IOP Roadmap: Semiempirical methods

Christoph Bannwarth
Institute of Physical Chemistry; RWTH Aachen University, Aachen, Germany

Ben Hourahine
SUPA, Department of Physics, University of Strathclyde, Glasgow, UK

Jonathan Moussa
Molecular Sciences Software Institute, Blacksburg, VA 24060, USA

Status

Semiempirical electronic structure methods reduce the cost of solving the many-body Schrödinger equation by simple models and approximate solutions and mitigate the resulting errors with parameters fitted to reference data, either from experiments or higher levels of theory. Typically, they use a minimal atomic orbital basis set, parameterized multi-center integral approximations, and mean-field calculations based on Hartree-Fock (HF) theory or density-functional theory (DFT). The semiempirical Hückel method for π electrons was proposed only a year after HF theory in 1931, and it inspired more general models based on the zero-differential overlap (ZDO) approximation in the 1950's. By the 1980's, this had been further refined into the neglect of diatomic differential overlap (NDDO) approximation and developed into popular thermochemistry models such as AM1 and PM3, which are implemented in the MOPAC program [1].

The popularity of DFT in the early 1990's shifted most semiempirical method development from minimal-basis models to semiempirical density functionals with fitted parameters, and the last few decades of development has produced hundreds of new semiempirical density functionals but relatively few new minimal-basis models. While large-basis DFT calculations are typically more accurate than minimal-basis models, this accuracy comes at a roughly three orders of magnitude increase in computational cost. Semiempirical methods were also able to reduce the cost of DFT, and extended Hückel theory (EHT) from the 1960's inspired the development of density functional tight binding (DFTB) in the late 1990's, as implemented in software such as DFTB+ [2].

Even with steady growth in computing power, scientists still have limited computational budgets and often seek lower-cost methods, particularly when the size or number of systems is large or the required time to solution is short. Currently, semiempirical models are mainly used for explorations of conformational and chemical spaces and interactive quantum

Model name	PM7	GFN2-xTB	DFTB3/3OB-D4
Model family	MNDO	GFN	DFTB3
Parent software	MOPAC	xTB	DFTB+
Primary output	heat of formation	total energy	total energy
Reference data	heats, geometries, dipole moments, ionization potentials	geometries, forces, vibrational frequencies, non-covalent energies	energies, geometries, vibrational frequencies, barrier heights
Elemental coverage	H-La, Lu-Bi	H-Rn	H, C-F, Na, Mg, Zn, P-Cl, K, Ca, Br, I
Orbital type	orthogonal	non-orthogonal	non-orthogonal
Hopping integrals	Wolfsberg-Helmholz approximation of Slater-type orbitals	generalized Wolfsberg-Helmholz approximation of STO-nG orbitals	tabulated Slater-Koster matrix elements from atomic and diatomic DFT calculations
Coulomb integrals	NDDO approximation	multipole approximation	monopole approximation
Exchange energy	Fock exchange	density functional	density functional
Dispersion energy	short-range DH+ model	self-consistent D4 model	self-consistent D4 model

Table 1. Basic features and approximations of several popular semiempirical models.

mechanical studies, which continue to drive semiempirical model development. As shown in Table 1, the GFN family of models in the recent xTB program [3] combines the DFTB formalism with some design elements from EHT and atomic multipole expansions up to quadrupoles. There is also progress towards more unified software, with SCINE Sparrow [4] providing implementations of both NDDO-based and DFTB-based methods.

Current and Future Challenges

The applicability of semiempirical methods remains constrained for the following reasons: limited availability of suitable reference data combined with the employed Hamiltonian simplifications hinders their accuracy and transferability. Linear scaling of parameters with the number of elements has been a very successful strategy for the PM6/PM7 and the GFN-xTB methods in their Wolfsberg-Helmholz-type expressions to cover 70 and 86 elements of the periodic table, respectively. In contrast, the original DFTB models use the pairwise parameterized Slater-Koster tight-binding formalism, which has limited its model coverage of the periodic table. Nowadays, a plethora of quantum chemistry packages and powerful computers are available, enabling the fast generation of theoretical reference data at large scale. With enough data, parameters for nearly arbitrary elements and, possibly, element combination can be generated.

