doi:10.2533/chimia.2014.620 Chimia 68 (2014) 620–623

Crossing the Boundaries within Computational Chemistry: From Molecular Dynamics to Cheminformatics and back

Sereina Riniker*

Abstract: The research in the group for computational chemistry at the ETH Zurich focuses on the development of methods and software for classical molecular dynamics simulations and cheminformatics, and their application to biological and chemical questions. Here, important advances and challenges in these subfields of computational chemistry are reviewed and potential opportunities for cross-fertilization are outlined.

Keywords: Cheminformatics · Molecular dynamics simulations

1. Introduction

Molecular dynamics (MD) simulations and cheminformatics are two subdisciplines of computational chemistry. Classical cheminformatics deals with the development of algorithms for handling, searching and mining large databases of small organic molecules, *i.e*. the focus is on high-throughput with a two-dimensional, static description of the molecules. MD on the other hand is typically performed on a single biological system, *i.e.* it is low-throughput, three-dimensional and dynamic. What is common to all areas of computational chemistry is that the methodological development is strongly coupled to the advances in computer power. MD simulations for example started out in 1977 with 8.8 picoseconds of bovine pancreatic trypsin inhibitor (BPTI, 58 residues) in vacuum.[1] Afterwards the increasing speed of the central processing units (CPUs) pushed the accessible time scales higher. Parallelization using graphics processing units (GPUs) accelerated the performance even further, and is now standardly employed in MD programs.[2–5] The millisecond simulation of the same BPTI in explicit water on a special-purpose supercomputer by the Shaw group in 2009 so far presents the culmination of this development.[6] Due to these advances, MD simulations have started to become feasible for applications in computer-aided drug design that have previously been out of reach due to computational limitations.

A similar development can be observed for areas in cheminformatics and computer-aided drug design. First of all, the size of the data sets available has been growing tremendously in the past decade, both within the pharmaceutical industry and in the public space. High-throughput screening routinely employed in industry enables the testing of 1–5 million compounds within a few weeks.[7] PubChem[8] introduced in 2004 now contains over 50 million compounds and close to 200'000 confirmatory and primary bioactivity screens with several million data points.[9] The current version of ChEMBL,[10] version 18, contains 1.5 million compounds with 12 million activity values extracted from scientific publications.[11] The advances in computer power enable the efficient processing and searching of these large data sets. Machine-learning (ML) methods are already heavily used for cheminformatics applications as will be discussed in detail in the following, but only with massive computing power do approaches such as deep learning with artificial neural networks become meaningful for bioactivity prediction using these large data sets.[12]

The advances in computer power also had a great influence on the method complexity employed in structure-based virtual screening (VS) or molecular docking. In VS, new potentially active molecules are fished for in a large library using known active molecules as bait.[13] The structure-based approach uses thereby the known three-dimensional structure of the protein and its binding pocket (usually a crystal structure) to generate docking poses of the different potential ligands and compare them using a scoring function.[14–16] In the early days of docking, both the ligand and the protein were kept rigid, but with increasing computer power algorithms to search the accessible conformational space of the ligand in the binding pocket and more sophisticated scoring functions have been introduced.[17] Now, the focus has shifted to the efficient treatment of protein flexibility, which will be discussed in detail in the following.

Here, several research areas will be highlighted where the advances in computer power have opened up new opportunities for cross-fertilization between different subdisciplines of computational chemistry.

2. The Hunt for Protein Conformations

In recent years, the importance of not only considering ligand flexibility but also protein flexibility in molecular docking has been increasingly recognized.[18–23] The various proposed strategies fall into four categories, with increasing degree of protein flexibility included. Firstly, soft-docking approaches consider only small local movements of the protein implicitly by treating the protein as a soft body, which can be penetrated to a small degree by the ligand atoms.[24] In the second class of methods, the side chain flexibility is evaluated explicitly while keeping the backbone rigid, using for example rotamer libraries.[25,26] In the third category, certain domains of the protein become flexible,[27] while methods in the fourth category rep-

