doi:10.2533/chimia.2019.1024 Chimia 73 (2019) 1024–1027

Machine Learning with and for Molecular Dynamics Simulations

Sereina Riniker*, Shuzhe Wang, Patrick Bleiziffer, Lennard Böselt, and Carmen Esposito

Abstract: From simple clustering techniques to more sophisticated neural networks, the use of machine learning has become a valuable tool in many fields of chemistry in the past decades. Here, we describe two different ways in which we explore the combination of machine learning (ML) and molecular dynamics (MD) simulations. One topic focuses on how the information in MD simulations can be encoded such that it can be used as input to train ML models for the quantitative understanding of molecular systems. The second topic addresses the utilization of machine learning to improve the set-up, interpretation, as well as accuracy of MD simulations.

Keywords: Machine learning · Molecular dynamics

Sereina Riniker completed her Master's degree in chemistry at ETH Zurich in 2008. After an internship in the research department of Givaudan AG and a research stay at the University of California Berkeley, she returned in 2009 to ETH Zurich to obtain a PhD in molecular dynamics simulations. From 2012 to 2014, she held a postdoctoral position in cheminformatics at the Novartis Institutes for BioMedical Research in Basel, Switzerland and Cambridge, Massachusetts. Since June 2014, Sereina Riniker is an Assistant Professor (with tenure track) of Computational Chemistry at the Department of Chemistry and Applied Biosciences at ETH Zurich. Picture credit: ETH Zurich/ Giulia Marthaler.

1. Introduction

Already before the advent of the current neural networks, machine-learning (ML) methods have been used in different areas of chemistry, most widely in cheminformatics.[1,2] In pharmaceutical industry, ML models are routinely being trained to predict binding affinities,[3–6] physicochemical properties of potential drug candidates (*e.g.* partition coefficients,[7,8] aqueous solubility[8–11]), or toxicological effects.[12–14] More recently, approaches to train ML models based on quantum-mechanical (QM) data have emerged to predict molecular atomization energies,[15,16] forces,[17,18] potential energies[19–21] or properties of materials.[22] The use of ML methods in the area of classical molecular dynamics (MD) simulations, on the other hand, is less explored.

In MD simulations, the motion of particles in a system is calculated with classical mechanics, *i.e.* Newton's equations of motion are solved numerically. The physical interactions between the particles can thereby be described by methods based on quantum mechanics (QM), classically, or by using a mixture of both, *i.e.* QM/MM[23] (for reviews on the different approaches see *e.g.* refs. [24–27]). The model for describing the particle–particle interactions determines both the accuracy of the simulations as well as the spatial and time scales that can be reached. In the following, we show how ML approaches can be used to either exploit the information inherent in MD simulations for the prediction of physicochemical properties, or to improve the accuracy of classical force fields.

*Correspondence: Prof. S. Riniker, E-mail: sriniker@ethz.ch
Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, CH-8092 Zurich

2. Learning With Molecular Dynamics

The time evolution of a system in an MD simulation is recorded in the form of coordinate and energy trajectories. Historically, ensemble averages are calculated to compare different properties with experimental values for the purpose of either interpreting experimental results or validating molecular simulations and force fields.[28,29] Such properties can be either thermodynamic (*e.g.* density, heat of vaporization, heat capacity, free energy of solvation), dynamic (*e.g.* self-diffusion coefficient, viscosity), or structural (*e.g.* radial distribution function, radius of gyration, NMR observables). However, by averaging over the trajectories potentially valuable information on kinetics and the relative population of conformational states is discarded. In the past years, kinetic modeling, *i.e.* Markov state modeling[30–33] has emerged as an approach to harness the time information stored in MD trajectories. Furthermore, the information from separate simulations can be combined with this technique. Briefly, to construct a Markov state model (MSM), the snapshots in the trajectories are clustered first structurally into microstates (discretization step), and subsequently the microstates are clustered kinetically into so-called metastable sets. MSMs have been used to gain insights into the conformational changes of a wide variety of biological systems.[33–39] In order to obtain converged kinetic models, the total simulation time is, however, typically in the range of microseconds or even milliseconds. Different ML approaches have been developed in the past few years to assist with kinetic modeling. For example, deep neural networks have been proposed as alternative to the commonly used principal component analysis (PCA) and time-lagged independent component analysis (TICA) to reduce the feature space and identify optimal collective variables for the building of MSMs.[40–42] Alternatively, McGibbon and Pande have developed an algorithm that learns the optimal geometric distance metric to classify structures based on kinetic proximity.[43] Recently, Noé and co-workers proposed the concept of Boltzmann generators to sample equilibrium states of complex systems employing deep learning.[44]

