Electronic Supplementary Information (ESI)

Robust molecular representations for modelling and design derived from atomic partial charges

A. R. Finkelmann[a], A. H. Göller[b] and G. Schneider[a]

[a] Department of Chemistry and Applied Biosciences, ETH Zürich, Vladimir-Prelog-Weg 4, CH-8093 Zürich, Switzerland; [b] Bayer HealthCare AG, Global Drug Discovery, Aprather Weg 18a, D-42096 Wuppertal, Germany

Contents **page**

S.1 Computational Details

Initial 3D coordinates of all structures were generated with CORINA.[1]

Sildenafil and Sildenafil+ conformers were sampled with the systematic search algorithm implemented in MOE.[2] This algorithm was not feasible for Atorvastatin anymore. Thus, the low mode MD search[3] was invoked. The energy cutoff for conformations was chosen to be 15 kcal/mol (force field potential energy). Conformers differing less than 0.75 Å were deemed equal. Conformers were optimized and scored with help of the MMFF94x force field as implemented in MOE.

Gaussian09 [4] DFT calculations were accelerated with the density fitting approximation.[5] Dispersion corrections were calculated with the Grimme D3 scheme employing Becke—Johnson damping. [6,7] Relative energies of conformers in the colored distribution plots (Figs. S10 – S38) are obtained from the DFT optimized structures. Note, that the DFTB energy comprises the D3 correction for dispersion interactions. ESP partial charges were calculated with the Hu, Lu, Yang scheme[8] but applying Gaussian standard atomic densities. The charges were restrained to reproduce the dipole moment. For method evaluation we chose in addition to the TPSS exchange-correlation functional[9] the Hartree—Fock method (HF), the PBE exchange-correlation functional[10,11] and the B3LYP exchange-correlation functional[12,13,14] in combination with the STO-3G,[15,16] SV,[17] def2-SVP (abbreviated SVP) and def2-TZVP (abbreviated TZVP)[18,19] basis sets. All charge values are given in Hartree atomic units (a.u.). AM1[20] and PM6[21] calculations were conducted with Gaussian09.

DFTB single point calculations (DFTB3 variant) were performed with the DFBT+ program package and latest available parameters.[22,23,24,25] DFTB energies include the D3 dispersion interaction correction.[26]

Plots, linear regressions and autocorrelation function calculations were conducted/accomplished with in-house python programs and the SciPy, Numpy[27], Matplotlib[28], Scikit-learn[29] python libraries.

S.2 Structural depiction of Sildenafil+ and its sampled conformations

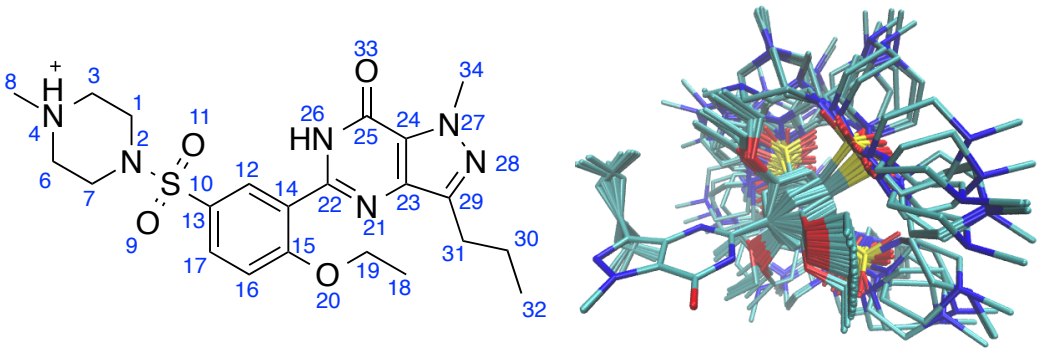

Fig. S1: Lewis representation and superimposed conformational ensemble of Sildenafil+.

S.3 Boxplots of charge distributions on carbon atoms for the conformational ensembles of Atorvastatin and Sildenafil+

