machine learning only. This is demonstrated mainly by the excellent transferability from the small systems used in the training to much larger ones that are important for practical applications.

In order to facilitate the use of PM6-ML, we have implemented a direct interface between the ML correction and MOPAC, which is the leading software in the field of SQM calculations. The code, as well as the parameters for the ML model, are available at `https://github.com/Honza-R/mopac-ml`.

2 Methods

2.1 PM6 and Corrections for Non-Covalent Interactions

The foundation of PM6-ML is the classical semiempirical method PM6, which is based on the NDDO approximation. There were two primary reasons for utilizing a method from this class: first, these methods are highly robust, exhibiting no convergence issues in large systems, which is a common occurrence in density functional tight binding methods. Secondly, when computed in the MOPAC software, they can be coupled with the MOZYME

linear scaling algorithm, which makes them faster than any competing method in very large systems (thousands of atoms). Moreover, we have extensive experience with PM6 from our previous work.

PM6-ML is intended as a replacement for the empirical corrections for SQM methods that we developed previously. It will therefore be compared to the most recent version of PM6 with corrections for London dispersion, hydrogen, and halogen bonds, PM6-D3H4X.[4,5] Specifically, the H4 and X corrections, with parameters updated in Refs. 24 and 26, and the additional repulsive correction introduced in Ref. 16, are employed throughout the paper. This method is designated PM6-D3H4X' to differentiate it from the default version.

All the PM6 calculations presented in this work were conducted using MOPAC.[28]

2.2 PM6-ML Correction

The PM6-ML method combines unmodified PM6, the D3 dispersion correction, and the machine learning potential, repurposed to serve as a short-ranged correction for the remaining errors. The ML correction (ΔE_{ML}) is trained to reproduce the difference between the DFT reference and PM6, both without the D3 dispersion correction. The dispersion correction ΔE_{D3}, as parametrized for the DFT reference, is then added back, so that the final PM6-ML energy is assembled as:

$$E_{PM6-ML} = E_{PM6} + \Delta E_{D3} + \Delta E_{ML} \tag{1}$$

The ΔE_{ML} correction is based on the TorchMD-NET/ET ML potential, as detailed in Ref. 18. This potential employs the equivariant transformer architecture,[29,30] affording it significant advantages. The appropriate handling of spatial symmetries enables the training of more robust potentials with only moderate requirements on the size of the training dataset. In contrast to the original setup, which was intended as a standalone potential, we reduce the radial cutoff defining the environment of each atom to 5 Å and use only a single atom type for each element, regardless of the charge of the atom. This is due to the fact that the electrostatics is handled by the underlying SQM calculation. Also, in cases

where a correction is required at the short range, the charge information can be inferred from the environment of the atom. Finally, the original TorchMD-NET/ET model exhibited a discontinuity in the potential at the cutoff distance. This was fixed in collaboration with the authors of the TorchMD-NET software (see the Acknowledgements) and all the models presented here are free of this error.

The D3 dispersion correction in the ωB97M-D3BJ functional employs the Becke-Johnson damping with parameters $s_8 = 0.3908$, $a_1 = 0.566$, and $a_2 = 3.128$, with these values being utilized here without modification. In addition, the three-body dispersion term is included in the calculation of ΔE_{D3}, which is not used in the ωB97M-D3BJ functional. This term is negligible in the small molecules used for training, but it should improve the description of larger systems where PM6-ML will likely be used.

2.3 Training Data

The training data were obtained from two sources - the SPICE[19] and the Non-Covalent Interaction Atlas[16,24–27] (NCIAtlas) databases. The SPICE database (version 1.1.2), comprising approximately 1.1 million conformations of small molecules (including peptides, amino acids, drug-like molecules, and ions) and their non-covalent complexes, serves as the backbone of the training data. To enhance the representation of a broader range of non-covalent interactions, the NCIAtlas datasets were further added to the training data. The NCIAtlas is comprised of seven datasets that map disparate classes of non-covalent interactions within an expanded chemical space, and it provides multiple points along the dissociation curve of each complex. From each of these datasets, 50 systems were removed for subsequent use in the validation dataset (see Section 2.5 for details). Additionally, systems that fell outside the PM6-ML chemical space were excluded. The final composition of the training data is presented in Table 1.

The methods presented here are trained to reproduce the reference DFT atomization energies, $\Delta_{at}E_{DFT\text{-}D3}$. Furthermore, the associated DFT gradients are also used in the training, providing a substantial additional data that characterize the potential energy surface of the molecules. The SPICE database provides both the energies and gradi-

ents computed at the ωB97M-D3BJ/def2-TZVPPD level.[19] The NCIAtlas datasets were calculated at the same level in Orca, version 5.0.3.[31,32]

The PM6-ML correction is trained to reproduce the difference between the DFT and PM6-ML atomization energies and gradients. For this purpose, the whole training set was recalculated with PM6 in MOPAC. The training data are then constructed as a difference between the DFT and PM6 atomization energies. Since the D3 dispersion correction is already included in the DFT results, the same D3 correction is added to the PM6 side (see the previous section, denoted as PM6-D3) prior this subtraction. The quantity used for the training of the PM6-ML correction, denoted as ΔE_{train}, is thus defined as:

$$\Delta E_{train} = \Delta_{at} E_{DFT\text{-}D3} - \Delta_{at} E_{PM6\text{-}D3} \tag{2}$$

and its gradient is constructed in the same way.

Table 1: The training data employed in the development of PM6-ML comprises a combination of the SPICE and NCIAtlas databases.

Database	Dataset	Molecules	Conformations
SPICE	Dipeptides	677	33850
	Solvated Amino Acids	26	1300
	DES370K	3864	364376
	PubChem	14643	731856
	Ion Pairs	28	1426
NCIAtlas	D1200	752	752
	D442×10	230	2300
	HB300SPX×10	250	2500
	HB375×10	325	3250
	IHB100×10	50	500
	Rep739×5	504	2520
	SH250×10	128	1280
Total		21477	1145910

2.4 Training Procedure

The PM6-ML correction, as well as a reparametrization of the standalone TorchMD-NET/ET potential (see Section 2.7 below), had been trained using the "torchmd-train"' tool from the TorchMD-NET project.[33] This tool employs the "Trainer" module of Py-

Torch[34] to perform the optimization of the model.

The hyperparameters utilized for training are derived from those used to train the original TorchMD-NET/ET model on the SPICE dataset, as provided in the TorchMD-NET GitHub repository.[33] However, a few modifications have been made: The majority of the training was performed on RTX 3090 graphics cards, which limited the batch size to 62. The upper cutoff was reduced to 5 Å, as the model is mainly meant to correct shorter-range errors. For both the PM6-ML correction and the standalone ML potential, 40 models were trained starting from different randomized initial parameters (keeping all the hyperparameters but the random seed the same).