The time information in MD trajectories can also be exploited on a much shorter time scale using ML. For this, we developed the concept of MD fingerprints (MDFPs),[45] which encode MD trajectories into a 'ML readable' format. MDFPs are built as follows. The distributions of different terms extracted from short MD simulations (*e.g.* energetic terms, the solvent-accessible surface area, *etc*.) are described by statistical moments such as the average, variance, and skewness (or simpler the average, standard deviation, and median), and stored in a floating-point vector (Fig.

1). The variance contains thereby some entropic information. The MDFPs of different molecules can then be used as input features to train an ML model against experimental data. The MDFP-ML approach was tested on solvation free energies (ΔG_{solv}) of small organic molecules in solvents with different polarities: water, octanol, hexadecane, and cyclohexane. We could show that a single fingerprint based on simulations in water together with simple counts from the 2D topology of the molecule (*i.e.* number of heavy atoms, number of rotatable bonds, number of N, O, S, and halogen atoms) can be used to predict ΔG_{solv} in all four solvents. The prediction accuracy was thereby comparable to more rigorous MD-based methods but at a fraction of the computational cost.[45]

From solvation free energies, partition (or distribution) coefficients between two solvents *S1* and *S2* can be calculated straightforwardly,

$$\log P_{S2/S1} = \frac{\Delta G^{S1}_{solv} - \Delta G^{S2}_{solv}}{RT \ln(10)} \quad (1)$$

where R is the gas constant and T the absolute temperature. Thus, from ML models trained for the individual solvents, partition coefficients in all pairs of solvents can be calculated. Again, we found that the MDFP-ML approach performed similarly to the MD-based methods for predicting partition coefficients in octanol/water, hexadecane/water, and cyclohexane/water. Furthermore, when applied retrospectively to the molecules in the SAMPL5 blind challenge for cyclohexane/water distribution coefficients,[46] the MDFP-ML approach outperformed the more rigorous methods (Fig. 2). Recently, we participated in the SAMPL6 blind challenge[47] to predict octanol/water partition coefficients $\log P_{oct/wat}$ of a small set of molecules, resulting in a top 10 ranking.[48] There is significantly more $\log P_{oct/wat}$ data available than ΔG_{solv} data for octanol and water (10^4 range versus 10^2 range), because $\log P_{oct/wat}$ is routinely measured in medicinal chemistry as an indicator for the aqueous solubility and passive membrane permeability of a compound. We compared ML models with MDFPs as input trained against experimental ΔG_{solv} values with those trained against experimental $\log P_{oct/wat}$. The results indicated that the former are more robust (*i.e.* better accuracy with smaller training sets) while the latter profit from the large amount of experimental $\log P_{oct/wat}$ values available.

The concept of MDFPs is very general and versatile. Depending on the property to be learned, different types of simulations (*e.g.* solute in solvent, pure liquids, crystals) can be performed (and combined), and different terms can be extracted to construct the fingerprints. In the future, we want to explore the use of the MDFP-ML approach for the prediction of other physicochemical properties such as melting point, vapour pressure, and aqueous solubility. Furthermore, we plan to extend ligand-based MDFPs with terms from protein–ligand simulations to predict activity/affinity for target proteins.

