Supplement for "Gaussian Processes for Finite Size Extrapolation of Many-Body Simulations"

Edgar Josué Landinez Borda,[1, a)] Kenneth Berard,[1] Annette Lopez,[2] and Brenda Rubenstein[1, b)]

[1)] *Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA*

[2)] *Department of Physics, Brown University, Providence, Rhode Island 02912, USA*

(Dated: 25 March 2024)

I. MOTIVATING THE USE OF GPR BASED UPON THE ACCURACY OF ALTERNATIVE MODELS

Given the widespread use of relatively simple scaling laws often grounded in system physics for finite-size extrapolation, a useful question to ask is how our GPR method's accuracy compares to that of simpler 'fits' or regression models. As alluded to in the Introduction, the energies of systems with fixed geometries can be extrapolated to their thermodynamic limits using power laws that are a function of the system size. Interpolating among energies for different systems sizes has traditionally been approached using piece-wise polynomials such as cubic splines.[1] The accuracy of piece-wise polynomials is controlled by the number of knots, which in turn controls the number of pieces used to represent the function, and the order of the derivatives used to ensure the continuity of the knots. The number of nodes and their locations can be determined via cross-validation and feature selection of the most relevant parameters.

To analyze how these techniques perform, we apply them here just to our CCSD(T) data since they should perform similarly on our AFQMC data. We use these methods to fit E/N to the systems' natural geometric parameters: for the homogeneous chain, the intra-dimer distance, a, and the inverse of the system size, $1/N$; and for the inhomogeneous chain of dimers, a, $1/N$, and the inter-dimer distance, b. We use exactly the same training sets as used in the GPR regressions for consistency.

As an initial attempt, we tried using cubic and natural splines to perform our regressions, but these fits resulted in substantial errors: errors of 1 mHa for interpolations on systems with the same N and errors of up to 100 mHa for extrapolations to large N when $1/N$ was also used as a parameter, making them impractical for the accuracy targeted. The best technique we identified for the automatic selection of the knots and functions was the Multivariate Adaptive Regression Splines (MARS) algorithm.[1] We chose a Bayesian (BMARS) version of this method as implemented in pyBASS[2] in order to ensure robustness.