*Correspondence: Prof. Dr. S. Riniker
Laboratory of Physical Chemistry
ETH Zurich
Vladimir-Prelog-Weg 2
CH-8093 Zurich
E-mail: sriniker@ethz.ch

resent the full flexibility of the protein by a conformational ensemble of rigid structures.[28] Considering the full flexibility of a protein is of course the most desirable approach, but also the computationally most expensive one. Multiple protein conformations can be obtained, for example, from a set of crystal structures of the same protein with different ligands bound,[29,30] but this implies that the target has been well studied in the literature such that the necessary variety of crystal structures is available. Another source of protein conformations for docking is the trajectory generated by a MD simulation. Early examples have been the docking study of the M603 Fab fragment of immunoglobulin McPC603[31,32] and of the immunophilin FKBP.[33] Very recently snapshots from MD simulations biased towards the experimental crystal structure were used for docking of cyclin-dependent kinase 2 (CDK2) and Factor Xa.[34] This last example indicates that with the advances in computer power, the use of MD simulations as a source for protein conformations for docking on a larger scale becomes feasible. A general limitation of ensemble-based docking is, however, the number of protein structures that can be used due to combinatorial explosion. This shifts the focus to the choice of clustering algorithm used to extract the appropriate conformations from an MD trajectory, as well as on the development of alternative strategies to consider the conformational ensemble of the protein efficiently.

3. The Happiness of Water

Another problem in docking is posed by the water molecules in the binding pocket. The importance of considering them during the docking process has been increasingly recognized during the past years and a large number of strategies have been proposed with varying computational effort, but all aiming at distinguishing between 'happy' water molecules (*i.e.* displacement decreases in binding affinity) and 'unhappy' water molecules (*i.e.* displacement increases binding affinity).[35–39] As information of the dynamics of the water molecules in the binding pocket is required for this analysis, several approaches employ MD simulations for this task. The popular commercial program WaterMap[36,40] for example runs a short MD simulation of the rigid protein in explicit water to identify the preferred positions of water molecules in the binding pocket and subsequently estimates the enthalpic and entropic differences between these water molecules compared to those in the bulk. A different approach applied a recently published free energy perturbation (FEP) method called enveloping distribution sampling[41,42] to calculate ΔG between having a water molecule or a hydrophobic moiety at certain positions in the binding pocket of a fully flexible protein.[37] Although many methods for analyzing water molecules in binding pockets have been proposed, the validation of new methods and comparison between existing methods remains a major challenge. Apart from retrospective case studies, a systematic prospective validation would be needed.

4. How Well Does it Bind?

The scoring functions used in molecular docking aim at estimating the absolute binding free energy, ΔG_{bind}, of a ligand, which is the difference between the free energy of the ligand in solution and bound to the protein.[43] As the docking process contains no information about the dynamics of the system, the accurate estimation of the entropic contribution to the absolute binding free energy is a major problem. MD simulations could in principle provide this information, but this would require the occurrence of a spontaneous binding/unbinding event during the course of the simulation, which is unfortunately not yet feasible with unbiased simulations. An alternative approach is therefore to focus on the difference between the absolute binding free energy of two ligands, $\Delta\Delta G_{bind}$, also termed the relative binding free energy.[44] Here, FEP methods using MD simulations belong to the most accurate, but also computationally most demanding approaches. $\Delta\Delta G_{bind}$ can be calculated using the MD simulations by employing so-called 'alchemical' perturbations, where one ligand is transformed into the other once in solution and once bound to the protein.[44] A robust and widely applied FEP method is thermodynamic integration (TI), dating back to the work of John Kirkwood in 1935.[45] In TI, the system is perturbed in small steps along an artificial coupling parameter λ, which connects the two end states, and the resulting curve is integrated to yield ΔG. The system needs to be at equilibrium at each λ-step for accurate estimates of ΔG. Thus, sampling can be an issue in TI calculations – especially for the protein–ligand complexes. In many cases though sampling can be improved by combining TI with the sampling enhancement technique replica exchange (RE).[46] RE was first introduced in 1997[47] and two major variants are known. In temperature RE simulations, multiple copies (replicas) of the system are run in parallel at different temperature values and the exchange of conformations between copies is attempted at defined time intervals.[48] Hamiltonian RE, where the copies are simulated at different λ-values, is standardly used for TI.[46] Recently a combination of temperature and Hamiltonian RE called REST (replica exchange with solute tempering)[49] was proposed as an alternative to use with TI calculations. In REST, the end states (corresponding to λ=0 and λ=1) are at ambient temperature and the intermediate states at higher temperature values with λ=0.5 being the hottest replica. This enables faster sampling during the alchemical transformation process, while keeping the end states unchanged. Developments such as these in combination with the advances in computer power facilitate the accurate estimation of relative binding free energies of a chemical series within a practically useful time frame.