3. Learning for Molecular Dynamics

In classical MD simulations, the physical interactions between atoms are described with an empirical force field that consists of different bonded (*i.e.* bond stretching, bond-angle bending, and dihedral-angle torsion) and non-bonded (van der Waals and electrostatic) energy terms (for a review see *e.g.* ref. [26]). This involves a large number of parameters for each molecule, which are fitted to quantum-mechanical (QM) or available experimental data. The neglect of electronic degrees of freedom and the constraining of high frequency modes (*e.g.* C–H vibration) reduce the computational cost of the calculations dramatically such that much larger spatial scales and longer time scales can be reached. This comes, however, at the cost of accuracy and transferability. There is thus a need for more accurate and general force fields. In this context, ML approaches can be applied in several ways. ML models can be trained to learn the potential-energy surface such that they replace the force field entirely (see *e.g.* refs. [49,50]). A major challenge for this approach is the generation of a diverse enough training set in terms of chemical and conformational space.[51] Furthermore, computing MD trajectories with ML-based potentials is slower than with a classical force field (but faster than the QM calculations which were used as reference).

A different use case for ML is in force-field development, *i.e.* the aim is to improve the accuracy of force fields instead of replacing them. We have recently developed an ML approach to predict partial charges of organic molecules.[52] As reference partial charges extracted from density-functional theory (DFT) calculations including implicit solvent (with different dielectric constants) were used. The input features depend solely on the 2D topology of a molecule represented by an atom-centered atom-pairs (AP) fingerprint,[53] thus the learned partial charges are conformation independent and averaged over similar substructures in different molecules. The latter aspect should improve the transferability. The workflow for training of the ML models is shown schematically in Fig. 3. Once the ML models are trained, the generation of the ML-based partial charges of a new molecule is extremely fast and scales linearly with the number of atoms in the molecule. An individual QM calculation per new molecule as typically done with the current force fields is no longer required. The usability of ML models relies heavily on the training set. For the partial charges, a large training set of more than 130'000 lead-like molecules was compiled, which represent the substructures present in the lead-like parts of the public databases ChEMBL[54] and ZINC.[55,56] We chose lead-like compounds because these are large enough to contain multiple functional groups but small

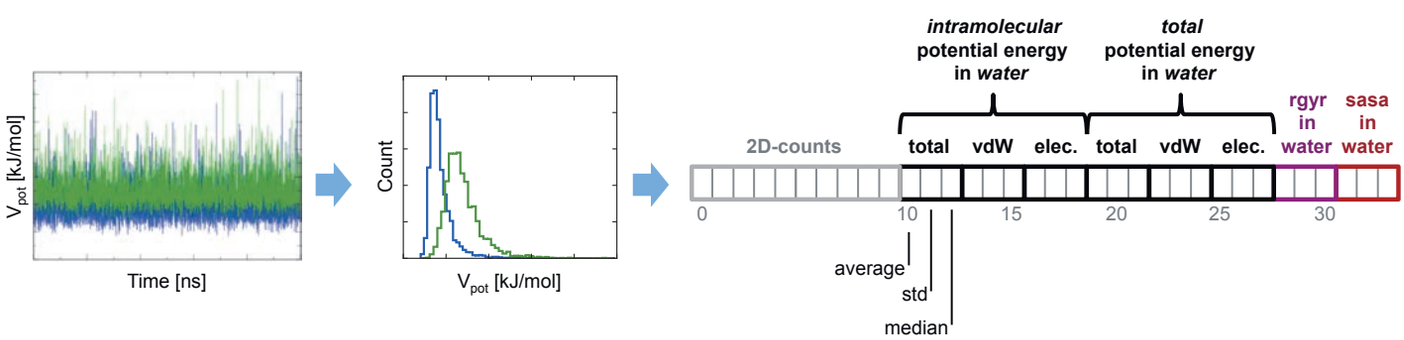

Fig. 1. Schematic representation of the construction process of the MDFP used in ref. [45]. From left to right: (i) The time series of properties (*e.g.* energetic terms, radius of gyration (rgyr) and solvent-accessible surface area (sasa)) are extracted from the MD simulations, (ii) the time series are converted into distributions, and (iii) the distributions are encoded in the MDFP as average, standard deviation (std) and median, and combined with simple counts from the 2D topological structure of the molecule (*e.g.* number of heavy atoms, number of rotatable bonds, number of oxygens, nitrogens, *etc.*).

enough for QM calculations in a reasonably large bases set. The best performance was obtained using an AP fingerprint with a maximum bond length of four, *i.e.* atom pairs up to four bonds apart are recorded in the fingerprint. The highest accuracy was obtained for hydrogen and fluorine, for which all data points were within an absolute error of 0.05 e. On the other hand, the lowest accuracy was found for phosphorus (79% of the data points within an absolute error of 0.05 e), which is likely due to the relatively small number of data points in the training set (around 1'000 compared to 1'000'000 for hydrogen). The database behind the ML-based partial charges can be continually expanded to further improve the coverage of the organic chemical space.