All models were trained until convergence, as defined by the hyperparameters. Furthermore, we verified that continuing the training beyond this point did not result in any noticeable improvements. It is important to note, however, that the quality of the trained models exhibited significant variation depending on the initial conditions (defined by the random seed used to generate the starting point). Consequently, the performance of all resulting models was subsequently analyzed in more detail, and the best candidates were selected as described in the Results and Discussion, Section 3.1.

2.5 Validation Datasets

The PM6-ML method is benchmarked and compared to other Δ-ML, SQM, and ML approaches using multiple validation datasets covering its primary area of applications. These include non-covalent interactions and conformation energies of organic compounds and biomolecules.

Non-covalent interactions. The first part of the validation set comprises subsets of all the NCIAtlas datasets, specifically D442, HB375, HB300SPX, R739, SH250, and IHB100, which have been excluded from the training set. It should be noted that, in contrast to the training phase, only the equilibrium geometries are employed in the validation step. This is indicated by dropping the suffix "×..." from the names of the datasets. For each of the aforementioned datasets, the validation subset is constructed by taking the predefined subset of the 50 most diverse systems (obtained by a clustering analysis, as

reported in the original publications) and removing systems containing elements outside the PM6-ML chemical space. This results in a reduction of the number of systems to 26 in D442, 22 in R739, and 43 in SH250. These datasets facilitate the interpretation of the results in terms of different classes of non-covalent interactions and the chemical composition of the systems. In some tests, these datasets had been complemented by the widely used, but less diverse S66 set. In all these datasets, the benchmark interaction energies were computed at the coupled cluster with single, double, and perturbative triple excitations (CCSD(T)) level and extrapolated to the complete basis set (CBS) limit. The interaction energies are computed on the fixed structure of the complex, and thus do not include the deformation energy.

Non-covalent interactions in large systems. As the developed method is intended for applications to large molecular systems, it is essential to validate its accuracy in such context. To this end, we employ a series of datasets of increasing size. First, for the widely used L7 and S12L datasets,[35,36] we utilize the highest-quality benchmark interaction energies, computed using domain-based local pair natural orbital coupled clusters (DLPNO-CCSD(T)), extrapolated to the complete basis set (CBS) limit.[37] This benchmark is only available for six systems from the S12L set (2a, 2b, 4a, 5a, 6a, and 7b), and these are the only ones used. The systems in question range in size from 72 to 177 atoms. Secondly, the PLA15 set, as outlined in reference 9, is employed, comprising fifteen protein-ligand complex models. There, the DLPNO-CSCD(T) benchmark is constructed from fragment-based calculations, and the systems comprise 283 to 584 atoms. In certain instances, we also examine these fragments, designated as PLF547 datasetset, independently. Finally, we assess the performance of PM6-ML in larger protein-ligand complex models (up to 1,000 atoms) from the PL-REX dataset,[11] for which DFT calculations are available (136 systems, after zinc-containing complexes had been excluded). These employ a DFT functional similar to the one utilized for PM6-ML training, namely ωB97X-D3BJ, but with a more compact DZVP-DFT basis set[38] (which had been validated to perform exceptionally well for non-covalent interactions[39]).

Conformation energies and torsional profiles. Another important domain in

which the Δ-ML approach can offer substantial enhancements to SQM methods is in the calculation of relative energies of conformers. This is validated in multiple established benchmark datasets. In this context, the relative energies are evaluated with respect to the lowest-energy minimum of each molecule. The MPCONF196 set comprises small peptides and peptidic macrocycles, with conformation energies computed at the DLPNO-CCSD(T) level.[17] The Amino20x4 set, obtained from the GMTKN55 database,[40] comprises selected conformers of biogenic amino acids computed at the CCSD(T)-F12/CBS level. The SCONF dataset, drawn from the same database,[40] features conformers of sugars, a particularly challenging problem for SQM methods. Finally, we examine the source of the errors in the conformation energies by analyzing CCSD(T)/CBS torsional profiles of diverse drug-like molecules from the dataset of Sellers[41] (denoted "Torsions" in the remainder of the text).

Evaluation of the results. The primary measure of the performance of the studied methods in the validation datasets is the root mean square error (RMSE). When assessing multiple datasets simultaneously, the RMSE is evaluated in each of them and averaged. In some cases, it is more useful to discuss relative error, which we define as the RMSE divided by the average magnitude of the studied quantity at the benchmark level and express in percent. Additional statistical measures are available in the outputs provided in the Supporting Information.

2.6 SQM and DFT Methods Included in the Benchmarking

For the purpose of comparison with earlier SQM approaches, a selection of these was incorporated into the benchmarking process. Firstly, we utilise PM6 without the additional corrections.[12] Secondly, PM6 is employed with the most recent corrections for non-covalent interactions, described above and denoted PM6-D3H4X'. Thirdly, PM7 is included, which incorporates analogous corrections for dispersion and hydrogen bonding.[7] However, it has already been demonstrated to be unsuitable for the application to larger systems.[42] The calculations were performed using MOPAC[28] with the D3H4X' corrections added using the Cuby framework.[43] Finally, the extended tight binding method

GFN2-xTB was also tested, as this represents a different family of SQM methods and is known to perform well in a wide range of benchmarks.[8]

The PM6-ML method had been trained on DFT-D3 reference data; however, the majority of the validation datasets had been computed at a higher level. It is thus necessary to benchmark also the DFT method used to generate the training set with respect to the high-accuracy QM reference. The validation sets were thus recalculated using the same DFT setup, employing the ωB97M-D3BJ functional[22] and the def2-TZVPPD basis set.[23] The aforementioned calculations were performed using Orca, version 5.0.3.[31,32]

2.7 Δ-ML and ML Methods Included in the Benchmarking

TorchMD-NET/ET. First, in order to enable a comparison between our Δ-ML approach and a pure ML potential with the same setup, the TorchMD-NET/ET model was reparametrized with the same hyperparameters that were used to develop PM6-ML on the same training set. Analogously, the most accurate yet well-balanced model was selected from among the 40 candidates. The resulting model demonstrated superior performance compared to the published models trained on the SPICE dataset.[44] Additionally, the original model fails in large systems because of internal limits of the number of atoms defining the atomic environment. Consequently, we have chosen to utilize only our reparametrization in the benchmarking.

AIMNet2[45] is a machine learning potential that incorporates physics-based components, specifically separate terms for electrostatics and London dispersion. It encompasses a relatively broad chemical space. These properties render it a viable candidate for applications to large biomolecular systems. A variety of parameter sets are available for use. In this study, we employ the variant that was trained on the higher-level reference data, ωB97M-D3 DFT calculations, in the ensemble version (averaging over the ensemble of models) which is supposedly the best parameter set available. The code and parameters were downloaded from the AIMNet2 GitHub repository.[46]

MACE-OFF23[47] is a very recent equivariant message-passing ML potential covering sufficiently large chemical space. However, it is constrained to neutral molecules,

which renders it applicable only to a subset of the benchmarks presented in this paper. MACE-OFF23 had been trained on a dataset that was primarily derived from the SPICE database, which has also been utilized in this work. The calculations were performed using software[48,49] and model obtained from the respective GitHub repositories.[50,51] In this paper, we consistently use the most accurate MACE-OFF23 model labeled "large".