5. How Well Does it Dissolve?

The accurate prediction of aqueous solubility of compounds is an important and yet unresolved problem in drug discovery.[50,51] A strategy often used is to build predictive models based on experimental solubility data using for example simple descriptors based on the molecular structure[52,53] or machine learning methods.[54] Recently, deep learning with artificial neural networks was applied to the problem of solubility prediction and exposed the fundamental limitations of these approaches due to noisy experimental data used for training of the models.[55] The use of FEP calculations based on MD simulations would present an alternative approach that does not depend on experimental data as input. The thermodynamic cycle for this involves the ligand in vacuum because the direct transition between crystal and solution would require the occurrence of a dissolving/crystallization event during the course of the simulation, which is not yet feasible.[56–58] A first step towards solubility prediction by MD is therefore the accurate estimation of free energy of solvation in water, *i.e.* the free energy difference between the ligand in vacuum and in solution.[56,59] One important issue is the requirement to have a fast and accurate way to generate the force field parameters for the ligands. There have been many advances during the past decades in this area. Mobley *et al.*[59] used the general Amber force field (GAFF)[60] for ligand parameters to participate in the SAMPL blind test[61] containing 52 small drug-like molecules, and obtained errors of up to 14 kJ mol^{-1}. In the latest edition of this blind test, SAMPL4, with 47 molecules,[62] RMS errors in the range of 5 kJ mol^{-1} were found using GAFF,[63] the OPLS-AA force field,[64] or the GROMOS force field.[65] In an other study, the latest OPLS force field, OPLS2.0, has been found to perform better than other force fields like OPLS_2005,

GAFF, and CHARMm-MSI,[66] yielding errors in ΔG_{solv} between 0.5–7 kJ mol^{-1} (average error = 2.9 kJ mol^{-1}) over a different set of 239 small organic molecules where the experimental values were known.[67] There is clearly still room for improvement.

6. Leaving Flatland

Conformations generated by MD can be attractive for three-dimensional similarity search in addition to docking. In similarity search or ligand-based VS, structural information of the protein of interest is not available and therefore the similarity principle is employed, which states that similar molecules should have similar properties.[68,69] The similarity of two molecules, however, can be described in many different ways, with no globally best description found so far.[70,71] The majority of descriptors are based solely on the two-dimensional structure of the compound, which is consistently found to perform surprisingly well in retrospective studies[72–75] and, interestingly, generally better than three-dimensional similarity search methods.[76–80] The latter methods compare the shape of a specific conformation of a ligand with the shape of another ligand, with or without including pharmacophoric feature information.[81–87] They depend therefore crucially on the quality and number of the conformations used as well as the biological relevance of these conformations. The shortcomings of three-dimensional similarity methods in comparison with simple two-dimensional descriptors indicate that it is not yet possible – in a robust manner – to predict and compare the ligand conformations relevant for binding. The conformer generation approaches typically employed can be divided into two main categories based on their search algorithm: systematic and stochastic.[88] In systematic algorithms, conformations are generated by incrementing the torsion angles in small steps. These methods are therefore limited by the number of rotatable bonds in a molecule due to combinatorial explosion. In stochastic methods on the other hand, the conformational space is searched in a random manner, *e.g.* using genetic algorithms[89] or distance geometry.[90,91] As the conformations are generated without any information of the binding site, the diversity of the generated conformations is an important quality criterion when comparing different methods as it increases the chance of finding the biologically relevant ones among them.[92,93] MD has the advantage compared to these simpler methods to directly generate a Boltzmann-weighted ensemble of ligand conformations that considers concerted motions. This is especially important for macrocycles and peptides where the low-energy landscape cannot be sufficiently well approximated by rotatable bond combinations, causing the faster conformation generation methods therefore to fail.[94] An early example is here the study of 18-crown-6 ether by Sun and Kollman in 1992 using classical MD.[95] Later, the combination of MD with low-mode velocity filtering for the initial atom velocities was proposed to efficiently generate conformations.[94] The efficient sampling of low-strain energy conformations of macrocycles and peptidomimetics gains a new importance as these classes of molecules are currently rediscovered as potential drug candidates.[96–99] With the advances in computer power, MD is becoming applicable on a large, high-throughput scale to generate conformations of small organic molecules, macrocycles and peptides, given automated ligand parametrization and set-up procedures.