In a next step, we tested the combination of the ML-based partial charges with the van der Waals parameters of general force fields such as GAFF[57] or OPLS-AA.[58,59] These force fields typically use partial charges derived from individual QM calculations, which is relatively computationally expensive and limits transferability among similar substructures in different mole-

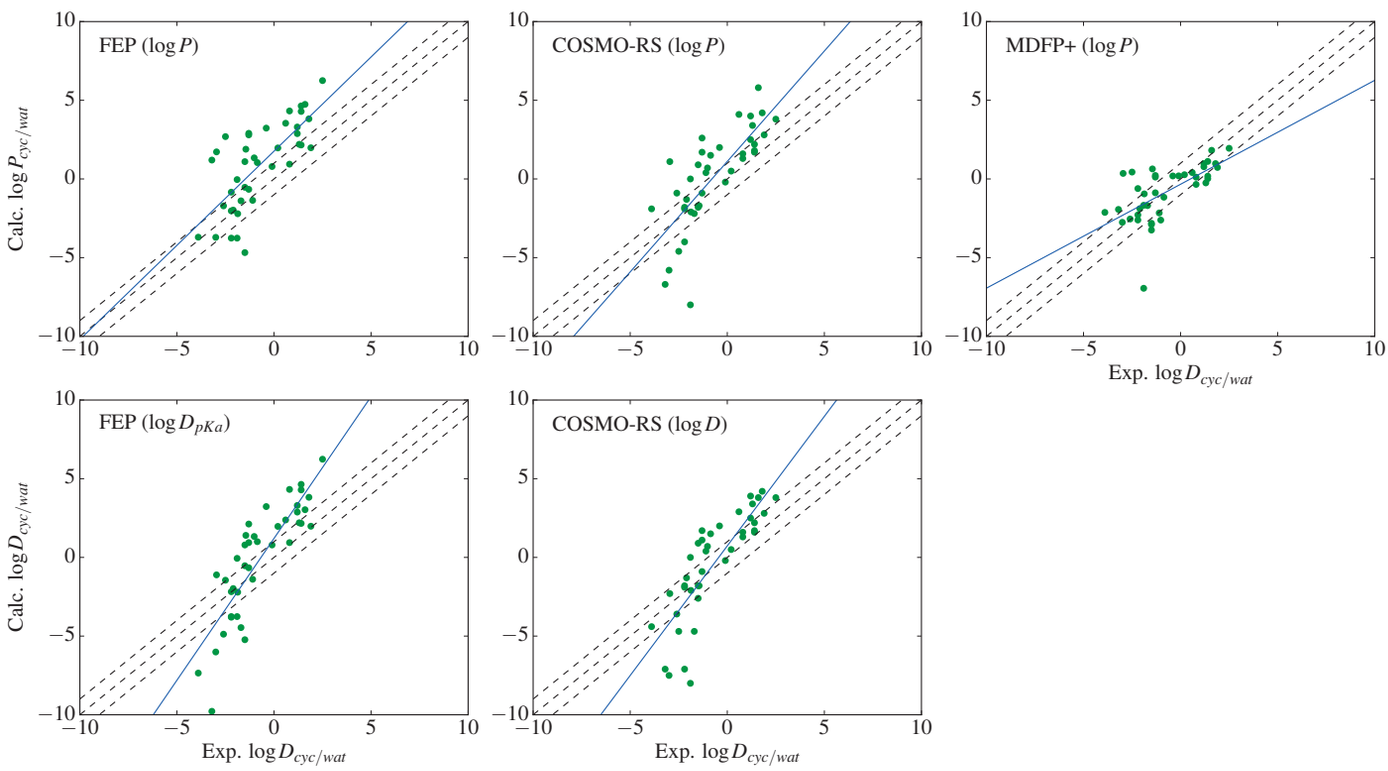

Fig. 2. Comparison of computed and experimental cyclohexane/water distribution coefficients log $D_{cyc/wat}$ of 40 SAMPL5 compounds. Predictions were generated for ΔG_{solv} using a fusion model (average between a linear LASSO model and a GTB model) trained on MDFPs. From the predicted ΔG_{solv} values, log*P* values were calculated using Eqn. (1). For comparison, the log*P* and log*D* results from FEP[46] and COSMO-RS[68] calculations are shown. The linear regression lines are shown in blue. The black dashed lines represent y = x and an interval ±1 log unit. Reprinted with permission from ref. [45]. Copyright 2017 American Chemical Society.

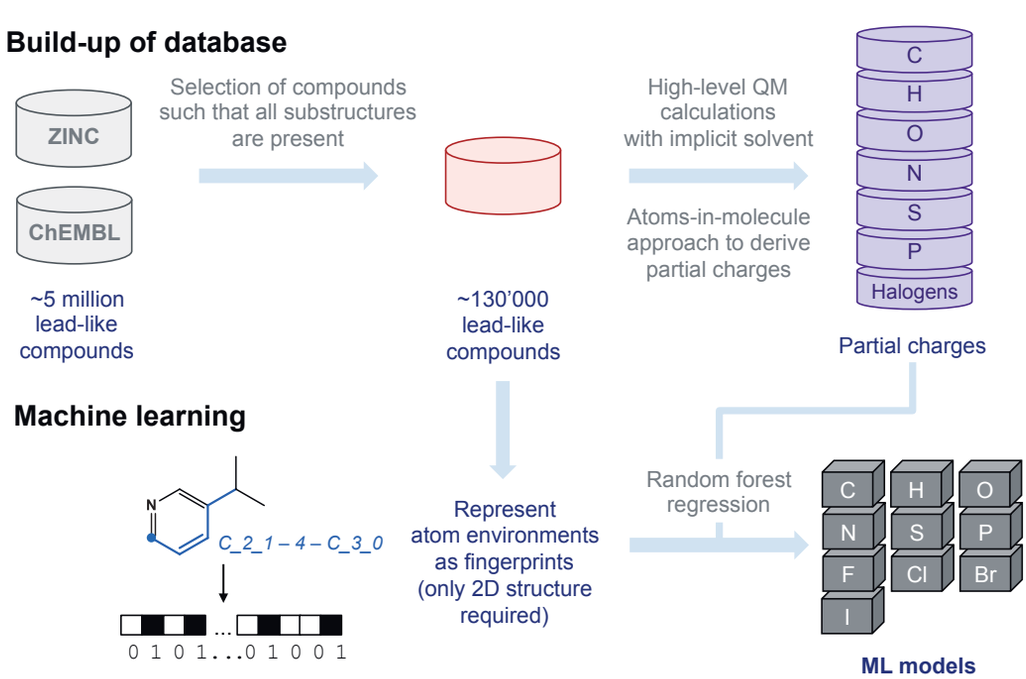

Fig. 3. Schematic representation of the workflow to train ML models to predict partial charges of organic molecules as used in ref. [52].

cules. The results showed that (a) the ML-based partial charges performed similarly as the commonly used ones in these force fields, and that (b) the ML-based partial charges derived with an implicit solvent model with a dielectric permittivity of 4 resulted in the best reproduction of thermodynamic properties. Due to the speed of the generation and their accuracy, the ML-based partial charges can also be interesting as descriptors in cheminformatics, *e.g.* for quantitative structure-activity relationship (QSAR) or quantitative structure-property relationship (QSPR) models.[60–62] Besides partial charges[52,63,64] it is also possible to learn multipole coefficients, which capture the dispersion interaction and can be directly connected to the C6 coefficients in the Lennard-Jones potential-energy function.[65–67]

4. Outlook

We have highlighted the methods developed in our group and others to exploit the combination of machine learning and molecular dynamics for the application in property prediction as well as for the improvement of the analysis, accuracy and sampling efficiency of MD simulations. Such combined approaches have the potential to influence fundamentally the field of computational chemistry. The ongoing advances in the field of machine learning and the increase in computational power will continue to impact the existing approaches and create opportunities for new applications and developments.