Other ML methods either do not provide a general model that can be readily used, do not cover the chemical space of our benchmarks, or are not freely available. The latter applies, for example, to the ML and Δ-ML methods described in Ref. 52, which would be interesting to include in our comparison, but access to them is limited.

Table 2: Δ-ML and ML methods used in the paper

Δ-ML	Ref.	Elements	Notes
PM6-ML	This work	15: H, C, N, O, P, S, F–I, Li–K, Mg, Ca	SQM + Equiv. transformer
AIQM1	20	4: H, C, N, O	SQM + ANI NN
QDπ	21	4: H, C, N, O	DFTB3 + ML

ML	Ref.	Elements	Notes
TorchMD-NET/ET	18, this work	15: H, C, N, O, P, S, F–I, Li–K, Mg, Ca	Equivariant transformer
AIMNet2	45	14: H, B, C, N, O, F–I, P, S, Si, As, Se	ML + Coulomb + D3
MACE-OFF23	47	10: H, C, N, O, F, P, S, Cl, Br, I	ML, neutral molecules only

2.8 Software Implementation

For the development and testing of PM6-ML, we interfaced the TorchMD-NET model to the Cuby framework[43,53], which already provides an interface to PM6 calculations in MOPAC and can combine arbitrary methods. The D3 dispersion correction is also provided by Cuby. This implementation is also useful for benchmarking, as it provides access to a wide range of predefined datasets and automates the calculations on them.[54] It is therefore the reference implementation used in the development of the method and all the results presented here had been computed using this code. The latest version

of Cuby, includes this interface, and an example input for performing PM6-ML calculations is provided in the documentation at `http://cuby4.molecular.cz/interface_torchmdnet.html`.

To simplify the use of PM6-ML for existing MOPAC users and to provide access to all the functionality of MOPAC, we also developed a direct interface between MOPAC (written in Fortran) and the code implementing the ML correction (written in Python and using PyTorch as the ML backend) with D3 dispersion provided by simple-dftd3 library.[55] It consists of a wrapper layer that initializes the ML model and passes control to a modified version of MOPAC, which can then request the computation of the ML correction whenever the SQM energy or gradient is evaluated. This code is available in a GitHub repository `https://github.com/Honza-R/mopac-ml`. It was tested to closely reproduce the results in the validation datasets introduced above (Section 2.5).

3 Results and Discussion

3.1 PM6-ML Training and Model Selection

A total of 40 models were trained to convergence for both the TorchMD-NET/ET and PM6-ML methods, initiating the training process from different, randomly generated initial conditions. The considerable variability in the outcomes necessitated a more comprehensive examination.

This variance is evident in the final value of the loss function (the error measure minimized by the optimization), with some models exhibiting significantly superior performance compared to others. It became evident that only a select few of the 40 models should be chosen for the final use. However, employing training loss as the sole criterion for model selection would not yield the optimal method for practical applications. The composition of the training set is significantly biased (over 60% of the data comprises drug-like molecules sourced from PubChem), and the overall error does not guarantee that the method will perform equally well in cases with lower representation in the training set.

Therefore, an additional metric was evaluated, reflecting both the accuracy of the model and the balance of the description of the different subclasses of the training set. A root mean square error (RMSE) was calculated for each subset of the training data, as detailed in Table 1. Furthermore, to enhance the model's generalizability to larger systems beyond those included in the training set, we assessed also the RMSE for the interaction energies in the L7 and S12L datasets from the validation set. The final error measure used for the selection of the optimal models is a product of the aforementioned RMSEs. In general, this RMSE product correlates well with the overall loss function; however, it alters the ordering of the top-performing models, favoring those with greater balance.

A concise statistical comparison between the best models and the median is presented in Table 3, which also illustrates the superiority of the Δ-ML approach, PM6-ML, over the pure ML model. The individual results for all models are provided in the Supporting Information, Tables Tables S5 to S10. All the results presented in the remaining part of the paper were computed with the best model selected using the aforementioned procedure. For PM6-ML, it is the model labeled "seed8" in the tables in the Supporting Information. This model will also be released along with the code implementing the PM6-ML method. For TorchMD-NET/ET, it is the model denoted "seed25".

Table 3: The metric used to select the final model, a product of root mean square errors computed in the individual subsets of the training set, and the final value of the loss function used in the training, for five best models and the median of all the 40 models trained.

Method	Model	RMSE product	Training loss
PM6-ML	1. (seed8)	0.09	108.93
	2. (seed15)	0.27	110.71
	3. (seed21)	0.28	113.92
	4. (seed17)	0.31	117.30
	5. (seed3)	0.39	114.18
	median	1.09	116.16
TorchMD-NET/ET	1. (seed25)	840	184.52
	2. (seed2)	879	171.22
	3. (seed19)	1126	163.28
	4. (seed31)	1361	178.67
	5. (seed30)	1449	174.02
	median	20788	185.81

3.2 Comparison to Earlier SQM/Δ-ML Approaches

First, we compare PM6-ML with the two other available Δ-ML methods based on SQM calculations, AIQM1[20] and QD-π.[21] They are both applicable to molecules with only four elements (H, C, N and O), which limits the choice of validation sets for this comparison. The results are plotted in the Figure 1 and are available in the supplementary information as a Table S1.

In the datasets of small, neutral non-covalent complexes (D442, HB375, R739, and S66), PM6-ML consistently yields root-mean-square errors (RMSE) under 1 kcal/mol, followed by AIQM1, which exhibits a larger error in the D442 dataset (RMSE 1.84 kcal/mol). The QD-π method yielded larger errors (RMSE between 2 and 3 kcal/mol) in all datasets except for S66. When charged molecules are considered in the IHB100 dataset of ionic hydrogen bonds, PM6-ML provides accurate results with an RMSE of 0.75 kcal/mol, representing a significant improvement over the earlier corrections for PM6. The RMSE yielded by AIQM1 is 2.3 kcal/mol, and this error is not systematic. Conversely, QD-π consistently overestimates the strength of these interactions, resulting in an RMSE of up to 7.48 kcal/mol. In the next two sets of larger non-covalent complexes, L7 and S12L, PM6-ML yields larger errors (RMSE 2.47 and 6.36 kcal/mol, respectively), but these are not significantly different from the errors of the DFT calculations (0.92 and 5.98 kcal/mol) that this method has been trained to reproduce. It is also noteworthy that the interaction energies in this dataset are an order of magnitude larger than in the smaller systems. Both AIQM1 and QD-π are unable to accurately describe the interaction energies in the larger systems, with errors of 19.5 and 62.4 kcal/mol in the S12L dataset.

