Next-generation semiempirical quantum mechanics

Jonathan E. Moussa
Molecular Sciences Software Institute, Blacksburg, VA

1 What I am trying to do

As I observed in a recent paper [DOI:10.1088/2516-1075/ab2022], available atomistic simulation capabilities span 10 orders of magnitude in computational cost per simulated atom. The cheapest 3 orders of magnitude are spanned by molecular mechanics (MM) simulations based on simple, inexpensive interatomic potentials, and the most expensive 3 orders of magnitude are spanned by various approximations of a full quantum mechanical (QM) description. In between is a handful of semiempirical QM (SQM) methods, all based on a common strategy of minimal-basis tight-binding models combined with interatomic potentials. This large gap in atomistic simulation capabilities is a substantial barrier to attempts at multiscale modeling connecting macroscopic continuum models to MM and then on down to the "first principles" of QM. QM/MM coupling is an ongoing challenge because of the large disparity between their costs and resolution of physical details. SQM should be a natural interface between QM and MM, but it is not yet capable of filling that role.

MM and QM both have active developer communities that are polarized by their emphasis on either maximizing performance (MM) or accuracy (QM). SQM focuses on performance/accuracy compromises that preserve the model transferability of QM, but such compromise is unpopular in academic research. Michael Dewar's group was the main chemistry SQM developer in the 1970's and 1980's, and one of his former group members, Jimmy Stewart, has continued its development to the present day with MOPAC [http://openmopac.net], outside of academia and mostly alone.

I have a technical plan for a new generation of SQM methods that build on the last 40 years of methodological development in electronic structure and adhere to modern standards in statistical modeling and data science. I will develop the key components of this new simulation capability, one at a time, publishing a paper and releasing an open-source software implementation for each component. Once I have all of the components for a minimum viable product, I will release new SQM simulation software. My primary simulation target is large, electronically inhomogeneous systems that are too expensive for QM tools and very difficult to design interatomic potentials for to enable MM tools. This is the single biggest deficiency in our atomistic simulation capability.

2 How it is done today and the limits of current practice

Historically, SQM has been used as a "proving ground" for QM method development in chemistry and physics. QM simulations in chemistry are usually based on Gaussian atomic orbitals (GAO), but a large number of orbitals per atom are needed to converge basis errors and a large number of matrix elements must be computed per orbital. SQM models based on atomic orbitals (AO) were able to reduce both the number of orbitals per atom and the number of matrix elements per orbital while retaining enough accuracy through parameter fitting to maintain experimental relevance. QM and SQM were competing capabilities in chemistry, and eventually density functional theory (DFT) tipped the balance in favor of QM in the 1990's. QM simulations in physics are usually based on

planewaves and pseudopotentials (PP). Unlike in chemistry, there was a gradual transition from SQM to QM in physics. Soft pseudopotentials fit to experimental data and requiring small basis sets [DOI:10.1103/PhysRev.141.789] gave way to hard pseudopotentials fit to simulation data and requiring large basis sets [DOI:10.1103/PhysRevLett.43.1494]. The pragmatic use of Hartree-like theories in physics was formalized by Kohn-Sham DFT and cross-bred with Hartree-Fock theory in chemistry during the 1990's to form the modern notion of hybrid density functionals.

The most active period for AO-SQM in chemistry was from 1965 to 1977 when 3 generations of models were developed by relaxing approximations from Complete to Intermediate and finally to Modified Neglect of Differential/diatomic Overlap (CNDO, INDO, and MNDO). The MNDO form [DOI:10.1021/ja00457a004] has persisted in every SQM thermochemistry model developed since 1977 with refinements of functional forms, reparameterizations, and extensions to d orbitals. Relative to GAO-DFT, these models are $\approx$ 1000$\times$ cheaper and $\approx$ 5$\times$ less accurate. According to Google Scholar, DFT is presently mentioned in 3$\times$ more scientific publications than SQM, which is suggestive of the large premium that is placed on accuracy by QM/SQM simulation users.

Besides AO-SQM in chemistry, there are only a few other actively-developed SQM simulation capabilities. The closest physics equivalent of MNDO-based models are density functional tight-binding (DFTB) models [DOI:10.1103/PhysRevB.58.7260] and tight-binding models of excitation are used in NEMO-3D to simulate semiconductor devices [DOI:10.1109/TED.2007.902879]. No total-energy models or actively-developed simulation capabilities are available for PP-SQM.

There is now far more method development in QM rather than SQM. Both GAO-based and PP-based QM have developed a substantial amount of methodological infrastructure, and any new development must operate effectively against a backdrop of existing infrastructure. Unfortunately, there are many examples of promising theoretical concepts such as linear-scaling solver algorithms [DOI:10.1103/RevModPhys.71.1085] and random-phase approximation (RPA) correlation models [DOI:10.1063/1.2977789] that don't perform well enough in practice to see widespread use in the context of GAO/PP-QM. The reduced computational overhead per atom in SQM could make some of these concepts more practically viable. For example, Fock exchange is at least 10$\times$ the cost of semilocal DFT in PP-DFT, but it is effectively free in MNDO-based SQM models.