7. Machine Learning for MD

Unsupervised ML methods such as clustering algorithms are standardly used in MD to analyze the generated trajectories.[100,101] Supervised ML methods such as random forests (RF),[102] naïve Bayes or support-vector machines[103] are heavily used in cheminformatics for similarity search,[104–107] quantitative structure–activity relationships (QSAR),[108–110] and physiochemical property prediction.[54,55,111,112] They are, however, underrepresented in the MD community. The input of supervised ML methods is a set of features (*e.g.* substructures present in a molecule) and a class label (*e.g.* 0 or 1) in the case of classifiers or a continuous variable in the case of regression models. Given a new set of features the model predicts the class or the value of the continuous variable for this set. A possible application of ML methods for MD could therefore be the use in force field parametrization. Rai and Bakken[113] reported recently the use of RF regression models to predict the partial charges of small organic molecules. The RF models were trained using atom environment descriptors and partial charges derived from *ab initio* calculations for a training set of 80'000 molecules taken from the Pfizer library.[113] Other examples from the area of quantum-chemical methods are the prediction of molecular atomization energies using different regression models,[114,115] or the use of artificial neural networks to predict density function theory energies[116] or to represent the *ab initio* potential-energy surface of sodium.[117] These examples show that supervised ML methods could be a powerful tool for applications in the area of MD simulations.

8. Conclusions

Despite the many methodological advances in computational chemistry achieved during the past decades, many challenges remain open such as the handling of flexibility, *i.e.* dynamics, in docking and similarity search, or the fast and accurate prediction of binding free energies and free energies of solvation. In addition, new possibilities for cross-fertilization between the disciplines of computational chemistry emerge with the continuous increase in computer power, bringing them closer together and thus helping to improve accuracy and applicability.

Acknowledgements

I thank Gregory Landrum for helpful discussions and the ETH Zürich for its generous support.

Received: June 30, 2014

[1] J. A. McCammon, B. R. Gelin, M. Karplus, *Nature* **1977**, *267*, 585.
[2] J. E. Stone, D. J. Hardy, I. S. Ufimtsev, K. Schulten, *J. Mol. Graph. Model*. **2010**, *29*, 116.
[3] M. S. Friedrichs, P. Eastman, V. Vaidyanathan, M. Houston, S. Legrand, A. L. Beberg, D. L. Ensign, C. M. Bruns, V. S. Pande, *J. Comput. Chem.* **2009**, *30*, 864.
[4] N. Schmid, M. Bötschi, W.F. van Gunsteren, *J. Comput. Chem*. **2010**, *31*, 1636.
[5] A. W. Götz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, R. C. Walker, *J. Chem. Theory Comput.* **2012**, *8*, 1542.
[6] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K.J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, B. Towles, *Proc. Conf. High Performance Computing Networking, Storage and Analysis*, IEEE, **2009**.
[7] L. M. Mayr, D. Bojanic, *Curr. Opin. Pharmacol.* **2009**, *9*, 580.
[8] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, S. H. Bryant, *Nucleic Acids Res.* **2009**, *37*, W623.
[9] PubChem, *http://pubchem.ncbi.nlm.nih.gov*, accessed June 23, **2014**.
[10] 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.
[11] ChEMBL, *http://www.ebi.ac.uk/chembl/*, accessed June 23, **2014**.
[12] G. E. Dahl, J. Navdeep, R. Salakhutdinov, *arXiv: 1406.1231* [*stat.ML*], **2014**.
[13] G. Schneider, *Nat. Rev. Drug Discov.* **2010**, *9*, 273.
[14] J. M. Blaney, J. S. Dixon, *Perspect. Drug Discov. Des*. **1993**, *1*, 301.
[15] T. P. Lybrand, *Curr. Opin. Struct. Biol.* **1995**, *5*, 224.
[16] G. Jones, P. Willett, *Curr. Opin. Biotech.* **1995**, *6*, 652.
[17] R. D. Taylor, P. J. Jewsbury, J. W. Essex, *J. Comput. Aided Drug Des.* **2002**, *16*, 151.
[18] J. A. Erickson, M. Jalaie, D. H. Robertson, R. A. Lewis, M. Vieth, *J. Med. Chem.* **2004**, *47*, 45.
[19] D. Schneidman-Duhovny, R. Nussinov, H. J. Wolfson, *Proteins* **2007**, *69*, 764.