Additionally, existing approximations in contemporary semiempirical methods may require revision for improved accuracy, transferability to more diverse chemical environments, or extended applicability to a broader set of physical properties. One direction is to better understand and systematically improve established concepts such as the NDDO approximation [5]. Another direction is to incorporate more information and concepts from first-principles calculations as in done in composite methods such as PBEh-3c [6] and avoid the approximation of multi-center integrals altogether. Furthermore, the inclusion of more basis functions or core electrons to minimal-basis models may enable new spectroscopic applications like NMR or XAS. However, increasing the number of basis functions in semiempirical methods also increases their cost and thus reduces their computational advantage over first-principles methods.

Lastly, the computational scaling and efficiency of semiempirical models needs to be improved for both existing and future models. For all semiempirical schemes, the linear algebra necessary to solve for the density matrix is the rate-determining step. To compete with existing force-field methods, this step needs to be accelerated. Different schemes relying on fragmentation, sparse linear algebra and highly parallel computing architectures have been suggested [7], but only a few of them have been successfully applied in a black-box fashion on commodity computers at large scale [8].

Advances in Science and Technology to Meet Challenges

At the core of any model improvement in semiempirical methods will be the availability of more reference data: well-balanced, in large amounts, and preferably of high quality. The development of semiempirical methods will greatly benefit from the ongoing efforts to generate large data for machine-learning (ML) models. The ML priorities will likely be different and the resulting data might not be ideally suited for fitting new semiempirical models. Particularly, semiempirical models are different from purely geometry-based ML models, especially when extrapolation beyond the reference data space is important, such as in chemical space exploration and photochemistry. For this, it will be important that semiempirical Hamiltonians can be applied with appropriate wavefunctions for both the ground and excited states. While some software implementations of semiempirical methods already include excited-state and multi-determinant functionality, semiempirical models are primarily fit to reproduce single-determinant calculations of electronic ground states because that is what the vast majority of reference data is available for.

Even with sufficient data available, it may be challenging to choose between different model ingredients. ML machinery is effective at high-dimensional interpolation, and it is possible to generate semiempirical model parameters as the output of ML models, which improves the interpretability of the overall model relative to black-box ML predictions of total electronic energies [9]. Semiempirical models may also benefit in other ways from ML developments, particularly in accelerating rate-determining steps: improved initial guesses for SCF calculations and case-specific semiempirical parameter adjustments can both be aided by ML schemes. Alternatively, the framework of statistical model selection and tools such as the Akaike Information Criterion might be useful for selecting between semiempirical models with differing numbers of parameters. An improved formal understanding of semiempirical methods can also make these choices easier.

Similar to classical force fields, semiempirical models are well-suited to benefit from heterogeneous computing architectures that can leverage mixed-precision such as commodity GPUs, which enable much faster calculations than standard computing architectures [10]. This will likely increase the relevance of GPUs in quantum chemistry, which correspondingly follows their growth in ML applications.

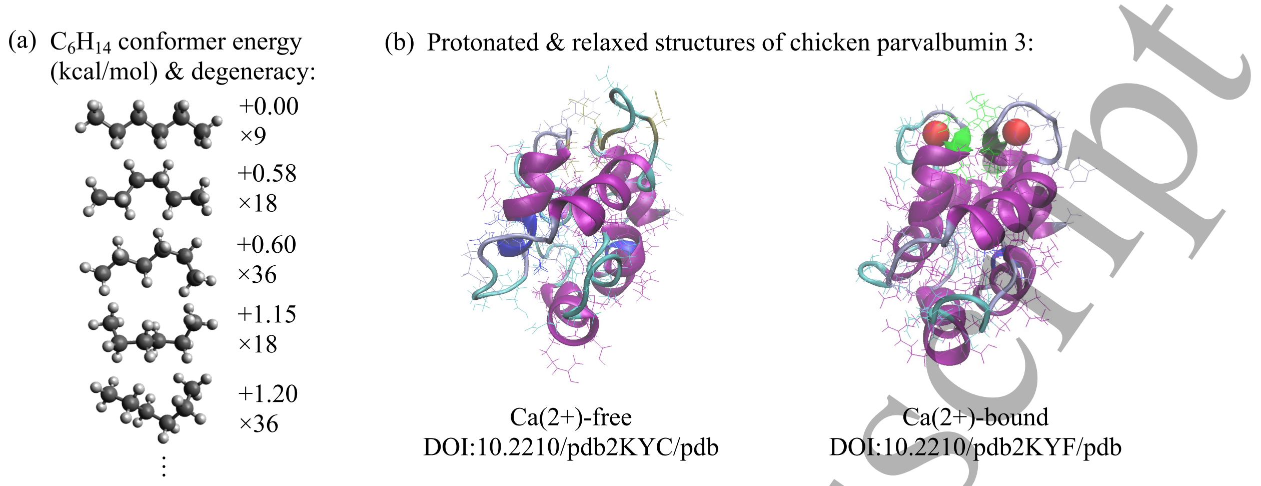

Figure 1. The low cost of semiempirical models enables novel functionality such as (a) conformer searches using GFN2-xTB and GBSA implicit water with CREST and (b) protein modeling using PM7 and COSMO implicit water with the MOZYME solver in MOPAC.