[a)] Electronic mail: edgar_landinez_borda@brown.edu

[b)] Electronic mail: brenda_rubenstein@brown.edu

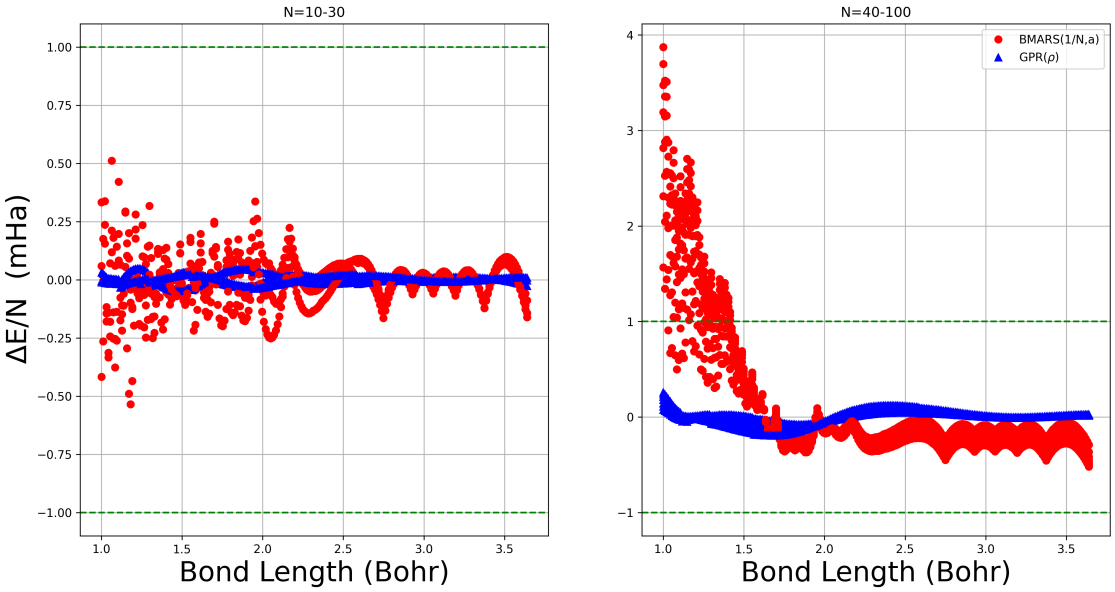

FIG. 1. (Left) Energy difference per atom, $\Delta E/N$, between the training and test data using BMARS on a data set of homogeneous hydrogen chains comprised of 10 to 30 atoms. (Right) $\Delta E/N$ for the validation set for hydrogen chains comprised of 40-100 atoms. The dashed lines denote the range between -1 and 1 mHa, which marks the range within chemical accuracy.

We thus begin by comparing the BMARS and GPR fits of the homogeneous chains of hydrogen atoms, as presented in Figure 1. Both methods perform well at interpolation for the 10-30-atom systems, making predictions with less than milliHartree errors. Both methods encounter greater difficulties extrapolating 40-100-atom energies, but GPR clearly outperforms BMARS across the validation dataset.

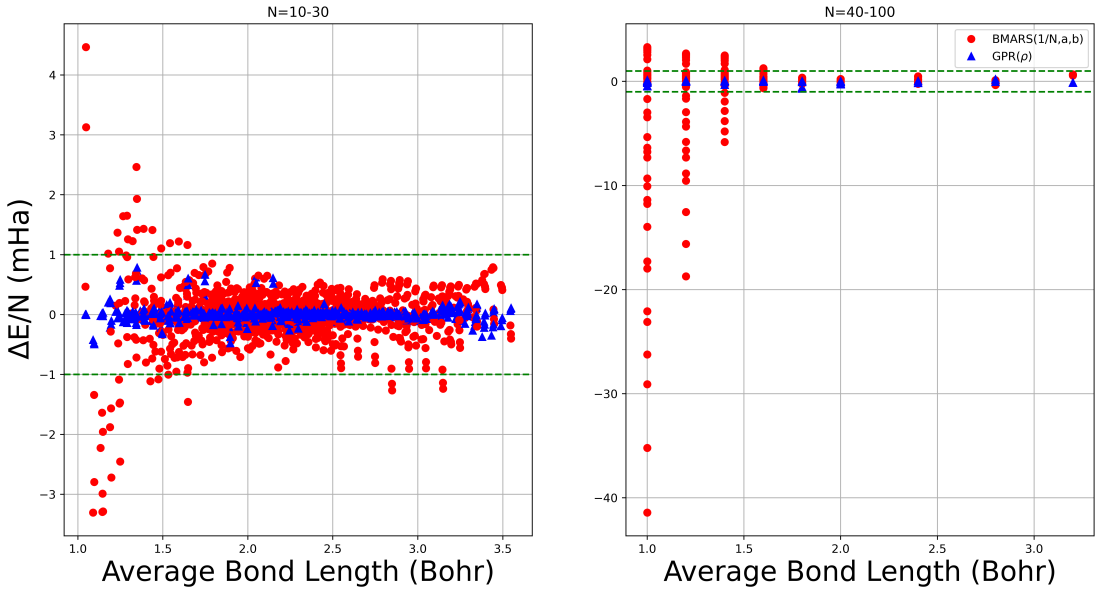

FIG. 2. (Left) Energy difference per atom, $\Delta E/N$, between the training and test set energy using GPR and BMARS on two-dimensional, heterogeneous hydrogen chains comprised of 5-15 dimers. (Right) Energy difference per atom for the validation set of two-dimensional hydrogen chains comprised of 20-50 dimers. The dashed lines denote the range between -1 and 1 mHa, which marks the range within chemical accuracy.