[20] C. Beier, M. Zacharias, *Expert Opin. Drug Discov.* **2010**, *5*, 347.
[21] E. Yuriev, M. Agostino, P. A. Ramsland, *J. Mol. Recogn.* **2010**, *24*, 149.
[22] E. Yuriev, P. A. Ramsland, *J. Mol. Recogn.* **2012**, *26*, 215.
[23] K. M. Elokely, R. J. Doerksen, *J. Chem. Inf. Model.* **2013**, *53*, 1934.
[24] A. M. Ferrari, B. Q. Wei, L. Costantino, B. K. Shoichet, *J. Med. Chem.* **2004**, *47*, 5076.
[25] A. R. Leach, *J. Mol. Biol.* **1994**, *235*, 345.
[26] S. C. Lovell, J. M. Word, J. S. Richardson, D. C. Richardson, *Proteins* **2000**, *40*, 389.
[27] I. W. Davis, D. Baker, *J. Mol. Biol.* **2009**, *385*, 381.
[28] S.-Y. Huang, Y. Zou, *Proteins* **2007**, *66*, 399.
[29] R. Shashidhar, P. C. Sanschagrin, J. R. Greenwood, M. P. Repasky, W. Sherman, R. Farid, *J. Comput. Aided Mol. Des.* **2008**, *22*, 621.
[30] I. R. Craig, J. W. Essex, K. Spiegel, *J. Chem. Inf. Model.* **2010**, *50*, 511.
[31] A. Di Nola, D. Roccatano, H. J. C. Berendsen, *Proteins* **1994**, *19*, 174.
[32] M. Mangoni, D. Roccatano, A. Di Nola, *Proteins* **1999**, *35*, 153.
[33] J.-H. Lin, A. L. Perryman, J. R. Schames, J. A. McCammon, *J. Am. Chem. Soc.* **2002**, *124*, 5632.
[34] A. J. Campbell, M. L. Lamb, D. Joseph-McCarthy, *J. Chem. Inf. Model.* **2014**, 54, online.
[35] C. Barillari, J. Taylor, R. Viner, J. W. Essex, *J. Am. Chem. Soc.* **2007**, *129*, 2577.
[36] R. Abel, T. Young, R. Farid, B. J. Berne, R. A. Friesner, *J. Am. Chem. Soc.* **2008**, *130*, 2817.
[37] S. Riniker, L. J. Barandun, F. Diederich, O. Krämer, A. Steffen, W. F. van Gunsteren, *J. Comput. Aided Mol. Des.* **2012**, *26*, 1293.
[38] I. Maffucci, A. Contini, *J. Chem. Theory Comput.* **2013**, *9*, 2706.
[39] A. Bortolato, B. G. Tehan, M. S. Bodnarchuk, J. W. Essex, J. S. Mason, *J. Chem. Inf. Model*. **2013**, *53*, 1700.
[40] L. Wang, B. J. Berne, R. A. Friesner, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1326.
[41] C. D. Christ, W. F. van Gunsteren, *J. Comput. Chem.* **2009**, *31*, 1664.
[42] S. Riniker, C. D. Christ, N. Hansen, A. E. Mark, P. C. Nair, W. F. van Gunsteren, *J. Chem. Phys.* **2011**, *135*, 024105.
[43] H. Gohlke, G. Klebe, *Curr. Opin. Struct. Biol*. **2001**, *11*, 231.
[44] W. F. van Gunsteren, H. J. C. Berendsen, *J. Comput. Aided Mol. Des.* **1987**, *1*, 171.
[45] J. G. Kirkwood, *J. Chem. Phys.* **1935**, *3*, 300.
[46] C. J. Woods, J. W. Essex, M. A. King, *J. Phys. Chem. B* **2003**, *107*, 13703.
[47] U. H. E. Hansmann, *Chem. Phys. Lett.* **1997**, *281*, 140.
[48] Y. Sugita, Y. Okamoto, *Chem. Phys. Lett.* **1999**, *314*, 141.
[49] P. Liu, B. Kim, R. A. Friesner, B. J. Berne, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 13749.
[50] A. J. Hopfinger, E. X. Esposito, A. Llinàs, R. C. Glen, J. M. Goodman, *J. Chem. Inf. Model.* **2009**, *49*, 1.
[51] J. Wang, T. Hou, *Comb. Chem. HTS* **2011**, *14*, 328.
[52] J. S. Delaney, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000.
[53] J. Wang, T. Hou, X. Xu, *J. Chem. Inf. Model.* **2010**, *49*, 571.