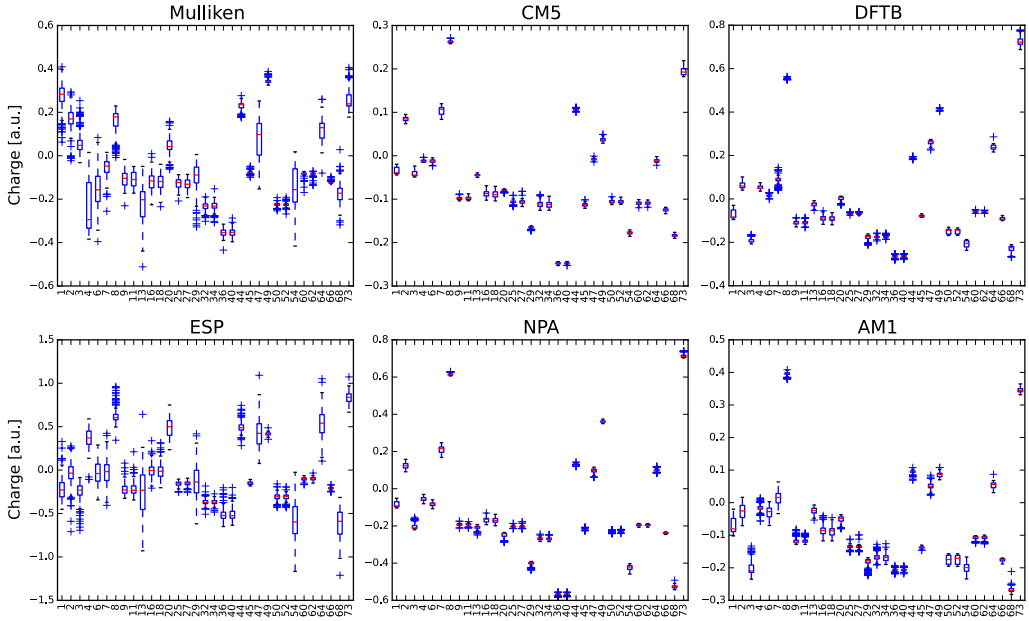

Fig. S2: Boxplots of partial charge distributions on carbon atoms of Atorvastatin. Atom indices are given on the x-axis.

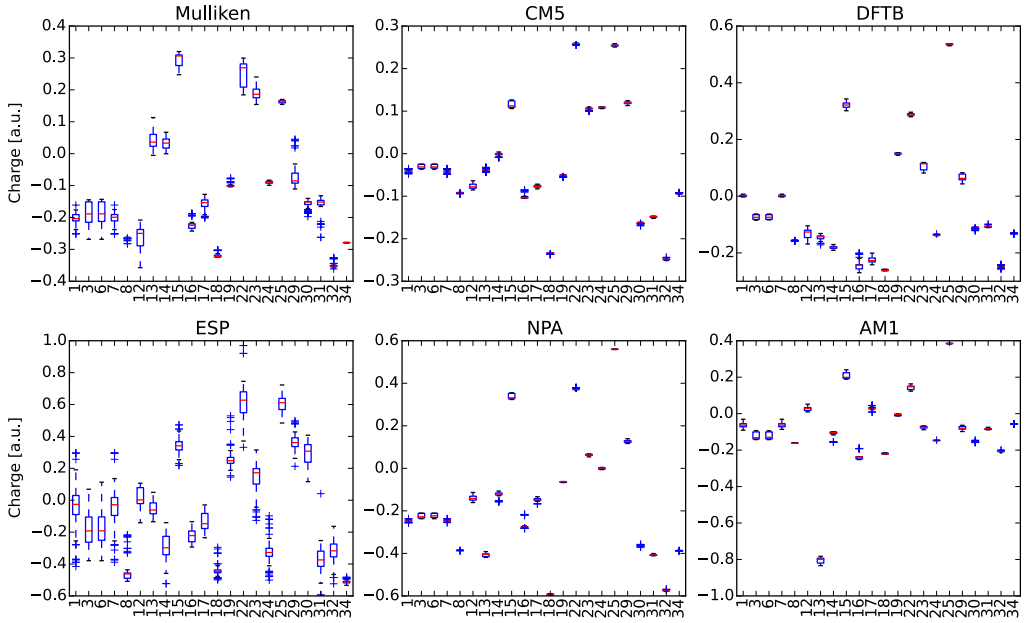

Fig. S3: Boxplots of partial charge distributions on carbon atoms of Sildenafil+. Atom indices are given on the x-axis.