This brief analysis demonstrates that, in addition to their limited coverage of chemical space, both AIQM1 and QD-π are substantially less accurate than PM6-ML and are not transferable to larger molecular systems.

3.3 Validation: Non-Covalent Interactions

Due to their importance in all large molecular systems, non-covalent interactions are a key target of the PM6-ML method. They are well represented in the training set, and

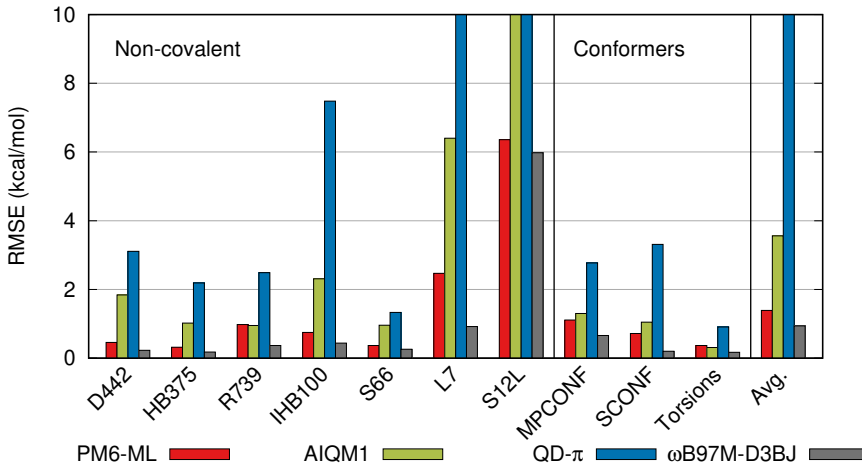

Figure 1: Comparison of PM6-ML to the earlier SQM Δ-ML methods, AIQM1 and QD-π, and to the ωB97M-D3BJ DFT calculations used for training PM6-ML. Subsets of the validation datasets containing only H, C, N and O elements were used. RMSE in kcal/mol.

PM6-ML is expected to provide significant improvement over the empirical corrections previously used with SQM methods.

Small non-covalent complexes. First, we analyze non-covalent interactions in smaller complexes, similar in size to the systems used in the training. Here, we take advantage of the NCIAtlas database, which comprises datasets representing separate classes of non-covalent interactions, providing further insight into the results. Additionally, these datasets feature reference data computed at a true benchmark level, CCSD(T)/CBS. The errors of the tested methods are plotted in Figure 2 and listed in Table S2 in the Supporting Information. The error averaged over the six NCIAtlas datasets is also listed it Table 4.

PM6-ML is unquestionably the best-performing method in this test. Not only does it yield the lowest error on average (RMSE averaged over the six datasets is 0.90 kcal/mol), but it also provides the best result in each of them. It is followed by AIMNet2 (avg. RMSE 1.65 kcal/mol) with consistently good results in all the datasets, TorchMD-NET/ET (avg. RMSE 1.96 kcal/mol), and GFN2-xTB (avg. RMSE 1.96 kcal/mol). All the SQM methods have some weak points, most often in the description of σ-hole interactions (SH250 set), repulsive contacts (R739), and ionic hydrogen bonds (IHB100). Among them, GFN2-xTB performs the best, with only the IHB100 set standing out with an

error of 3.9 kcal/mol. The standalone ML potentials generally perform better than the SQM methods, with the exception of the very large error of MACE-OFF23 in the SH250 dataset (7.3 kcal/mol), which is likely caused by the lack of relevant systems in its training set.

In these validation sets, PM6-ML benefits from having similar systems in the training set. This is also the reason why the TorchMD-NET/ET ML potential, trained on the same data, yields relatively small errors. On the other hand, it is clearly visible in the results that despite this, TorchMD-NET/ET performs worse specifically in the more exotic interactions (σ-hole bonds, hydrogen bonds involving heavier elements), which are represented more sparsely in the training set. The better description of these in PM6-ML indicates that it is the synergy of the SQM physics with the ML correction that enables higher accuracy with the same training data.

However, the analysis of the NCIAtlas datasets demonstrates only the accuracy that can be reached under very favorable conditions, i.e., in systems of similar size to those in the training set. It tests the ability of the models to interpolate the chemical space, but not their ability to extrapolate to larger molecular systems.

Table 4: Root mean square errors of the tested methods, in kcal/mol, averaged over the six NCIAtlas validation sets, non-covalent interactions in large systems (L7 and S12l datasets) and datasets of conformation energies (MPCONF196, SCONF, Amino20x4). Three best results in each column (neglecting the DFT) are highlighted.
[a] Excluding the IHB100 dataset.

Method	NCIAtlas	NCI/Large	Conformers
PM6	3.66	17.94	3.45
PM6-D3H4X'	2.80	**4.47**	4.04
PM7	3.22	11.41	3.37
GFN2-xTB	1.96	**2.71**	2.00
PM6-ML	**0.90**	**4.42**	**0.77**
TorchMD-NET/ET	**1.96**	17.06	**1.05**
AIMNet2	**1.65**	40.57	1.23
MACE-OFF23	2.38[a]	7.34	**0.71**
ωB97M-D3BJ	0.38	3.45	0.40

Large non-covalent complexes. Since the intended application of all the studied methods is larger systems, it is necessary to validate their transferability from small model complexes to larger ones where long-range interactions become more important. To do

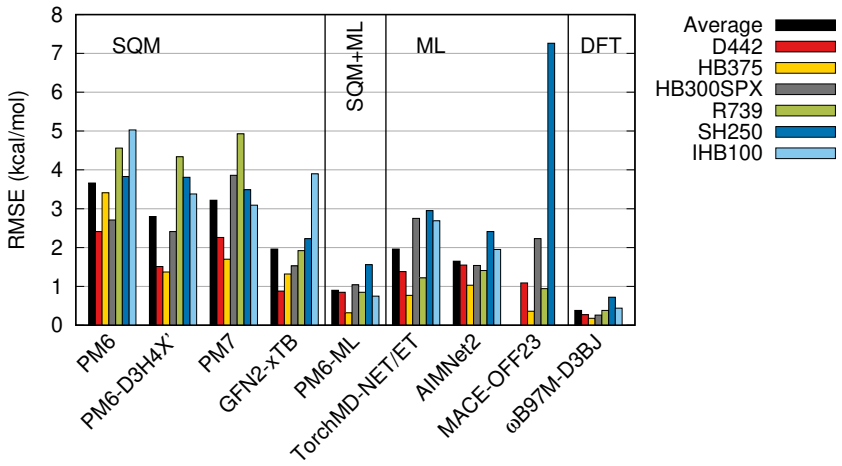

Figure 2: Errors of the tested methods in the validation subset of the Non-Covalent Interactions Atlas datasets. Root mean square error of interaction energies in kcal/mol. MACE-OFF23 is not applicable to the IHB100 dataset.

so, we employ several datasets where quality benchmarks (DPLNO-CCSD(T)) are still available. The L7 and S12L are datasets of larger non-covalent complexes with emphasis on π-π stacking. The PLA15 set features even larger models of protein-ligand complexes (15 ligands with surrounding amino acid residues extracted from an experimental structure of the complex). The pairwise interactions in this dataset form the PLF547 set, which we include here as well. The results are summarized in Figure 3 and listed in the Supporting Information, Table S3. Additionally, the average of the RMSE in the L7 and S12L sets is reported in Table 4 (PLA15 results are not available for all the methods).