[54] D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2007**, *47*, 150.
[55] A. Lusci, G. Pollastri, P. Baldi, *J. Chem. Inf. Model*. **2013**, *53*, 1563.
[56] J. Westergren, L. Lindfors, T. Hoglund, K. Lüder, S. Nordholm, R. Kjellander, *J. Phys. Chem. B* **2007**, *111*, 1872.
[57] K. Lüder, L. Lindfors, J. Westergren, S. Nordholm, R. Kjellander, *J. Phys. Chem. B* **2007**, *111*, 1883.
[58] K. Lüder, L. Lindfors, J. Westergren, S. Nordholm, R. Kjellander, *J. Phys. Chem. B* **2007**, *111*, 7303.
[59] D. L. Mobley, C. I. Bayly, M. D. Cooper, K. A. Dill, *J. Phys. Chem. B* **2009**, *113*, 4533.
[60] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.
[61] A. Nicholls, D. L. Mobley, J. P. Guthrie, J. D. Chodera, C. I. Bayly, M. D. Cooper, V. S. Pande, *J. Med. Chem*. **2008**, *51*, 769.
[62] D. L. Mobley, K. L. Wymer, N. M. Lim, J . P. Guthrie, *J. Comput. Aided Mol. Des.* **2014**, *28*, 135.
[63] H. S. Muddana, N. V. Sapra, A. T. Fenley, M. K. Gilson, *J. Comput. Aided Mol. Des.* **2014**, *28*, 277.
[64] O. Beckstein, A. Fourrier, B. I. Iorga, *J. Comput. Aided Mol. Des.* **2014**, *28*, 265.
[65] K. B. Koziara, M. Stroet, A. K. Malde, A. E. Mark, *J. Comput. Aided Mol. Des.* **2014**, *28*, 221.
[66] F. A. Momany, R. Rone, *J. Comput. Chem.* **1992**, *13*, 888.
[67] D. Shivakumar, E. Harder, W. Damm, R. A. Friesner, W. Sherman, *J. Chem. Theory Comput.* **2012**, *8*, 2553.
[68] P. Willett, J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983.
[69] A. Bender, R. C. Glen, *Org. Biomol. Chem.* **2004**, *2*, 3204.
[70] R. P. Sheridan, S. K. Kearsley, *Drug Discov. Today* **2002**, *7*, 904.
[71] A. Bender, *Expert Opin. Drug Discov.* **2010**, *5*, 1141.
[72] J. Hert, P. Willett, D. J. Wilton, P. Ackling, K. Azzaoui, E. Jacoby, A. Schuffenhauer, *Org. Biomol. Chem.* **2004**, *2*, 3256.
[73] M. Sastry, J. F. Lowrie, S. L. Dixon, W. Sherman, *J. Chem. Inf. Model.* **2010**, *50*, 771.
[74] K. Heikamp, J. Bajorath, *J. Chem. Inf. Model*. **2011**, *51*, 1831.
[75] S. Riniker, G. A. Landrum, *J. Cheminf.* **2013**, *5*, 26.
[76] G. B. McGaughey, R. P. Sheridan, C. I. Bayly, J. C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, W. D. Cornell, *J. Chem. Inf. Model.* **2007**, *47*, 1504.
[77] M. von Korff, J. Freyss, T. Sander, *J. Chem. Inf. Model*. **2009**, *49*, 209.
[78] V. Venkatraman, V. I. Pérez-Nueno, L. Mavridis, D. W. Ritchie, *J. Chem. Inf. Model.* **2010**, *50*, 2079.
[79] G. Hu, G. Kuang, W. Xiao, W. Li, G. Liu, Y. Tang, *J. Chem. Inf. Model.* **2012**, *52*, 1103.
[80] M. Sastry, V. S. Sandeep Inakollu, W. Sherman, *J. Chem. Inf. Model*. **2013**, *53*, 1531.
[81] A. C. Good, W. G. Richards, *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112.
[82] J. A. Grant, M. A. Gallardo, B. T. Pickup, *J. Comput. Chem.* **1996**, *14*, 1653.
[83] P. J. Ballester, W. G. Richards, *J. Comput. Chem.* **2007**, *28*, 1711.
[84] P. Tosco, T. Balle, F. Shiri, *J. Comput. Aided Mol. Des.* **2011**, *25*, 777.
[85] M. Sastry, S. L. Dixon, W. Sherman, *J. Chem. Inf. Model.* **2011**, *51*, 2455.