Concluding Remarks

Within electronic structure theory, semiempirical methods remained successful because of their unmatched computational efficiency. In recent years, models covering most of the periodic table have consolidated their role among computational chemists and materials scientists alike. Particularly, for chemical and conformational space exploration, examples of which are highlighted in Fig. 1, they are in frequent use. With plentiful reference data within reach, many remaining limitations might be remedied in the near future. Via modular software implementations, semiempirical Hamiltonians will become more generalizable than existing models or, alternatively, case-specific reparametrization will be highly simplified. Due to the generally low precision requirements, semiempirical models are well-suited to be combined with consumer-grade GPUs and linearly scaling algorithms. This will push the limits of routine applications that are possible with semiempirical models. Overall, semiempirical methods are as popular as ever and will remain so for the foreseeable future.

Acknowledgements

J.E.M. and the Molecular Sciences Software Institute are supported by NSF Grant No. CHE-2136142. C.B. acknowledges funding from the Ministry of Culture and Science of the German State of North Rhine-Westphalia (MKW) via the *NRW-Rückkehrprogramm*.

References

[1] James J. P. Stewart. MOPAC: A semiempirical molecular orbital program. *Journal of Computer-Aided Molecular Design*, 4(1):1–103, 1990.

[2] B. Hourahine, B. Aradi, V. Blum, F. Bonafé, A. Buccheri, C. Camacho, C. Cevallos, M. Y. Deshaye, T. Dumitrică, A. Dominguez, S. Ehlert, M. Elstner, T. van der Heide, J. Hermann, S. Irle, J. J. Kranz, C. Köhler, T. Kowalczyk, T. Kubař, I. S. Lee, V. Lutsker, R. J. Maurer, S. K. Min, I. Mitchell, C. Negre, T. A. Niehaus, A. M. N. Niklasson, A. J. Page, A. Pecchia, G. Penazzi, M. P. Persson, J. Řezáč, C. G. Sánchez, M. Sternberg, M. Stöhr, F. Stuckenberg, A. Tkatchenko, V. W.-z. Yu, and T. Frauenheim. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *The Journal of Chemical Physics*, 152(12):124101, 2020.

[3] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science*, 11(2):e1493, 2021.

[4] Francesco Bosia, Peikun Zheng, Alain Vaucher, Thomas Weymuth, Pavlo O. Dral, and Markus Reiher. Ultra-fast semi-empirical quantum chemistry for high-throughput computational campaigns with Sparrow. *The Journal of Chemical Physics*, 158(5):054118, 02 2023.

[5] Tamara Husch and Markus Reiher. Comprehensive analysis of the neglect of diatomic differential overlap approximation. *Journal of Chemical Theory and Computation*, 14(10):5169–5179, 10 2018.

[6] Stefan Grimme, Jan Gerit Brandenburg, Christoph Bannwarth, and Andreas Hansen. Consistent structures and interactions by density functional theory with small atomic orbital basis sets. *The Journal of Chemical Physics*, 143(5):054107, 2015.

[7] D R Bowler and T Miyazaki. $O(N)$ methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, 2012.

[8] Robert Schade, Tobias Kenter, Hossam Elgabarty, Michael Lass, Ole Schütt, Alfio Lazzaro, Hans Pabst, Stephan Mohr, Jürg Hutter, Thomas D. Kühne, and Christian Plessl. Towards electronic structure-based ab-initio molecular dynamics simulations with hundreds of millions of atoms. *Parallel Computing*, 111:102920, 2022.

[9] Frank Hu, Francis He, and David J. Yaron. Treating semiempirical hamiltonians as flexible machine learning models yields accurate and interpretable results. *Journal of Chemical Theory and Computation*, 19(18):6185–6196, 09 2023.

[10] Xin Wu, Axel Koslowski, and Walter Thiel. Semiempirical Quantum Chemical Calculations Accelerated on a Hybrid Multicore CPU–GPU Computing Platform. *Journal of Chemical Theory and Computation*, 8(7):2272–2281, 2012.