Among the SQM methods, PM6 and PM7 yield very large errors, and the corrections in PM6-D3H4X' bring them down to an acceptable level, with an average RMSE of 4.5 kcal/mol in the L7 and S12L sets. GFN2-xTB performs excellently here with an RMSE of 2.7 kcal/mol but fails to converge in some of the PLA15 systems.

This test is much more challenging for the ML potentials. Our version of TorchMD-NET/ET, which lacks the description of long-range interactions, yields large errors increasing with the size of the system, up to an RMSE of 52 kcal/mol in the PLA15 set. It is more surprising that the AIMNet2 method does not perform better here, as it includes separate terms for long-range electrostatics and dispersion. Despite this, it exhibits the largest errors, with an RMSE of 65 kcal/mol in the S12L set (where the average interaction energy at the benchmark level is 41 kcal/mol, which translates to a relative error

of 160%). Interestingly, it also yields a large error in the PLA15 dataset (RMSE of 43 kcal/mol, which amounts to 25% of the average interaction energy in the set), while all the smaller fragments these systems comprise of, in the PLF547 dataset, are described much better, with the worst error in the subset of charged-charged species interactions having an RMSE of 1.5 kcal/mol, which translates to only 2.3% of the interaction energies there. Apparently, the model is able to describe smaller systems well but fails to scale to the larger ones. The only pure ML potential that performs better here is the MACE-OFF23 method, which yields an average RMSE of 7.3 kcal/mol in the L7 and S12L sets. It is, however, not applicable to charged systems, so it cannot be tested in the even larger complexes of the PLA15 set. The MACE-OFF23 potential does not include any explicit treatment of long-range interactions, but its training set was extended with a set of larger systems, which is likely the reason why it performs better.

The PM6-ML method works well here, with RMSEs of 2.47, 6.36, and 4.78 kcal/mol in the L7, S12L, and PLA15 sets, respectively. These results should be viewed in the context of the accuracy of the DFT on which the method is trained, which yields RMSEs of 0.92 and 5.98 kcal/mol in the L7 and S12L sets. The excellent result in PLA15 (where the relative error is only 2.9%) clearly demonstrates the transferability of the method to large systems and its ability to handle the strong long-range interactions present there. The advantage of the Δ-ML approach is that these interactions are handled independently at the SQM level, and the ML part of the potential is not forced to extrapolate into a territory not covered by the smaller systems used in training. Further tests in even larger protein-ligand complexes are discussed below.

3.4 Validation: Conformations and Torsions

Another quantity important in applications to large molecules is the relative energy of different conformers. This is a well-known weak point of all SQM methods that has not been satisfactorily addressed yet. In larger molecules, the conformation energies result from the interplay of non-covalent interactions, including repulsion defining the steric limits at short range, with the energetics of the torsions defined by the electronic structure

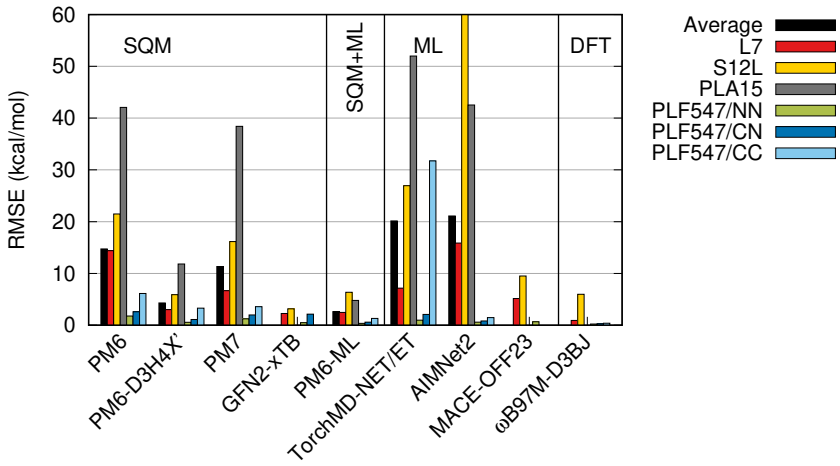

Figure 3: Errors of the tested methods in benchmark datasets of large non-covalent complexes. Root mean square error of interaction energies in kcal/mol. MACE-OFF23 is not applicable to the PLA15 and PLF547 datasets.

of the molecule. While the former can be addressed by corrections for non-covalent interactions, the description of the torsions is severely limited by the approximations inherent to the SQM methods, and there is no satisfactory solution to it. On the other hand, the latter contribution is localized to only a few atoms around each chemical bond, which makes it a good target for ML potentials.

In this paper, we examine the conformation energies in three benchmark datasets: MPCONF196, which comprises peptides and peptidic macrocycles,[17] SCONF, which represents sugars,[40] and Amino20x4, which covers selected conformers of the 20 biogenic amino acids.[40] The results are plotted in Figure 4 and listed in Table S4 in the Supporting information. It is clear that SQM methods do not yield very good results here. In particular, the sugar conformers in the SCONF dataset are very challenging, with the lowest RMSE achieved by GFN2-xTB being as high as 2.59 kcal/mol, which corresponds to 56% in relative terms. Similarly, in the MPCONF196 and Amino20x4 sets, GFN2-xTB performs the best among the SQM methods, but even these results are not very good. Its error averaged over the three datasets is 2.00 kcal/mol (see Table 4). It is, however, a significantly better result than that of the next method, PM7, which has an average RMSE of 3.37 kcal/mol.

On the other hand, these benchmarks are relatively easy for all the ML potentials tested. The results are consistently very good, with none of the ML potentials performing

worse than the best SQM method. Among the three methods, MACE-OFF23 performs the best, with an RMSE under 0.9 kcal/mol in all the datasets, and an average of 0.71 kcal/mol. Apparently, the conformer energetics are easy to learn, and the training data provide enough information for that. The same applies to the application of ML to correcting the SQM calculation in PM6-ML. It performs consistently well in all three datasets, with an average RMSE of 0.77 kcal/mol. However, even here, PM6-ML benefits from the synergy of SQM and ML, as the standalone TorchMD-NET/ET model does not perform as well, with an average RMSE of 1.05 kcal/mol.