S.4 Example of ESP charges resulting in large conformation dependent variations

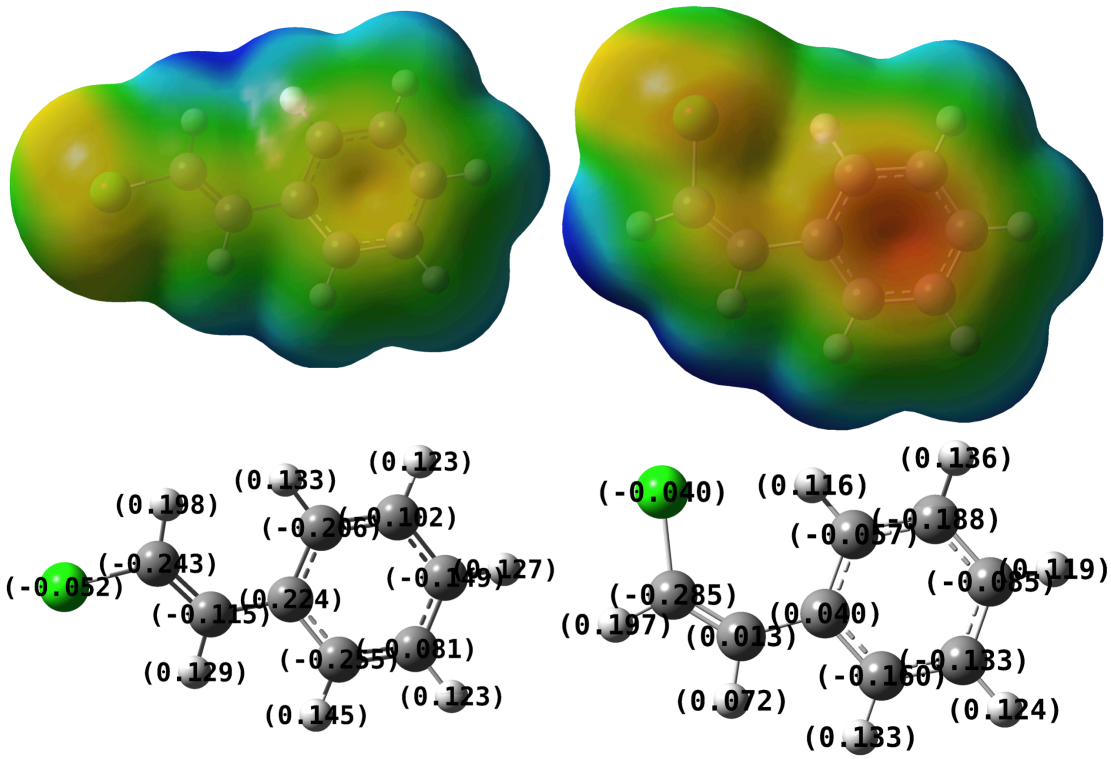

Fig. S4: Illustration of impact of cis/trans conformation of double bond on ESP charges. The cis/trans change of the double bond leads to large changes of charges on carbon atoms, especially in the phenyl ring. The ESP charges are given in a.u.

S.5 Basis set and density functional dependency of partial charge schemes

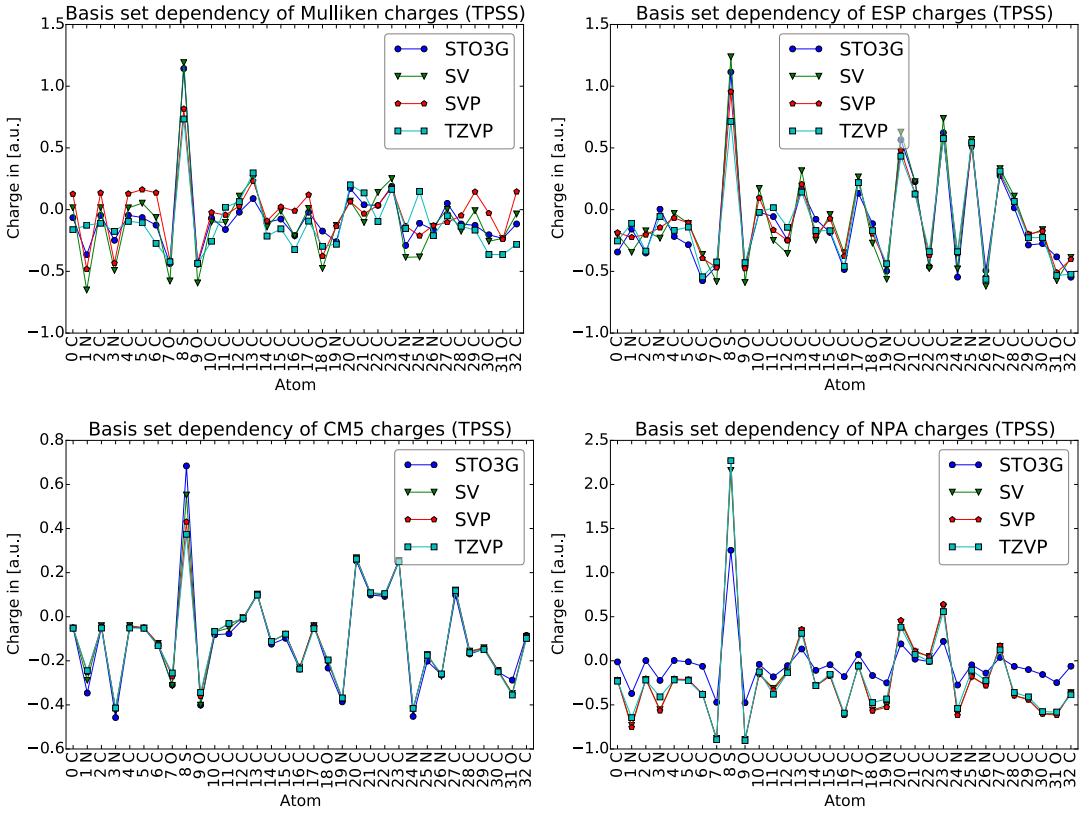

Fig. S5: Basis set dependency of partial charge schemes on most stable Sildenafil structure.

