Supporting Information:

Local Kernel Regression and Neural Network Approaches to the Conformational Landscapes of Oligopeptides

Raimon Fabregat,† Alberto Fabrizio,†,‡ Edgar A. Engel,¶,† Benjamin Meyer,†,‡
Veronika Juraskova,† Michele Ceriotti,¶,‡ and Clemence Corminboeuf*,†,‡

†Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
‡National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
¶Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

E-mail: clemence.corminboeuf@epfl.ch
Local Kernel Regression

In this Local Kernel Regression (LKR) model, a property $H_M$ of molecule $M$ is obtained as a sum of atomic contributions: $H_M \approx \sum_{j \in M} h_j$, where $h_j$ is the contribution to $H$ of atom $j$ in the molecule $M$, as usually done in most NN based models. We express each atomic contribution $h_j$ of the atomic environment $a_j$ using the common kernel regression formalism: $h_{a_j} = \sum_{a_i \in A_{train}} \alpha_j K(a_j, a_i)$, where $A_{train}$ is a set of atomic environments used to construct the kernel model. The quantity $H_M$ is then:

$$H_M \approx \sum_{a_i \in M} \delta h_{a_i} = \sum_{a_i \in M} \sum_{a_j \in A_{train}} \alpha_j K(a_j, a_i) = \sum_{a_j \in A_{train}} \alpha_j \sum_{a_i \in M} K(a_j, a_i) = \sum_{a_j \in A_{train}} \alpha_j S_{j,M}$$

(1)

The last expression means that for each atom in the training set we sum the kernel similarities with the atoms in the molecule, and then multiply by its contribution, the coefficient alpha. The regression coefficients $\alpha_i$ can be obtained using a linear regression from a set of $M$ molecules and their properties: $H \approx S \alpha$. Contrary to previous kernel based models to predict molecular properties, this model contains two training datasets: on one side, a set of atomic environments that compose the reference set of the regression, and on the other side, a training set of molecular structures and molecular properties which are used to fit the contribution of each reference atomic environment.

Concerning how we create the set of training atomic environments, from the 55378 training molecular structures we separated all the atomic environments of all the structures by atom type, obtaining more than 100,000 atomic environments for each atom type (carbon, hydrogen, oxygen, nitrogen and sulfur). The atomic environments, described by a vectorial quantity, were defined by the aSLATM$^S_1$ representation. We performed a 2-D Principle Component Analysis$^S_2$ (PCA) projection on the representation space for each atom type and then applied FPS to select the most different environments of each type, reducing the number to 39000 atomic environments. The projections $S_{n,m}$ of equation 1 were then computed on these atomic environments to solve the linear regression problem. As the number
Figure S1: **a)** Curves showing the accuracy vs. number of atomic environments chosen to generate the LKR models, using 70/30 train/test partition of the training data. In green is represented the accuracy obtained if random atomic environments are chosen to construct the LKR models instead of using OMP. In red, if FPS+ is used, i.e. FPS with the same composition in atomic environments as selected by OMP. **b)** Accuracy vs cost curve for the LKR model in the database of dipeptides. At each point the number marks the size of the training atoms chosen by OMP.

of dimensions is similar to the number of data points, some type of regularization is required to obtain a meaningful solution. Most kernel based models to predict molecular properties use either a Ridge regression \textsuperscript{83–85} to solve the linear equation. Here, instead, we use a greedy algorithm called Orthogonal Matching Pursuit \textsuperscript{86}(OMP) which generates the $\alpha$ coefficients in equation 1 by selecting one by one the coefficients that maximally reduce the error variance at each step. By setting only a fixed number of $\alpha$ coefficients to non-zero values, OMP generates a sparse representation, which in addition to acting as a regularization, has two more advantages. First, OMP does not require an inversion of the $S$ matrix in equation 1, which can be very computationally demanding both in time and in memory, and second, the sparse representation generated by OMP allows effective use of only a small fraction of the total atomic environments, which is critical for the overall cost of the model. Using OMP we are able to achieve a test MAE under 1 kcal/mol with as little as 300 atomic environments from the 39000 presampled with PCA (see the learning curves in Figure S1).
Figure S2: Energy predictions on the test set (y axis) v.s. target PBE (x axis). In blue is DFTB without ML correction, in orange DFTB + LKR and in green DFTB + NN. The number in the legend is the MAE between the predicted energies and the real values.
Figure S3: Isosurfaces of a scalar field representing the localization of the ML correction for two dipeptides, the aspartate (a) and the protonated histidine (b). The scalar field was generated with the LKR atomic corrections to the energy for that structure, convoluted with the atomic positions and a Gaussian filter of width 1 Å.