The effect of the energetics of torsions can be isolated using the Torsions dataset, which features torsional profiles of 62 bonds in model drug-like molecules.[41] There, the best SQM result is again obtained with GFN2-xTB, which has an RMSE of 0.93 kcal/mol. The best ML method is MACE-OFF23, with an RMSE of 0.29 kcal/mol, while PM6-ML yields an error of 0.36 kcal/mol. It should be noted that these results already approach the accuracy of the DFT used in the training of these methods, with respect to the CCSD(T) benchmark, which has an RMSE of 0.18 kcal/mol.

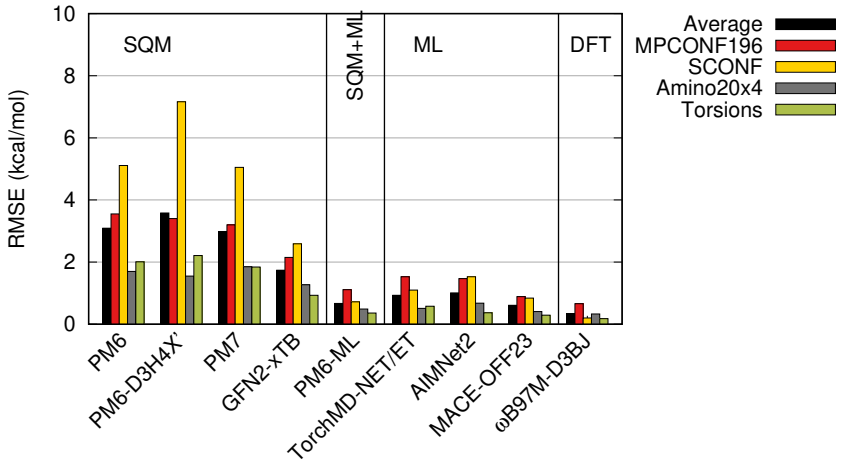

Figure 4: Errors of the tested methods in benchmark datasets of conformation energies and torsional profiles. Root mean square error of the relative energies in kcal/mol.

More importantly, this dataset provides additional insight into the phenomena when the torsional profiles are plotted. Two samples from the dataset are shown in Figure 5, and all the plots are available in the Supporting Information as Figure S1. It is clear that, in some cases, such as those shown here, PM6-D3H4X' describes the potential

qualitatively incorrectly, with missing barriers and artificial minima. The ML correction in PM6-ML is able to eliminate all these issues, and the resulting potential matches the benchmark almost perfectly. It is also important to note that this excellent result is achieved without reference data systematically sampling torsions in the training set – this information is recovered by the model from randomly sampled conformers only.

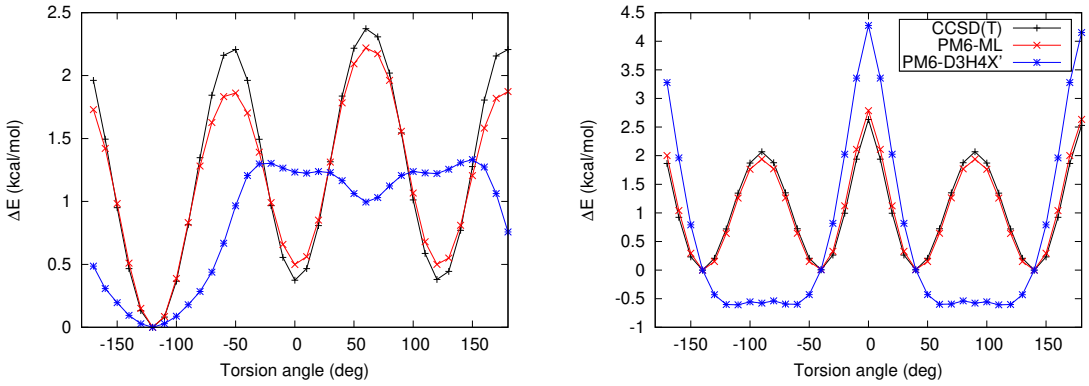

Figure 5: Examples of the torsional plots for 4-methyl-pentene (left) and biphenyl (right) computed with PM6-ML (red) and PM6-D3H4X' (blue), in comparison to the CCSD(T) benchmark (black).

3.5 Application to Protein–Ligand Complexes

The final test closely simulates the intended application to large biomolecular systems, specifically the evaluation of protein-ligand interactions. We take advantage of the PL-REX dataset we developed earlier,[11] which features DFT interaction energies in 164 models comprising the ligand and the surrounding part of the protein with $\sim$ 1000 atoms. Out of these, we use the 136 systems without zinc ions in this study.

The reference energies were computed using a DFT functional similar to the one used to train PM6-ML, ωB97X-D3BJ, but with a smaller basis set, DZVP-DFT.[38] Nevertheless, our previous results suggest that this basis set reproduces well the interaction energies obtained with a larger, triple-ζ basis set,[39] which makes the PL-REX DFT results a suitable reference. Furthermore, we omit the 3-body term in the D3 correction used in PM6-ML because the ωB97X-D3BJ reference data do not include it either.

We compare PM6-ML with PM6-D3H4X', which we have already applied successfully to scoring protein-ligand interactions,[11] and with two ML potentials applicable to

the chemical space spanned by the ligands, AIMNet2 and TorchMD-NET/ET (MACE-OFF23 is applicable only to neutral molecules). The key results are summarized in Table 5 and the individual data points are plotted in Fig. 6.

Table 5: Correlation and error of PM6-ML and other tested methods compared to the DFT reference, evaluated on 136 large models of protein–ligand complexes from the PL-REX dataset. The relative error is expressed as a percentage of the RMSE with respect to the average magnitude of the interaction energy in the set.