Acknowledgements

The authors gratefully acknowledge financial support by the Swiss National Science Foundation (Grant no. 200021-178762) and by ETH Zurich (Grant no. ETH-34 17-2).

Received: September 24, 2019

[1] J. B. O. Mitchell, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468.
[2] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241.
[3] D. Plewczynski, S. A. H. Spieser, U. Koch, *Comb. Chem. High Throughput Screening* **2009**, *12*, 358.
[4] S. Riniker, N. Fechner, G. A. Landrum, *J. Chem. Inf. Model.* **2013**, *53*, 2829.
[5] J. Jiménez, M. Skalic, G. Martínez-Rosell, G. De Fabritiis, *J. Chem. Inf. Model.* **2018**, *58*, 287.
[6] A. Mayr, G. Klambauer, T. Unterthiner, M. Steijaert, J. K. Wegner, H. Ceulemans, D.-A. Clevert, S. Hochreiter, *Chem. Sci.* **2018**, *9*, 5441.
[7] D. Eros, I. Kövesdi, L. Orfi, K. Takács-Novák, G. Acsády, G. Kéri, *Curr. Med. Chem.* **2002**, *9*, 1819.
[8] C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, K. F. Jensen, *J. Chem. Inf. Model.* **2017**, *57*, 1757.
[9] J. Wang, T. Hou, *Comb. Chem. High Throughput Screening* **2011**, *14*, 328.
[10] A. Lusci, G. Pollastri, P. Baldi, *J. Chem. Inf. Model.* **2013**, *53*, 1563.
[11] J. L. McDonagh, N. Nath, L. De Ferrari, T. van Mourik, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2014**, *54*, 844.
[12] A. B. Raies, V. B. Bajic, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2016**, *6*, 147.
[13] A. Mayr, G. Klambauer, T. Unterthiner, S. Hochreiter, *Front. Environ. Sci.* **2016**, *3*, 1.
[14] Y. Wu, G. Wang, *Int. J. Mol. Sci.* **2018**, *19*, 2358.
[15] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[16] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404.
[17] Z. Li, J. R. Kermode, A. De Vita, *Phys. Rev. Lett.* **2015**, *114*, 096405.
[18] V. Botu, R. Ramprasad, *Phys. Rev. B* **2015**, *92*, 094306.
[19] S. Lorenz, A. Gross, M. Scheffler, *Chem. Phys. Lett.* **2004**, *395*, 210.
[20] J. Behler, M. Parinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[21] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.
[22] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810.
[23] A. Warshel, M. Levitt, *J. Mol. Biol.* **1976**, *103*, 227.
[24] R. O. Jones, *Rev. Mol. Phys.* **2015**, *87*, 897.
[25] A. S. Christensen, T. Kubar, Q. Cui, M. Elstner, *Chem. Rev.* **2016**, *116*, 5301.
[26] S. Riniker, *J. Chem. Inf. Model.* **2018**, *58*, 565.
[27] H. M. Senn, W. Thiel, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198.
[28] W. F. van Gunsteren, A. E. Mark, *J. Chem. Phys.* **1998**, *108*, 6109.
[29] W. F. van Gunsteren, X. Daura, N. Hansen, A. E. Mark, C. Oostenbrink, S. Riniker, L. J. Smith, *Angew. Chem. Int. Ed.* **2018**, *57*, 884.
[30] C. Schütte, A. Fischer, W. Huisinga, P. Deuflhard, *J. Comput. Phys.* **1999**, *151*, 146.
[31] W. C. Swope, J. W. Pitera, F. Suits, *J. Phys. Chem. B* **2004**, *108*, 6571.
[32] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, F. Noé, *J. Chem. Phys.* **2011**, *134*, 174105.
[33] J. D. Chodera, F. Noé, *Curr. Opin. Struct. Biol.* **2014**, *25*, 135.
[34] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, V. S. Pande, *J. Am. Chem. Soc.* **2011**, *133*, 18413.