Right now, the most popular approach to building models in the gap between QM and MM is machine learning. While it is certainly worthwhile to study the effectiveness of these flexible, general-purpose models for a variety of applications, domain-specific models should ultimately be superior when building efficient approximations to a known underlying theory as measured by information criteria in statistical model selection (i.e. similar accuracy with fewer parameters).

3 My new approach and why I think it will be successful

QM and SQM methods have three main sources of error: solvers, basis sets/matrix elements, and electron correlation models. QM methods operate in a regime where these errors are well separated: solver errors are reduced to near machine precision, basis set errors are systematically reduced as necessary, and correlation errors persist as the dominant error. SQM methods can and should operate in a regime that balances these sources of error to minimize the overall error at fixed cost: approximate solvers and smaller basis sets can reduce the cost of more accurate correlation models. Model parameters are resources that are introduced and expended to mitigate model errors. I will address these sources of error, in order of largest (correlation) to smallest (solver).

First, I will develop a new SQM-compatible correlation model based on efficiently computable ingredients. I have previously demonstrated that RPA calculations are less than 10× the cost of mean-field calculations when applied to SQM models [DOI:10.1063/1.4855255], which makes RPA a viable ingredient. Historically, electron correlation in SQM has been modeled only through the modification of model parameters in an otherwise correlation-free mean-field model except for pairwise atomic corrections to model London dispersion interactions. However, physics-based models use Hartree-like mean fields intended mainly for highly polarizable solids, and chemistry-based models use Hartree-Fock-like mean fields intended mainly for weakly polarizable molecules. From my contribution to semiempirical hybrid functional development [DOI:10.1063/1.4722993], I am convinced that one-size-fits-all mean fields used in modern DFT are a major source of error. RPA correlation is the simplest known physical model that naturally include both dispersion and screened exchange. I will develop a self-consistent RPA correlation model by identifying and repairing its largest errors to squeeze as much accuracy as possible out of this level of theory.

Next, I will merge the MNDO-based and pseudopotential-based SQM formalisms to combine their best features and eliminate their biggest problems. By combining a coarse grid (or other uniform basis) with disjoint atomic orbitals, we can maintain most of the efficiency of a minimal atomic basis while fixing its known failures (e.g. electrides, weakly bound anions, metal surfaces, negatively-charged semiconductor vacancies). As in the ubiquitous multiple-ζ GAO formalism, the introduction of a smooth basis enables localization of atomic orbital basis functions, which systematically suppresses diatomic overlap and controls the MNDO approximation. The smooth-atomic matrix elements can be parameterized alongside atomic pseudopotentials, and an electron-electron pseudopotential can mitigate the slow basis-set convergence of RPA correlation. There is inherent redundancy in parameterizing 2-electron matrix elements that I will avoid by using an SQM-compatible resolution of identity based on fictitious pseudo-wavefunction values at each atomic nucleus together with an auxiliary-basis coarse grid. This basic framework can be adjusted by tuning both the coarse grid spacing and the number of atomic orbitals for each element. I will control coarse-grid translation errors by independently tuning grid spacing and Gaussian width of the cardinal basis functions for a uniform grid of Gaussians [DOI:10.1016/j.amc.2009.08.037].

Finally, I will refine my recently proposed algorithmic concept of a localization self-energy [DOI:10.1088/2516-1075/ab2022] into a fixed-cost linear-scaling solver. This solver will be based on matrix inversion instead of matrix diagonalization, which enables the decoupled submatrices of the Hamiltonian associated with disjoint atomic orbitals to be downfolded into the coarse-grid Hamiltonian at a negligible cost. It will solve for the smallest self-energy capable of localizing a Green's function (i.e. Hamiltonian matrix inverse) within a predetermined length scale, and then systematically expand in powers of the self-energy to approach the exact Green's function. Costs can be minimized by amortizing the iterative self-energy optimization against some non-iterative high-order terms in the self-energy expansion. Linear-scaling solvers usually do not work well for metals, but I will avoid this problem by solving at fixed cost rather than at fixed accuracy.

Much of QM and SQM method development is presently trapped in various local minima because typically only a single simulation component is optimized at a time. My proposed design will escape from the familiar minima in design space by simultaneously innovating in correlation models, basis sets, and solver algorithms. These innovations are much easier to develop for SQM instead of QM because unnecessary complications can be encapsulated by parameterized models. MNDO was a sensible design in 1977, but with all of the advances in scientific computing and electronic structure theory over the last 40 years, it is now time to develop new SQM methods.