Method	R^2	RMSE, kcal/mol	RMSE, relative
PM6-ML	0.990	8.1	7.2%
PM6-D3H4X'	0.985	8.4	7.4%
AIMNet2	0.821	84.0	74.7%
TorchMD-NET/ET	0.311	48.9	43.5%

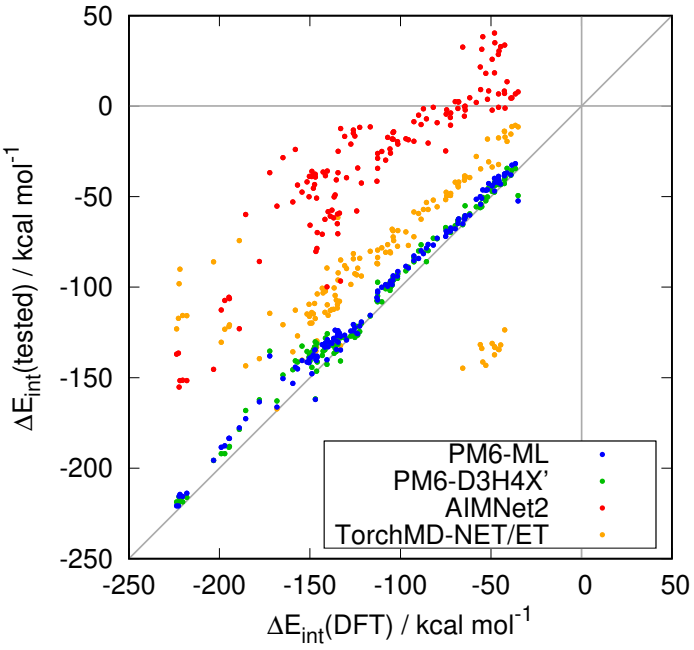

Figure 6: Plot of interaction energies obtained with PM6-ML and other tested methods against the DFT reference. The dots represent the 136 large models of protein–ligand complexes from the PL-REX dataset.

Both PM6-ML and PM6-D3H4X' reproduce the DFT results with a high degree of accuracy, with an RMSE of approximately 8 kcal/mol (equivalent to a relative error of approximately 7%, given the magnitude of the interaction energies). PM6-ML displays a slight advantage in this regard, however, the difference is inconsequential given that the reference DFT results themselves are subject to a non-negligible uncertainty due to

the utilisation of a relatively small basis set. A comparison of these two methods against more accurate benchmark is available in smaller systems of the same kind, namely the PLA15 dataset discussed above in Section 3.3. There, PM6-ML is clearly shown to be superior to PM6-D3H4X'. The principal conclusion to be drawn from the results in the PL-REX set is that PM6-ML maintains its high level of performance even when applied to significantly larger systems.

Both ML-only potentials are unable to adequately address this transition. It is unsurprising that the TorchMD-NET/ET potential is ineffective in this context. It is a simplified model that lacks the capacity to describe long-range effects, which manifests as a systematic bias towards weaker interaction energies. Additionally, a group of outliers is evident in Fig. 6, comprising all complexes of the BACE1 protein. The underlying cause of this discrepancy remains unclear. In contrast, the suboptimal performance of AIMNet2 is unexpected. This potential incorporates a physics-based description of long-range interactions (including both electrostatics and dispersion), suggesting that the observed large errors and systematic shift towards more positive interaction energies can be attributed to the parametrization, rather than the form, of the model. It is probable that including terms covering long-range effects is insufficient for developing a scalable model when the training set comprises only smaller systems.

4 Conclusions

The results presented in this paper show that the PM6-ML method, a Δ-ML approach based on efficient semiempirical QM calculations and a modern ML potential, has unique features that none of these components can currently achieve on their own. Although the accuracy of SQM methods has recently improved, it is now reaching the limits imposed by the fundamental approximations these methods are based on. On the other hand, even the best general ML potentials are limited by the scope of their training data more than methods that include parameter-free physically sound components. This is demonstrated not only by the ML methods' failure to describe larger molecular systems where long-

range interactions play an important role but also by other benchmarks discussed here.

PM6-ML addresses the most severe issue of existing SQM methods, the poor description of conformation energies of flexible molecules, and improves the accuracy of non-covalent interaction energies beyond what was achievable with the empirical corrections used previously. Compared to previous Δ-ML SQM methods, PM6-ML covers a much wider chemical space, enabling its application to real-world chemical problems. In this paper, we demonstrate its applicability to the study of protein-ligand interactions, where it achieves excellent accuracy.

The implementation of PM6-ML is freely available and includes a direct interface to MOPAC, the most widely used software for SQM calculations. This should facilitate its adoption in all cases where traditional SQM methods have been used.

The PM6-ML method presented here is, however, only the first step in its development. We are already working on extending its training set to include additional chemical elements and to cover other quantities and phenomena beyond those discussed in this paper.

5 Data and Code Availability

The trained models of both the PM6-ML correction and the standalone TorchMD-NET/ET potential used in this paper, as well as the wrapper enabling the PM6-ML calculations in MOPAC are available for download from a GitHub repository `https://github.com/Honza-R/mopac-ml`. The modified MOPAC code with the corresponding interface is available at `https://github.com/Honza-R/mopac/tree/pm6-ml`. Additionally, the PM6-ML calculations can be also carried out using an unmodified version of MOPAC interfaced to the Cuby framework, as described on the Cuby website at `http://cuby4.molecular.cz/interface_torchmdnet.html`.

6 Acknowledgements

We acknowledge the support of the Czech Science Foundation, Grant No. 22-17063S. We would also like to thank Prof. Saulo Vazquez who provided the data on which we discovered the discontinuity of the TorchMD-NET/ET potential, and Dr. Raul Pelaez from the TorchMD-NET development team who fixed this issue in the code.

Supporting Information Available

The Supporting Information comprises the following files:

SI_tables_and_figures.pdf - Additional tables and figures referenced in the main text, including tables of the errors presented in the paper only as plots.

SI_validation_outputs.zip - dataset calculation outputs for all the benchmarked methods and validation sets, which contain individual results as well as additional statistical measures.

References

(1) Thiel, W. Semiempirical quantum–chemical methods. *WIREs Comput Mol Sci* **2014**, *4*, 145–157.

(2) Akimov, A. V.; Prezhdo, O. V. Large-Scale Computations in Chemistry: A Bird's Eye View of a Vibrant Field. *Chem. Rev.* **2015**, *115*, 5797–5890.

(3) Řezáč, J.; Fanfrlík, J.; Salahub, D.; Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **2009**, *5*, 1749–1760.

(4) Řezáč, J.; Hobza, P. A halogen-bonding correction for the semiempirical PM6 method. *Chem. Phys. Lett.* **2011**, *506*, 286–289.

(5) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *J. Chem. Theory Comput.* **2012**, *8*, 141–151.

(6) Řezáč, J. Empirical Self-Consistent Correction for the Description of Hydrogen Bonds in DFTB3. *J. Chem. Theory Comput.* **2017**, *13*, 4804–4817.

(7) Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model* **2013**, *19*, 1–32.

(8) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(9) Kříž, K.; Řezáč, J. Benchmarking of Semiempirical Quantum-Mechanical Methods on Systems Relevant to Computer-Aided Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 1453–1460.

(10) Řezáč, J.; Stewart, J. J. P. How well do semiempirical QM methods describe the structure of proteins? *The Journal of Chemical Physics* **2023**, *158*, 044118.

(11) Pecina, A.; Fanfrlík, J.; Lepšík, M.; Řezáč, J. SQM2.20: Semiempirical quantum-mechanical scoring function yields DFT-quality protein–ligand binding affinity predictions in minutes. *Nat Commun* **2024**, *15*, 1127.