[35] D. Shukla, C. X. Hernández, J. K. Weber, V. S. Pande, *Acc. Chem. Res.* **2015**, *48*, 414.
[36] J. Witek, B. G. Keller, M. Blatter, A. Meissner, T. Wagner, S. Riniker, *J. Chem. Inf. Model.* **2016**, *56*, 1547.
[37] J. Witek, M. Mühlbauer, B. G. Keller, M. Blatter, A. Meissner, T. Wagner, S. Riniker, *ChemPhysChem* **2017**, *18*, 3309.
[38] J. Witek, S. Wang, R. Lingwood, A. Dounas, H.-J. Roth, M. Fouché, M. Blatter, O. Lemke, B. Keller, S. Riniker, *J. Chem. Inf. Model.* **2019**, *59*, 294.
[39] S. M. Hanson, G. Georghiou, M. K. Thakur, W. T. Miller, J. S. Rest, J. D. Chodera, M. A. Seeliger, *Cell Chem. Biol.* **2019**, *26*, 390.
[40] C. Wehmeyer, F. Noé, *J. Chem. Phys.* **2018**, *148*, 241703.
[41] A. Mardt, L. Pasquali, H. Wu, F. Noé, *Nat. Commun.* **2018**, *9*, 5.
[42] W. Chen, H. Sidky, A. L. Ferguson, *J. Chem. Phys.* **2019**, *150*, 214114.
[43] R. T. McGibbon, V. S. Pande, *J. Chem. Theory Comput.* **2013**, *9*, 2900.
[44] F. Noé, S. Olsson, J. Köhler, H. Wu, *Science* **2019**, *365*, eaaw1147.
[45] S. Riniker, *J. Chem. Inf. Model.* **2017**, *57*, 726.
[46] C. C. Bannan, K. H. Burley, M. Chiu, M. R. Shirts, M. K. Gilson, D. L. Mobley, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 927.
[47] M. Işık, D. Levorse, D. L. Mobley, T. Rhodes, J. D. Chodera, *bioRxiv* **2019**, 757393.
[48] S. Wang, S. Riniker, *J. Comput.-Aided Mol. Des.* **2019**, in press, doi:10.1007/s10822-019-00252-6.
[49] J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.
[50] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.
[51] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.
[52] P. Bleiziffer, K. Schaller, S. Riniker, *J. Chem. Inf. Model.* **2018**, *58*, 579.
[53] R. E. Carhart, D. H. Smith, R. Venkataraghavan, *J. Chem. Inf. Model.* **1985**, *25*, 64.
[54] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J. P. Overington, *Nucleic Acids Res.* **2012**, *40*, D1100.
[55] J. J. Irwin, B. K. Shoichet, *J. Chem. Inf. Model.* **2005**, *45*, 177.
[56] T. Sterling, J. J. Irwin, *J. Chem. Inf. Model.* **2015**, *55*, 2324.
[57] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.
[58] W. L. Jorgensen, D. S. Maxwell, J. Tirado-Rives, *J. Am. Chem. Soc.* **1996**, *118*, 11225.
[59] W. L. Jorgensen, J. Tirado-Rives, *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665.
[60] A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413.
[61] P. Polishchuk, *J. Chem. Inf. Model.* **2017**, *57*, 2618.
[62] E. N. Feinberg, D. Sur, Z. Wu, B. E. Husic, H. Mai, Li, Y., S. Sun, J. Yang, B. Ramsundar, V. S. Pande, *ACS Cent. Sci.* **2018**, *4*, 1520.
[63] B. K. Rai, G. A. Bakken, *J. Comput. Chem.* **2013**, *34*, 1661.
[64] B. Nebgen, N. Lubbers, J. S. Smith, A. E. Sifain, A. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, S. Tretiak, *J. Chem. Theory Comput.* **2018**, *14*, 4687.
[65] T. Bereau, D. Andrienko, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*, 3225.
[66] T. Bereau, R. A. DiStasio, A. Tkatchenko, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241706.
[67] E. Heid, M. Fleck, P. Chatterjee, C. Schröder, A. D. MacKerell, *J. Chem. Theory Comput.* **2019**, *15*, 2460.
[68] A. Klamt, F. Eckert, J. Reinisch, K. Wichmann, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 959.