We observe that BMARS performs similarly on the inhomogeneous hydrogen chains of dimers (see Figure 2). During training and validation, BMARS performs well at interpolation, but struggles to carry this accuracy over to extrapolation, exhibiting errors of tens of mHa at larger system sizes. Such errors are too large for the high-accuracy study of many material systems.

Altogether, these results demonstrate that, while such piece-wise polynomials can successfully interpolate among a given set of system sizes and geometries, they struggle to extrapolate with milliHartree precision outside of their training region using a, b, and $1/N$ as their set of parameters.

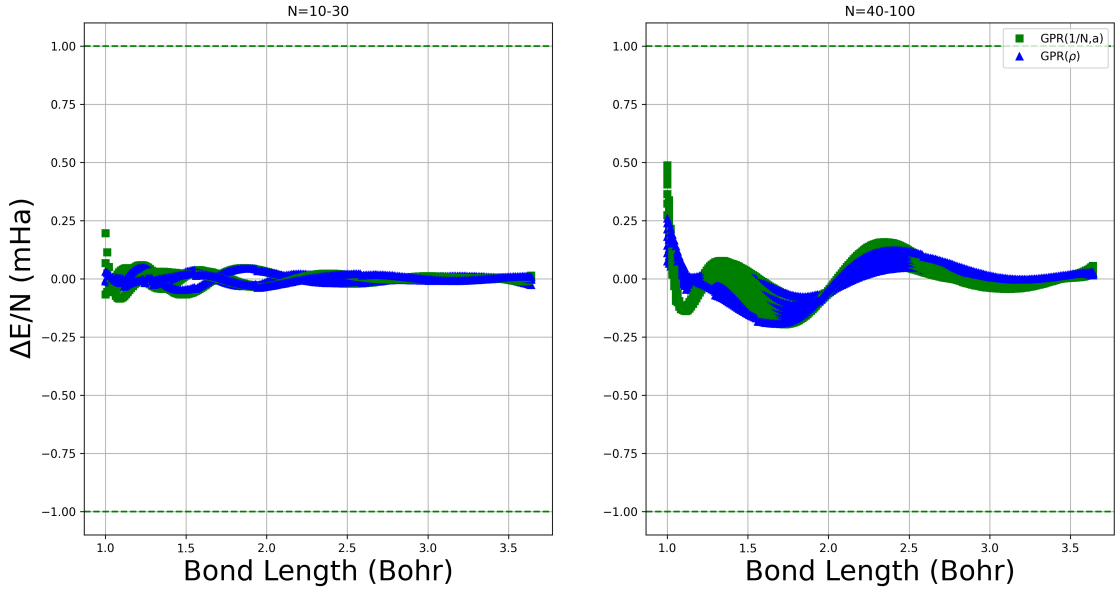

FIG. 3. (Left) Energy difference per atom, $\Delta E/N$, between the training and test sets for homogeneous hydrogen chains interpolated using GPR as a function of $1/N$ and a, the inverse of the number of atoms and the atom-atom separation, respectively. (Right) $\Delta E/N$ for the validation set of atomic hydrogen chains comprised of 40-100 atoms predicted using GPR as a function of $1/N$ and a. The dashed lines denote the range between -1 and 1 mHa, which marks the range within chemical accuracy.

This said, we can ask if there are more appropriate models than piece-wise polynomial regressions to extrapolate finite size effects with simple parameters. A more sophisticated set of regression techniques are those that employ polynomial kernels, including Kernel Ridge and Gaussian Process Regression. Given that we restricted our other regressions to using simple parameters, here, we also test the performance of our Gaussian or RBF kernel on the same simple set of parameters, a, b, and $1/N$. In Figure 3, we present the GPR results for the energy difference per atom using just these parameters, and compare it to our GPR results using the full charge density as featurized according to the main text, which we denote here as $GPR(\rho)$. GPR using just these parameters performs with greater than 1 mHa accuracy inside and outside the training region, as shown in the left- and right-hand panels of this figure, respectively.