(12) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J Mol Model* **2007**, *13*, 1173–1213.

(13) Stewart, J. J. P. Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *International Journal of Quantum Chemistry* **1996**, *58*, 133–146.

(14) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.

(15) Miriyala, V. M.; Řezáč, J. Testing Semiempirical Quantum Mechanical Methods on a Data Set of Interaction Energies Mapping Repulsive Contacts in Organic Molecules. *J. Phys. Chem. A* **2018**, *122*, 2801–2808.

(16) Kříž, K.; Nováček, M.; Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 3: Repulsive Contacts. *J. Chem. Theory Comput.* **2021**, *17*, 1548–1561.

(17) Řezáč, J.; Bím, D.; Gutten, O.; Rulíšek, L. Toward Accurate Conformational Energies of Smaller Peptides and Medium-Sized Macrocycles: MPCONF196 Benchmark Energy Data Set. *J. Chem. Theory Comput.* **2018**, *14*, 1254–1266.

(18) Thölke, P.; De Fabritiis, G. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. 2022; `http://arxiv.org/abs/2202.02541`.

(19) Eastman, P.; Behara, P. K.; Dotson, D. L.; Galvelis, R.; Herr, J. E.; Horton, J. T.; Mao, Y.; Chodera, J. D.; Pritchard, B. P.; Wang, Y.; De Fabritiis, G.; Markland, T. E. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci Data* **2023**, *10*.

(20) Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O. Artificial Intelligence-Enhanced Quantum Chemical Method with Broad Applicability. 2021; `https://chemrxiv.org/engage/chemrxiv/article-details/60f8e1630b093ee195e2bca0`.

(21) Zeng, J.; Tao, Y.; Giese, T. J.; York, D. M. QDπ: A Quantum Deep Potential Interaction Model for Drug Discovery. *J. Chem. Theory Comput.* **2023**, *19*, 1261–1275.

(22) Najibi, A.; Goerigk, L. The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738.

(23) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(24) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets: Hydrogen Bonding. *J. Chem. Theory Comput.* **2020**, *16*, 2355–2368.

(25) Řezáč, J. Non-Covalent Interactions Atlas Benchmark Data Sets 2: Hydrogen Bonding in an Extended Chemical Space. *J. Chem. Theory Comput.* **2020**, *16*, 6305–6316.

(26) Kříž, K.; Řezáč, J. Non-covalent interactions atlas benchmark data sets 4: σ-hole interactions. *Phys. Chem. Chem. Phys.* **2022**, *24*, 14794–14804.

(27) Řezáč, J. Non-Covalent Interactions Atlas benchmark data sets 5: London dispersion in an extended chemical space. *Phys. Chem. Chem. Phys.* **2022**, *24*, 14780–14793.

(28) Stewart, J. J. P. MOPAC 2016. 2016; `http://openmopac.net/`.

(29) Tai, K. S.; Bailis, P.; Valiant, G. Equivariant Transformer Networks. 2019; `http://arxiv.org/abs/1901.11399`.

(30) Bronstein, M. M.; Bruna, J.; Cohen, T.; Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 2021; `http://arxiv.org/abs/2104.13478`.

(31) Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. The ORCA quantum chemistry program package. *The Journal of Chemical Physics* **2020**, *152*, 224108.

(32) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Computational Molecular Science* **2022**, *12*, e1606.

(33) TorchMD-NET GitHub repository. 2024; `https://github.com/torchmd/torchmd-net`.

(34) PyTorch machine learning library. 2024; `https://pytorch.org`.

(35) Sedlák, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

(36) Risthaus, T.; Grimme, S. Benchmarking of London Dispersion-Accounting Density Functional Theory Methods on Very Large Molecular Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 1580–1591.

(37) Villot, C.; Ballesteros, F.; Wang, D.; Lao, K. U. Coupled Cluster Benchmarking of Large Noncovalent Complexes in L7 and S12L as Well as the C60 Dimer, DNA–Ellipticine, and HIV–Indinavir. *J. Phys. Chem. A* **2022**, *126*, 4326–4341.

(38) Godbout, N.; Salahub, D. R.; Andzelm, J.; Wimmer, E. Optimization of Gaussian-type basis sets for local spin density functional calculations. Part I. Boron through neon, optimization technique and validation. *Can. J. Chem.* **1992**, *70*, 560–571.

(39) Hostaš, J.; Řezáč, J. Accurate DFT-D3 Calculations in a Small Basis Set. *J. Chem. Theory Comput.* **2017**, *13*, 3575–3585.

(40) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(41) Sellers, B. D.; James, N. C.; Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **2017**, *57*, 1265–1275.

(42) Hostaš, J.; Řezáč, J.; Hobza, P. On the performance of the semiempirical quantum mechanical PM6 and PM7 methods for noncovalent interactions. *Chemical Physics Letters* **2013**, *568–569*, 161–166.

(43) Řezáč, J. Cuby: An integrative framework for computational chemistry. *J. Comput. Chem.* **2016**, *37*, 1230–1237.

(44) SPICE-Models models GitHub repository. 2022; https://github.com/openmm/spice-models.

(45) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs. 2023; https://chemrxiv.org/engage/chemrxiv/article-details/6525b39e8bab5d2055123f75.

(46) AIMNet2 GitHub repository. 2023; https://github.com/isayevlab/AIMNet2.

(47) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules. 2023; http://arxiv.org/abs/2312.15211.

(48) Batatia, I.; Batzner, S.; Kovács, D. P.; Musaelian, A.; Simm, G. N. C.; Drautz, R.; Ortner, C.; Kozinsky, B.; Csányi, G. The Design Space of E(3)-Equivariant Atom-Centered Interatomic Potentials. 2022.

(49) Batatia, I.; Kovacs, D. P.; Simm, G. N. C.; Ortner, C.; Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. Advances in Neural Information Processing Systems. 2022.

(50) MACE GitHub repository. 2024; https://github.com/ACEsuit/mace.

(51) MACE-OFF23 models GitHub repository. 2024; https://github.com/ACEsuit/mace-off.

(52) Jacobson, L. D.; Stevenson, J. M.; Ramezanghorbani, F.; Ghoreishi, D.; Leswing, K.; Harder, E. D.; Abel, R. Transferable Neural Network Potential Energy Surfaces for Closed-Shell Organic Molecules: Extension to Ions. *J. Chem. Theory Comput.* **2022**, *18*, 2354–2366.

(53) Řezáč, J. Cuby 4, software framework for computational chemistry. 2015; http://cuby4.molecular.cz/.

(54) Řezáč, J.; Kontkanen, O. V.; Nováček, M. Working with benchmark datasets in the Cuby framework. *The Journal of Chemical Physics* **2024**, *160*, 202501.

(55) Simple-dftd3 library GitHub repository. 2024; `https://github.com/dftd3/simple-dftd3`.