[86] A. M. Schreyer, T. Blundell, *J. Cheminf.* **2012**, *4*, 27.
[87] A. Kalaszi, D. Szisz, G. Imre, T. Polgár, *J. Chem. Inf. Model.* **2014**, *54*, 1036.
[88] J. Gu, P. E. Bourne, 'Structural bioinformatics', 2nd ed. Wiley-Blackwell: Hoboken, NJ, USA, **2009**, Chap. 27, p. 639.
[89] O. Mekenyan, D. Dimitrov, N. Nikolova, S. Karabunarliev, *J. Chem. Inf. Model*. **1999**, *39*, 997.
[90] J. M. Blaney, J. S. Dixont, 'Distance geometry in molecular modeling', in: 'Reviews in Computational Chemistry', Eds. K. B. Lipkowia, D. B. Boyd, VCH Publishers, Inc.: New York, USA, **1994**, Vol. V, pp. 299–335.
[91] T. F. Havel, *Encyclopedia of Computational Chemistry* **1998**, *120*.
[92] D. K. Agrafiotis, A. C. Gibbs, F. Zhu, S. Izrailev, E. Martin, *J. Chem. Inf. Model.* **2007**, *47*, 1067.
[93] J.-P. Ebejer, G. M. Morris, C. M. Deane, *J. Chem. Inf. Model.* **2012**, *52*, 1146.
[94] P. Labute, *J. Chem. Inf. Model.* **2010**, *50*, 792.
[95] Y. Sun, P. A. Kollman, *J. Comput. Chem.* **1992**, *13*, 33.
[96] E. M. Driggers, S. P. Hale, J. Lee, N. K. Terrett, *Nature Rev. Drug Discov*. **2008**, *7*, 608.
[97] E. Marsault, M. L. Peterson, *J. Med. Chem*. **2011**, *54*, 1961.
[98] J. Mallinson, I. Collins, *Future Med. Chem*. **2012**, *4*, 1409.
[99] W. S. Horne, *Expert Opin. Drug Discov.* **2011**, *6*, 1247.
[100] J. Shao, S. W. Tanner, N. Thompson, T. E. Cheatham, *J. Chem. Theory Comput.* **2007**, *3*, 2312.
[101] B. Keller, X. Daura, W. F. van Gunsteren, *J. Chem. Phys*. **2010**, *132*, 074110.
[102] L. Breiman, *Mach. Learn.* **2001**, *45*, 5.
[103] N. Cristianini, J. Shawe-Taylor, 'An introduction to Support Vector Machines and other kernel-based learning methods', Cambridge University Press: Cambridge, UK, **2000**.
[104] H. Geppert, T. Horváth, T. Gärtner, S. Wrobel, J. Bajorath, *Chem. Biol. Drug Des*. **2011**, *77,* 30.
[105] D. Plewczynski, S. A. H. Spieser, U. Koch, *Comb. Chem. HTS* **2009**, *12*, 358.
[106] A. Bender, 'Bayesian methods in virtual screening and chemical biology', in 'Chemoinformatics and computational chemical biology', 'Methods in Molecular Biology', Ed. J. Bajorath, Springer: Totowa, NJ, USA, **2011**, Vol. 672, pp. 175–196.
[107] S. Riniker, N. Fechner, G. A. Landrum, *J. Chem. Inf. Model.* **2013**, *53*, 2829.
[108] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947.
[109] F. Hammann, J. Drewe, *Expert Opin. Drug Discov*. **2012**, *7*, 341.
[110] B. Chen, R. P. Sheridan, V. Hornak, J. H. Voigt, *J. Chem. Inf. Model.* **2012**, *52*, 792.
[111] M. C. Hutter, *Curr. Med. Chem.* **2009**, *16*, 189.
[112] A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413.
[113] B. K. Rai, G. A. Bakken, *J. Comput. Chem.* **2013**, *34*, 1661.
[114] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett*. **2012**, *108*, 058301.
[115] 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.
[116] R. M. Balabin, E. I. Lomakina, *J. Chem. Phys.* **2009**, *131*, 074104.
[117] H. Eshet, R. Z. Khaliullin, T. D. Kühne, J. Behler, M. Parrinello, *Phys. Rev. B* **2010**, *81*, 184107.