Extrapolation test

To evaluate the transferability of the models, we use two sets of structures, which are themselves divided into two subsets (see Figure S6). The first set corresponds to the random selection of 1000 structures taken from the converged T-RE (300K) sampling at the DFTB level. Out of these 1000 structures, 100 are optimized at the same DFTB level (i.e., 0K optimization). The second set consists in the random selection of 1000 structures taken from the resH-RE 300K simulations at the DFTB + LKR level out of which 100 of them are optimized statically with PBE (i.e., 0K) (see . Such a sample is representative of the 300 K canonical ensemble for both the baseline and the target potential. As for the optimized geometries, they serve to assess the accuracy and robustness of the models in the lowest energy region of the PES.
Figure S4: Scatter plots of the predictions of DFTB, DFTB + LKR (left) and DFTB + NN (right).

Figure S5: Distribution of errors for the training set of dipeptides (left) and for the tripeptide test (right) for DFTB, DFTB + LKR and DFTB + NN trained without forces.

Figure S6: a) Regression slopes between atomization energies of DFTB and PBE for different test sets of the tripeptide. b) Histograms of shifted errors (systematic deviations in the atomization energy have been removed) made on the tripeptide test set. The data is divided according to the potential that was used to generate them: (left) DFTB, (right) PBE.
Figure S7: Regression slopes between atomization energies of PBE and DFTB (left) and histograms of shifted errors (systematic deviations in the atomization energy have been removed) (right) for the test sets of the tripeptide, using different dispersion corrections (-D3BJ(a) and -D3H5(b)).

Figure S8: Fluctuations of the potential and total energy of the tripeptide Gly-Phe-Gly during a short NVE simulation using DFTB (left) and DFTB + NN (right) as driving potentials. The initial structure and velocities were taken from NVT simulations at 300K.
Figure S9: Grid with 2D free energy landscapes for each pair of the selected collective variables. The lower diagonal contains results from T-RE simulations using DFTB. The upper diagonal contains the results of the resH-RE simulations using DFTB + LKR. In the diagonal are the probability distributions of each collective variable for DFTB (blue) and DFTB + LKR (orange).
Figure S10: Convergence check for the T-RE simulations with DFTB. The left figure shows structures sampled during the simulation, clustered based on the free energy basin they belong to. The right figure shows the evolution of the relative integrated free energy between basins. As these quantities represent the rarest events and slowest dynamics modes of the simulation (i.e. the crossing between basins), their convergence is a good prove that the obtained sampling follows the true canonical distribution.

Figure S11: Free energy landscape of the alanine dipeptide (top) obtained using DFTB, (left), DFTB + LKR (middle) and DFTB + NN (right). Landscapes for the alanine dipeptide obtained in similar conditions can be found in the work of Mineva, Parrinello and coworkers. However, the comparison of the results with previous literature is not straightforward, as most computational studies contain free energies computed either at different levels of theory (sub-DFT) or in different environmental conditions (with solvent, etc...).
References

(S1) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. Nat. Chem. 2020, 12, 945–951.

(S2) Hastie, T.; Friedman, J.; Tibshirani, R. The Elements of Statistical Learning, 2nd ed.; Springer, 2001; New York, USA.

(S3) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. Phys. Rev. Lett. 2012, 108, 058301.

(S4) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. J. Chem. Theory Comput. 2015, 11, 2087–2096.

(S5) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. J. Chem. Theory Comput. 2017, 13, 5255–5264.

(S6) Pati, Y. C.; Rezaaiifar, R.; Krishnaprasad, P. S. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. Proceedings of 27th Asilomar conference on signals, systems and computers. 1993; pp 40–44.

(S7) Cuny, J.; Korchagina, K.; Menakbi, C.; Mineva, T. Metadynamics combined with auxiliary density functional and density functional tight-binding methods: alanine dipeptide as a case study. J. Mol. Model. 2017, 23, 72.

(S8) Bonati, L.; Zhang, Y.-Y.; Parrinello, M. Neural networks-based variationally enhanced sampling. Proc. Natl. Acad. Sci. U.S.A. 2019, 116, 17641–17647.