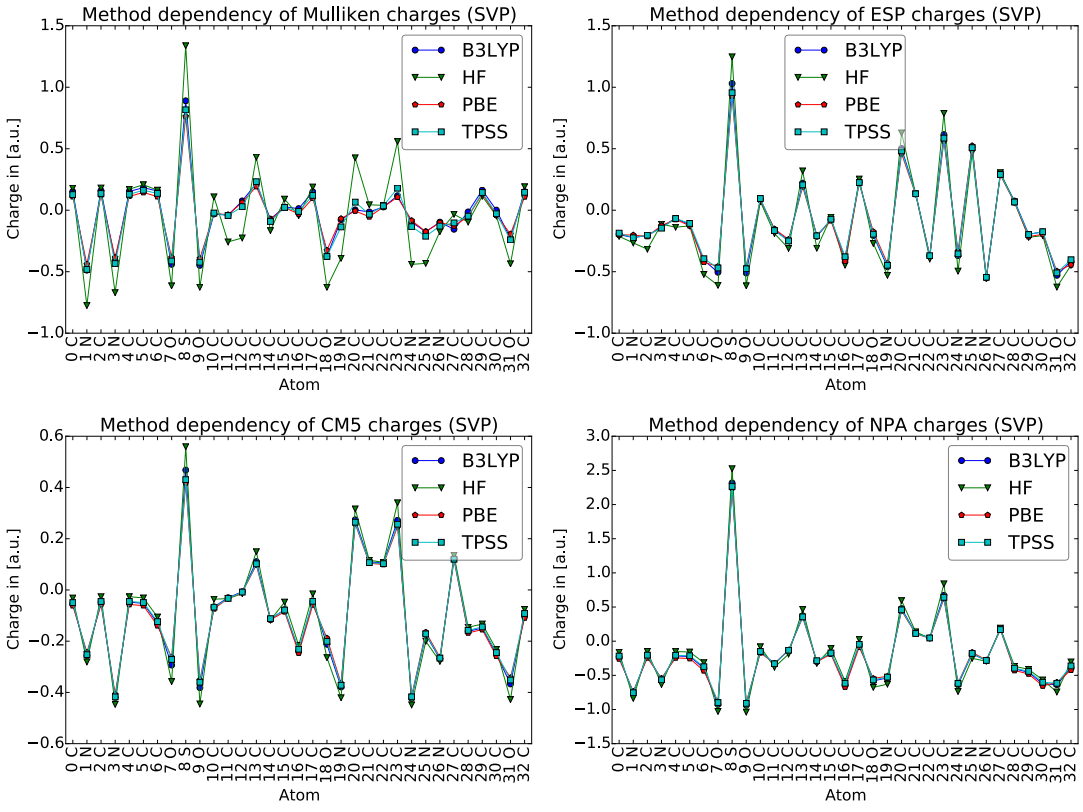

Fig. S6: Method dependency of partial charge schemes on most stable Sildenafil structure.

S.6 Correlation of conformer energies of AM1, PM6 and DFTB energies vs. DFT energies

S.6.1 Singlepoint calculations on DFT optimized structures

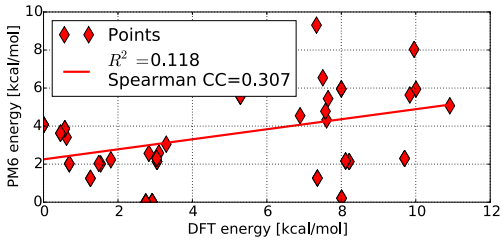

Fig. S7: Correlation of PM6 energy with DFT energy for conformational ensemble of Sildenafil. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

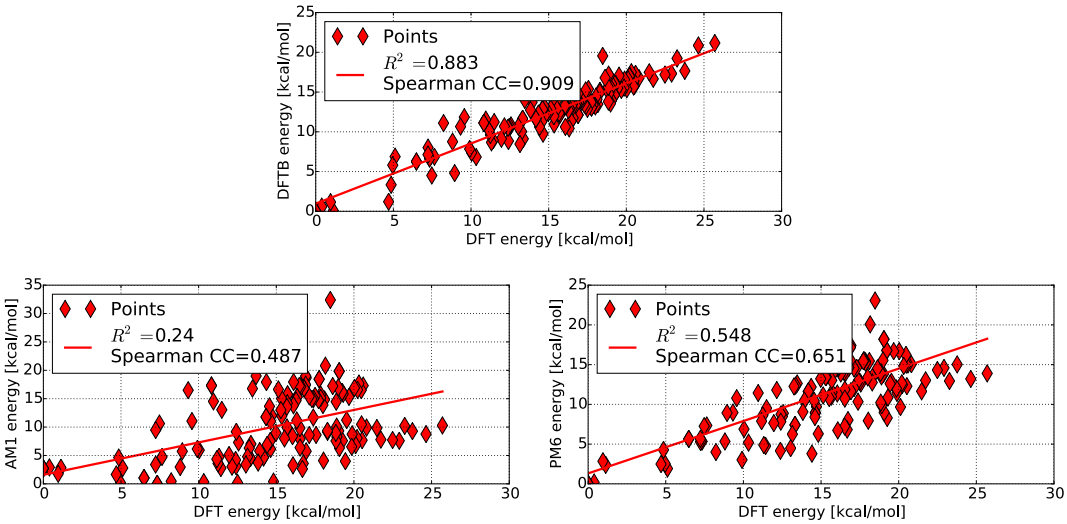

Fig. S8: Correlation of DFTB ,AM1 and PM6 energy with DFT energy for conformational ensemble of Atorvastatin. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