4 The difference it will make if I succeed and who should care

The main purpose of this project is to develop a new atomistic simulation capability, so its most direct impact would be on its prospective users. It will be useful to regular users of both MM and QM simulation software and help to bridge the gap between MM and QM simulation capabilities. MM simulations typically provide only structural properties, and small embedded QM regions are used to access localized electronic properties. Linear-scaling SQM with a low cost per atom will enable access to extended electronic properties by directly simulating full-system snapshots from long MM trajectories with the frequency of snapshots proportional to the relative cost of MM and SQM simulations. QM simulations usually require a cluster or supercell model to efficiently study small subsystems of large, inhomogeneous systems. SQM will be useful to independently test the convergence of these structural models and provide a sketch of a solution before committing to a substantially more expensive QM simulation. In both cases, SQM will provide direct simulation access to large, inhomogeneous systems with delocalized electronic properties that are otherwise not directly accessible with existing QM or MM simulation tools.

SQM has an important role in materials/chemical informatics as both a generator and consumer of data. The more organized and persistent SQM development efforts such as Jimmy Stewart's PMx models have critically depended on accumulating and organizing large amounts of high-quality reference data from experiment and high-level quantum chemistry. SQM developers should be cognizant of and active in modern efforts to organize data in chemistry and materials science and seek out their natural role in that ecosystem. In turn, SQM is capable of generating its own data at a much higher rate than QM methods. SQM is thus well-suited for both high-throughput screening and the fitting of interatomic potentials. While QM data will often be more accurate, SQM data is easier to generate "in situ", which is essential for fitting models that have low transferability and cannot easily incorporate data from the small, idealized systems accessible to QM simulations. For example, potentials for studying metal alloys could be generated from large, disordered supercells that are typical alloy realizations rather than smaller, more ordered supercells.

For the amount of use that it still gets, SQM has surprisingly few active developers. Renewed SQM activity might stimulate cooperation and competition with both QM and MM: more data to fit interatomic potentials while putting accuracy pressure on MM methods and prototypes for new QM methodological concepts while putting cost pressure on QM methods. SQM should seek out notions of compatibility with MM and QM methods to serve as a bridge in QM/MM coupling.

5 The risks

I am proposing to develop and implement 3 new methods, and each comes with risks. However, these methods have all been studied before in some form, and their adaptation to this project has already considered and mitigated their historic risks and problems. Overall, what I am proposing is more complicated than historic AO-SQM models and is intended to be more expensive ($\approx 10\times$) but more accurate (competitive with DFT). I foresee a lot of cost/accuracy flexibility in my design choices and many opportunities for introducing model parameters, but I cannot predict in advance the accessible cost-versus-accuracy phase space of SQM models. Also, there will be unforeseen challenges in making large-scale electronic structure simulations reliable, but this is an unavoidable problem that someone needs to confront head-on eventually, and I am ready to do so.

6 How much it will cost

Modern scientific research usually occurs within the context of established institutions, which often bear the primary costs of research that support their institutional mission and defray some costs for research that is at least aligned with their mission. I presently work at an NSF-funded institute that supports computational chemistry research through software development, support, and education but does not directly support internal basic research. In recognition of the importance of SQM to computational chemistry, the institute is supporting me to work with Jimmy Stewart to transition the MOPAC SQM software into an open-source software project as he prepares for retirement. As the most popular dedicated SQM software, the preservation of MOPAC will help maintain existing SQM simulation capabilities. However, I believe that new research is essential for SQM to attain its full modern potential and attract new developers who will build its future.

I am personally excited about and committed to the research project outlined in this white paper, regardless of its external support. An unpaid 20% of my work time is allocated to discretionary personal research efforts, and I am spending that all on this project. However, my rate of progress will be limited by this low level of activity, and I am seeking external funding to expand my commitment to this project to 40% of my work time. With my Virginia Tech research scientist salary and university overhead, the cost of this time is $\approx \$88,000$ per year. While this is a solo effort right now, I am looking for collaborations that might eventually justify a bigger budget.

7 How long it will take

The historic timescale of SQM development can be measured by the progress of Michael Dewar's research group and Jimmy Stewart's solo development. Pople's CNDO model from 1965 was Dewar's effective starting point, and his group produced MINDO in 1969, MINDO/3 in 1975, MNDO in 1977, AM1 in 1985, and SAM1 in 1993. AM1 was Stewart's effective starting point, and he produced PM3 in 1989, PM6 in 2007, and PM7 in 2013. The average SQM model development time is then 7 years, naïvely assuming that there were no gaps or breaks in development.

I believe that I can produce results faster than this historic rate for several reasons. I already have 20 years of experience in QM method development, and I already have published work (noted above) somewhat related to the goals of this project. There is now an abundance of QM reference data from the training of machine-learning models, I am aware of the historic pitfalls in SQM model parameterization, and I will continue to learn from Jimmy Stewart's experiences as I work with him on MOPAC. I am reasonably confident that I can deliver a minimum viable product of a cubic-scaling SQM model parameterized for H/C/N/O in 3 years and a linear-scaling version and parameterization of most other elements in 2 more years.

8 The metrics of short-term and long-term success

All of the technical milestones of this project will be accompanied by a research paper and a pilot software implementation that will become the basis for the new SQM simulation software. After a minimum viable software product has been released, I will measure success in terms of users that I can attract and retain and the research that they publish based on using this SQM software.