GPR with these parameters even performs well for the inhomogeneous chains with nontrivial a and b dependence (see Figure 4). This performance is particularly encouraging given that performing GPR with just these parameters instead of the full density makes it less prone to over-fitting and potential linear dependencies that may arise if more complicated atomic environment descriptors were employed instead. Nonetheless, the use of such simple parameters would likely not be able to be generalized to more complex systems, substantiating the need for the more sophisticated atomic environment descriptors described in the main text.

This comparison of the performance of different regression schemes highlights Gaussian kernels' exceptional ability to capture nontrivial nonlinear behavior, motivating our choice of this kernel in this work.

II. DATABASE OF UHF, UCCSD, AND AFQMC ENERGIES

In the following tables, we provide the UHF, UCCSD, and AFQMC energies we used to train our models as a function of bond and chain length. All data were produced using the STO-6G minimal basis set and open boundary conditions, as described in the main text. We additionally present comparisons of the energies obtained by the various methods presented in this work in

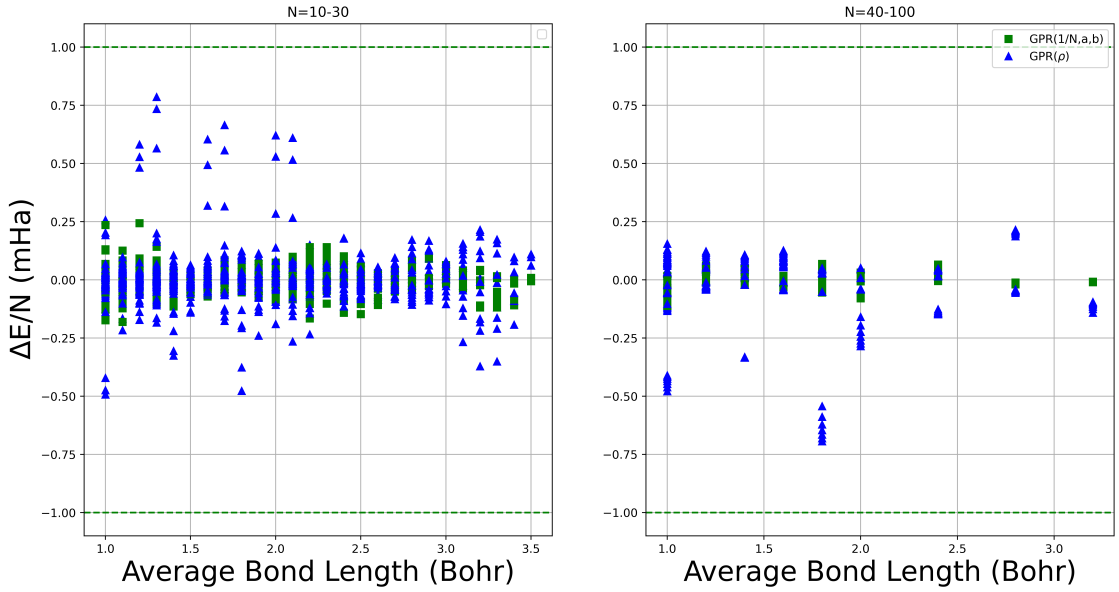

FIG. 4. (Left) Energy difference per atom, $\Delta E/N$, between the training and test sets using GPR with $1/N$, a, and b, the inverse of the number of atoms, intra-dimer spacing, and inter-dimer spacing for heterogeneous hydrogen chains of 5-15 dimers. (Right) $\Delta E/N$ using GPR with the same dscriptors for the validation set of hydrogen chains comprised of 20-50 dimers. The dashed lines denote the range between -1 and 1 mHa, which marks the range within chemical accuracy.

the thermodynamic limit to enable quantitative comparisons.