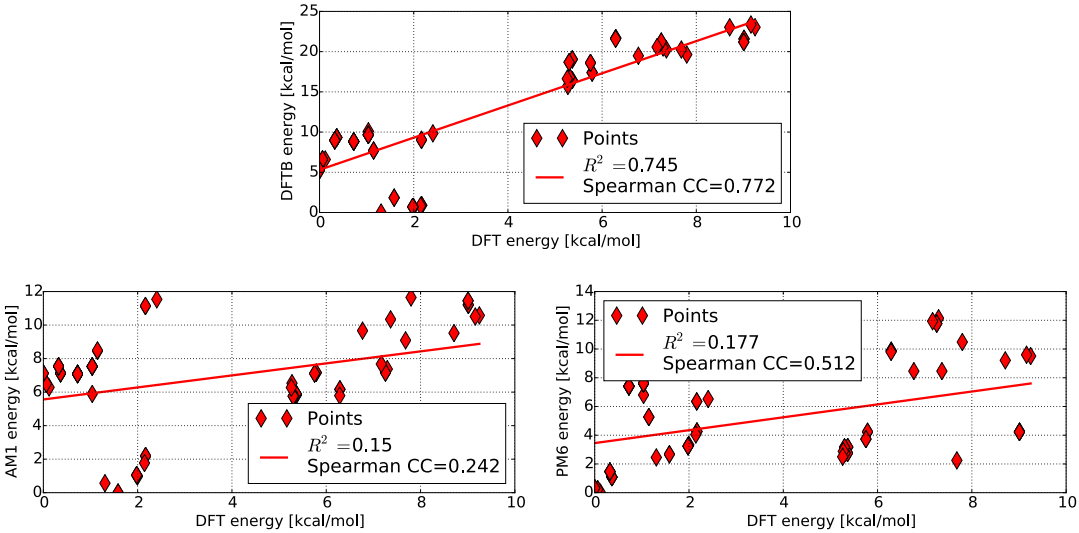

Fig. S9: Correlation of DFTB, AM1 and PM6 energy with DFT energy for conformational ensemble of Sildenafil+. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

S.6.2 Structure optimized with respective method

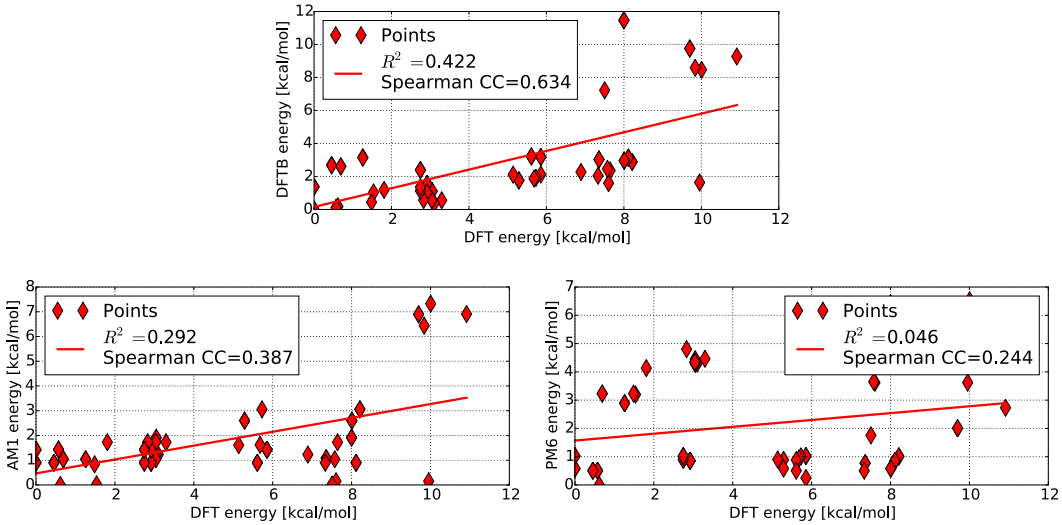

Fig. S10: Correlation of DFTB, AM1 and PM6 energy with DFT energy for conformational ensemble of Sildenafil. Structures were optimized with the respective method. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