TABLE I. UHF total energies in Hartrees for hydrogen chains of size N, H_N.

R (Bohr)	H_{10}	H_{30}	H_{50}	H_{70}	H_{90}	H_{100}
1.0	-3.751741	-10.313656	-16.864876	-23.415169	-29.967430	-33.243780
1.2	-4.678055	-13.569081	-22.468054	-31.370884	-40.274063	-44.725647
1.4	-5.098620	-15.071059	-25.058908	-35.047455	-45.035988	-50.030247
1.6	-5.256281	-15.688718	-26.130188	-36.571700	-47.013206	-52.233960
1.8	-5.277449	-15.835906	-26.396688	-36.957470	-47.518250	-52.798637
2.0	-5.231365	-15.734367	-26.237770	-36.741173	-47.244576	-52.496277
2.4	-5.065709	-15.249867	-25.434025	-35.618183	-45.802340	-50.894420
2.8	-4.911826	-14.773006	-24.634186	-34.495365	-44.356544	-49.287136
3.2	-4.813234	-14.460917	-24.108599	-33.756280	-43.403960	-48.227802
3.6	-4.760347	-14.291805	-23.823265	-33.354725	-42.886185	-47.651913

TABLE II. UCCSD total energies in Hartrees for hydrogen chains of size N, H_N.

R (Bohr)	H_{10}	H_{30}	H_{50}	H_{70}	H_{90}	H_{100}
1.0	-3.823875	-10.551319	-17.272170	-23.989079	-30.705048	-34.063330
1.2	-4.765611	-13.846666	-22.924540	-32.005486	-41.088450	-45.630295
1.4	-5.204019	-15.378910	-25.558884	-35.743183	-45.928500	-51.021250
1.6	-5.382921	-16.022041	-26.674374	-37.328835	-47.983486	-53.310820
1.8	-5.417419	-16.194029	-26.983278	-37.773026	-48.562782	-53.957660
2.0	-5.375040	-16.113564	-26.859707	-37.605904	-48.352104	-53.725200
2.4	-5.204174	-15.639478	-26.075703	-36.511925	-46.948150	-52.166264
2.8	-5.028207	-15.109411	-25.190659	-35.271904	-45.353150	-50.393776
3.2	-4.894827	-14.700113	-24.505402	-34.310688	-44.115980	-49.018623
3.6	-4.809517	-14.437078	-24.064636	-33.692196	-43.319756	-48.133537

TABLE III. UCCSD(T) total energies in Hartrees for hydrogen chains of size N, H_N.

R (Bohr)	H_{10}	H_{30}	H_{50}	H_{70}	H_{90}	H_{100}
1.0	-3.824320	-10.555888	-17.282262	-24.006498	-30.729299	-34.090722
1.2	-4.766261	-13.852863	-22.940150	-32.028065	-41.117762	-45.663033
1.4	-5.204920	-15.389607	-25.576883	-35.768427	-45.961339	-51.057937
1.6	-5.384141	-16.034008	-26.693585	-37.355965	-48.018644	-53.349999
1.8	-5.421750	-16.206099	-27.003137	-37.800911	-48.598704	-53.997600
2.0	-5.380497	-16.125340	-26.879149	-37.633042	-48.386939	-53.763884
2.4	-5.207534	-15.648310	-26.090298	-36.532284	-46.974273	-52.195268
2.8	-5.029597	-15.113596	-25.197649	-35.281701	-45.365752	-50.407782
3.2	-4.895217	-14.701345	-24.507475	-34.313603	-44.119735	-49.022801
3.6	-4.809595	-14.437327	-24.065058	-33.692789	-43.320521	-48.134389

TABLE IV. AFQMC total energies in Hartree for hydrogen chains of size N, H_N.