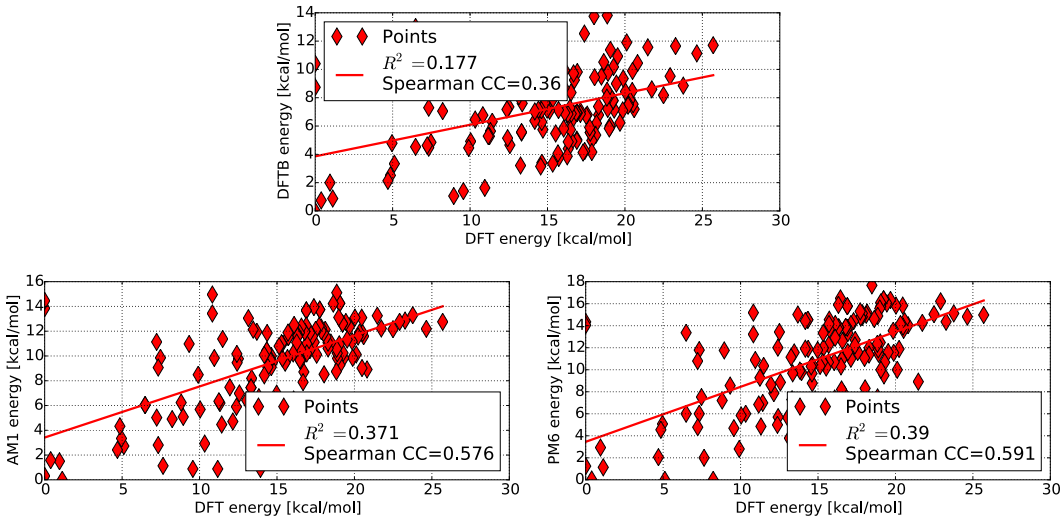

Fig. S11: Correlation of DFTB, AM1 and PM6 energy with DFT energy for conformational ensemble of Atorvastatin. Structures were optimized with the respective method. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

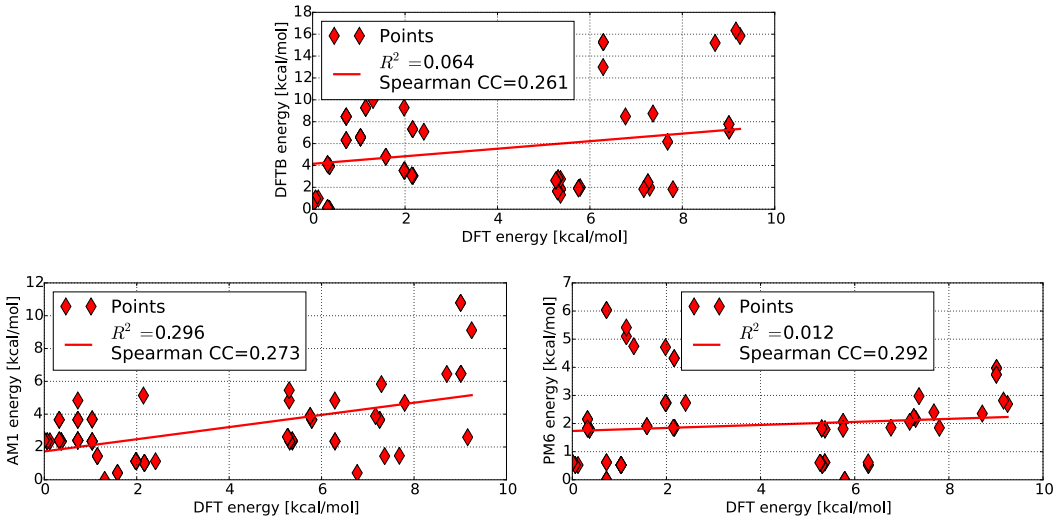

Fig. S12: Correlation of DFTB, AM1 and PM6 energy with DFT energy for conformational ensemble of Sildenafil+. Structures were optimized with the respective method. The energies are calculated with respect to the most stable conformer. The correlation coefficient R^2 and Spearman correlation coefficient (Spearman CC) are given.

S.7 pK_a prediction of substituted phenols

Gross et al. investigated correlation of partial charges of the alcohol group of substituted phenols and the amine group of substituted anilins with their pK_a value.[30] We investigated pK_a prediction for their test set of 19 substituted phenols. The structures are given in Fig. S14. Linear regressions (LR), multivariate linear regressions (MVLR), and regularized multivariate linear regression (RMVLR) with a regularization parameter of 0.0001 were performed with Scikit-learn.[29] The resulting models are gathered in Tab. S1. We plotted all 4 dimension of the O-centered descriptor calculated with CM5 charges in Fig. S13. The first dimension corresponds to the charge on the oxygen atom, the second dimension to the average charge of all atoms bound to the oxygen atom (including the H atom), and so on. The mean-squared error (MSE) is calculated according to $\frac{1}{n_{samples}}\sum_{i=1}^{n_{samples}}(y_i - \hat{y}_i)^2$, where y_i is the measured pK_a, $\hat{y}_i$ the predicted pK_a, and $n_{samples}$ the number of samples. To calculate the leave-one-out MSE (LOO-MSE) every sample is left out once and its pK_a predicted by a model trained on the $n-1$ remaining samples. The $\hat{y}_i$-values obtained with this procedure are used to calculate the LOO-MSE according to the formula above.

Tab. S1: Model results for pK_a prediction for different descriptors and partial charge schemes. In the first column the dimension of the descriptor, and if it is centered on the alcohol H atom or alcohol O atom, is indicated. A dimension of one (1D) corresponds to the atom partial charge only. R^2 is the coefficient of determination, MSE the mean-squared error and LOO-MSE the MSE determined with leave-one-out cross validation.

Descriptor	Charge scheme	R^2	MSE	LOO-MSE
1D, H-centered	NPA	0.84	0.11	0.15
2D	NPA	0.94	0.04	0.07
3D	NPA	0.96	0.03	0.06
4D	NPA	0.96	0.02	0.06
1D, H-centered	CM5	0.87	0.09	0.12
2D	CM5	0.91	0.06	0.10
3D	CM5	0.97	0.02	0.04
4D	CM5	0.98	0.01	0.03
1D, H-centered	Mulliken	0.83	0.12	0.17
2D	Mulliken	0.93	0.05	0.07
3D	Mulliken	0.95	0.03	0.07
4D	Mulliken	0.95	0.03	0.08
1D, H-centered	ESP	0.09	0.61	0.78
2D	ESP	0.39	0.41	0.63
3D	ESP	0.63	0.25	0.57
4D	ESP	0.90	0.07	0.16
1D, O-centered	NPA	0.94	0.04	0.07
2D	NPA	0.94	0.04	0.07
3D	NPA	0.96	0.03	0.06
4D	NPA	0.96	0.03	0.06
1D, O-centered	CM5	0.91	0.06	0.10
2D	CM5	0.96	0.03	0.05
3D	CM5	0.98	0.02	0.03
4D	CM5	0.98	0.02	0.03
1D, O-centered	Mulliken	0.93	0.05	0.07
2D	Mulliken	0.95	0.04	0.06
3D	Mulliken	0.95	0.03	0.07
4D	Mulliken	0.95	0.03	0.07
1D, O-centered	ESP	0.23	0.52	0.64
2D	ESP	0.29	0.47	0.77
3D	ESP	0.39	0.41	0.71
4D	ESP	0.60	0.27	0.67
AC (RMVLR)	CM5	0.92	0.05	0.09

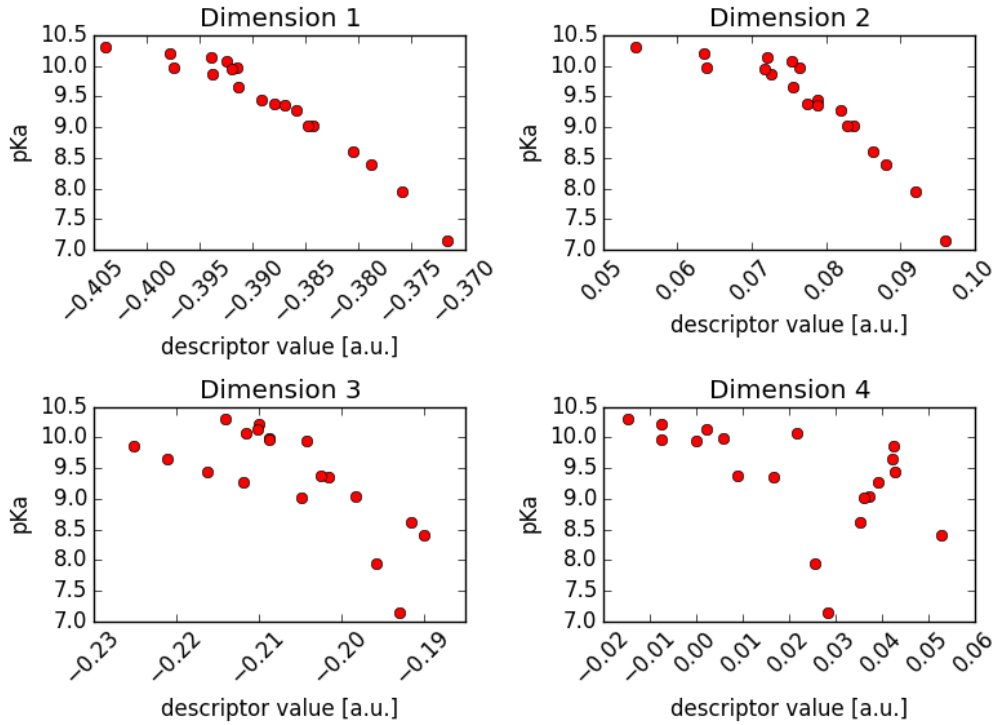

Fig. S13: Plot of all four dimensions of the O-centered radial descriptor calculated with CM5 charges.