R (Bohr)	H_{10}	H_{30}	H_{50}	H_{70}	H_{90}	H_{100}
1.0	-3.824841	-10.556912	-17.284995	-24.011174	-30.738827	-34.102049
1.2	-4.766713	-13.853941	-22.947219	-32.043003	-41.137732	-45.685495
1.4	-5.204092	-15.395489	-25.593646	-35.792978	-45.993014	-51.092832
1.6	-5.381873	-16.048077	-26.719823	-37.391189	-48.062583	-53.398691
1.8	-5.421827	-16.228916	-27.038958	-37.849788	-48.659628	-54.063537
2.0	-5.386939	-16.155198	-26.926919	-37.696240	-48.466989	-53.852305
2.4	-5.226785	-15.693020	-26.159033	-36.626692	-47.094404	-52.328345
2.8	-5.050066	-15.163181	-25.276574	-35.390147	-45.503771	-50.560391
3.2	-4.910878	-14.741508	-24.571581	-34.402390	-44.234783	-49.149157
3.6	-4.818507	-14.461193	-24.104561	-33.748037	-43.390617	-48.212736

TABLE V. AFQMC errors on the total energies in Hartrees hydrogen chains of size N, H_N.

R (Bohr)	H_{10}	H_{30}	H_{50}	H_{70}	H_{90}	H_{100}
1.0	0.000190	0.000251	0.000339	0.000430	0.000469	0.000391
1.2	0.000241	0.000265	0.000316	0.000338	0.000327	0.000370
1.4	0.000286	0.000278	0.000366	0.000357	0.000450	0.000473
1.6	0.000365	0.000275	0.000343	0.000501	0.001101	0.000625
1.8	0.000261	0.000344	0.000439	0.000479	0.000557	0.000887
2.0	0.000283	0.000401	0.000522	0.000651	0.001640	0.000674
2.4	0.000213	0.000791	0.000929	0.001097	0.001491	0.002093
2.8	0.000294	0.001183	0.001661	0.002078	0.002167	0.003005
3.2	0.000575	0.001543	0.002180	0.002609	0.003869	0.003401
3.6	0.001115	0.002181	0.003203	0.004396	0.005493	0.006210

III. COMPARISON OF POLYNOMIAL EXTRAPOLATIONS TO THE THERMODYNAMIC LIMIT

TABLE VI. Comparison of the energy per atom in Hartrees produced by the different methods studied in this work in the Thermodynamic Limit (TDL). TDL_{REF} denotes the thermodynamic limit obtained by Motta *et al.*[3], while TDL denotes the thermodynamic limit obtained based upon the data sets constructed in this work. In all of these cases, the limits were obtained by fitting to the polynomial $\frac{E}{N}(N\to\infty) = a_0 + a_1\frac{1}{N} + a_2\frac{1}{N^2}$. The regressions were fit using chains of size N=10, 30, and 50.

R(Bohr)	$\mathrm{TDL}_{REF}^{UCCSD(T)}$	$\mathrm{TDL}^{UCCSD(T)}$	$\mathrm{TDL}_{REF}^{AFQMC)}$	TDL^{AFQMC}
1.0	-0.336253	-0.336348	-0.336315	-0.336448
1.2	-0.454373	-0.454417	-0.454661	-0.454781
1.4	-0.509397	-0.509417	-0.509990	-0.510011
1.6	-0.533101	-0.533107	-0.533561	-0.533664
1.8	-0.540012	-0.540010	-0.540486	-0.540541
2.0	-0.537803	-0.537800	-0.538462	-0.538628
2.4	-0.522115	-0.522111	-0.523259	-0.523297
2.8	-0.504202	-0.504201	-0.505556	-0.505672
3.2	-0.490308	-0.490305	-0.491446	-0.491496
3.6	-0.481386	-0.481386	-0.482104	-0.482176

TABLE VII. Comparison of the energy per atom in Hartrees in the Thermodynamic Limit (TDL) produced using polynomial regression (Poly), Gaussian Process Regression (GPR), and the subtraction trick on chains of different lengths, $E_{ST}(N_1,N_2)$, for our UCCSD(T) database. $\pm 1.96\sigma$ denotes the error on the GPR regression.

R(Bohr)	Poly	GPR	$\pm(1.96 * sigma)$	$\mathrm{E}_{ST}(10,30)$	$\mathrm{E}_{ST}(10,50)$	$\mathrm{E}_{ST}(30,50)$
1.0	-0.336348	-0.336489	0.000472	1.004522	0.670611	0.336524
1.2	-0.454417	-0.454469	0.000144	1.281626	0.868086	0.454482
1.4	-0.509417	-0.509415	0.000095	1.464126	0.986769	0.509415
1.6	-0.533107	-0.532809	0.000095	1.599661	1.066011	0.532735
1.8	-0.540010	-0.539650	0.000091	1.710015	1.124563	0.539560
2.0	-0.537800	-0.537656	0.000090	1.806700	1.172070	0.537620
2.4	-0.522111	-0.522315	0.000070	1.983167	1.252894	0.522366
2.8	-0.504201	-0.504299	0.000030	2.156302	1.330374	0.504324
3.2	-0.490305	-0.490307	0.000024	2.335458	1.412884	0.490308
3.6	-0.481386	-0.481408	0.000039	2.522079	1.501760	0.481414

TABLE VIII. Comparison of the energy per atom in Hartrees in the Thermodynamic Limit (TDL) produced using polynomial regression (Poly), Gaussian Process Regression (GPR), and the subtraction trick on chains of different lengths, $E_{ST}(N_1,N_2)$, for our AFQMC database. $\pm 1.96\sigma$ denotes the error on the GPR regression.

R(Bohr)	Poly	GPR	$\pm(1.96 * sigma)$	$\mathrm{E}_{ST}(10,30)$	$\mathrm{E}_{ST}(10,50)$	$\mathrm{E}_{ST}(30,50)$
1.0	-0.336448	-0.336288	0.000421	1.004672	0.670360	0.336248
1.2	-0.454781	-0.454897	0.000133	1.282172	0.868621	0.454926
1.4	-0.510011	-0.509849	0.000089	1.465017	0.987311	0.509809
1.6	-0.533664	-0.533453	0.000057	1.600496	1.066816	0.533400
1.8	-0.540541	-0.540422	0.000041	1.710812	1.125528	0.540392
2.0	-0.538628	-0.538454	0.000031	1.807942	1.173068	0.538411
2.4	-0.523297	-0.523341	0.000025	1.984946	1.254176	0.523352
2.8	-0.505672	-0.505664	0.000023	2.158508	1.332080	0.505662
3.2	-0.491496	-0.491534	0.000022	2.337244	1.414418	0.491544
3.6	-0.482176	-0.482165	0.000035	2.523264	1.502706	0.482162

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2013. ISBN 9780387216065. URL https://books.google.com.co/books?id=yPfZBwAAQBAJ.

[2] Devin Francom and Bruno Sansó. Bass: An r package for fitting and performing sensitivity analysis of bayesian adaptive spline surfaces. *Journal of Statistical Software*, 94(8), 1 2020. ISSN 1548-7660. doi:10.18637/jss.v094.i08. URL https://www.osti.gov/biblio/1835765.

[3]Mario Motta, David M. Ceperley, Garnet Kin-Lic Chan, John A. Gomez, Emanuel Gull, Sheng Guo, Carlos A. Jiménez-Hoyos, Tran Nguyen Lan, Jia Li, Fengjie Ma, Andrew J. Millis, Nikolay V. Prokof'ev, Ushnish Ray, Gustavo E. Scuseria, Sandro Sorella, Edwin M. Stoudenmire, Qiming Sun, Igor S. Tupitsyn, Steven R. White, Dominika Zgid, and Shiwei Zhang. *Phys. Rev. X*, 7:031059, Sep 2017. doi:10.1103/PhysRevX.7.031059. URL https://link.aps.org/doi/10.1103/PhysRevX.7.031059.