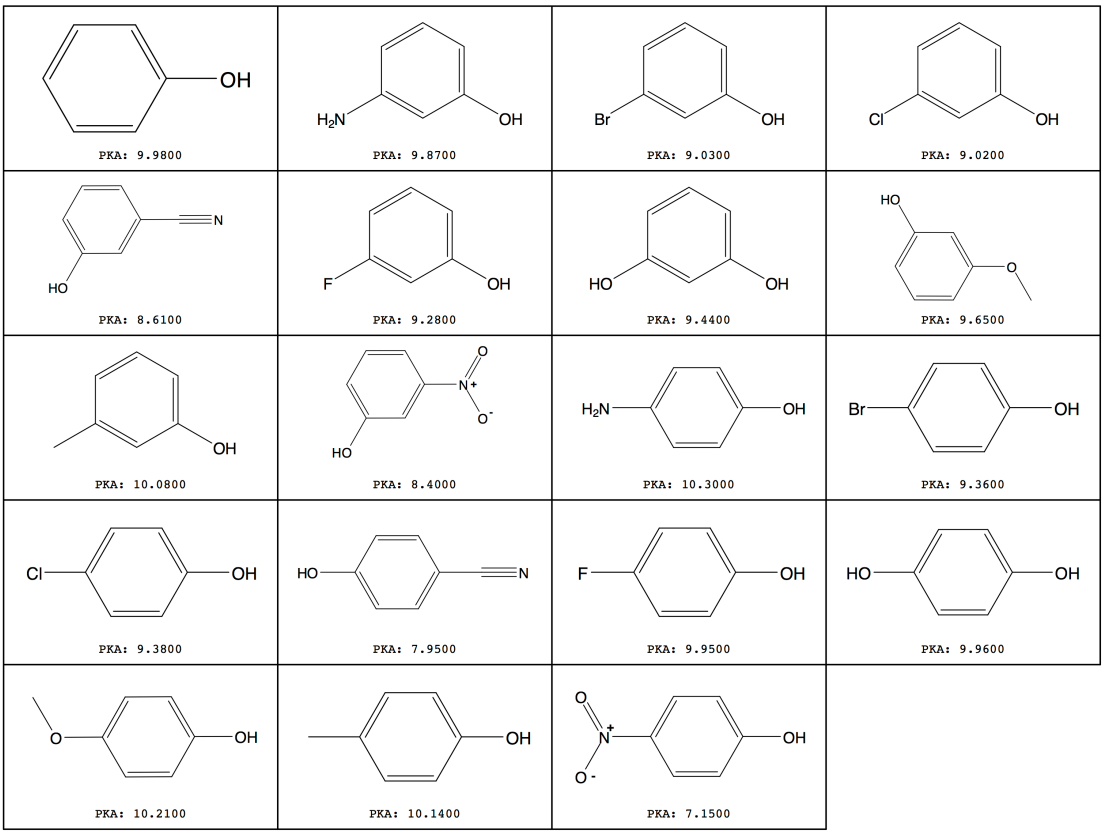

Fig. S14 Set of 19 substituted phenols for pK_a prediction.

S.8 Boxplots of charge distributions on heavy atoms of Sildenafil, Atorvastatin and Sildenafil+

S.8.1 Conformational ensemble of Sildenafil

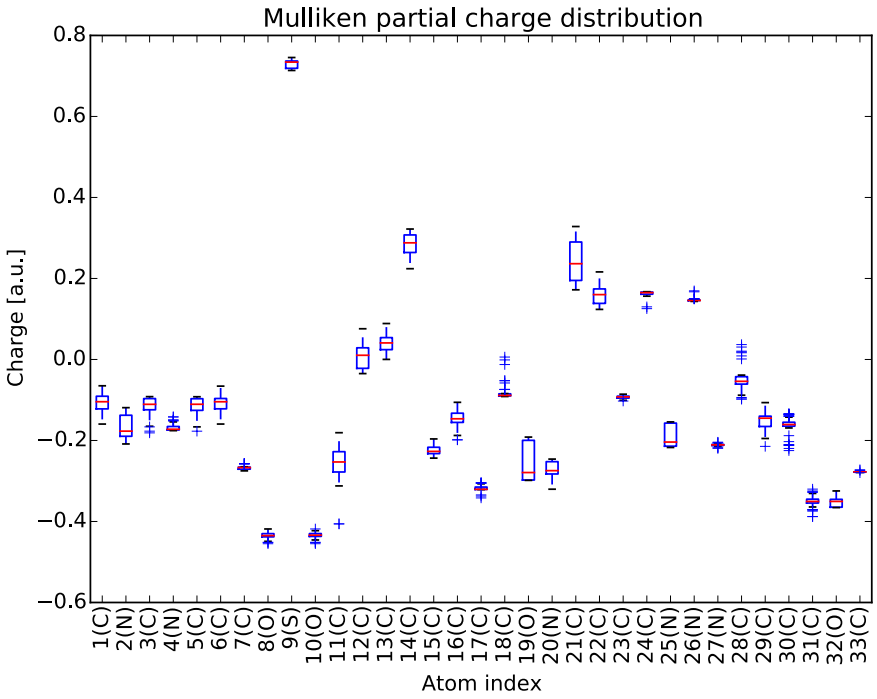

Fig. S15: Boxplot of Mulliken charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

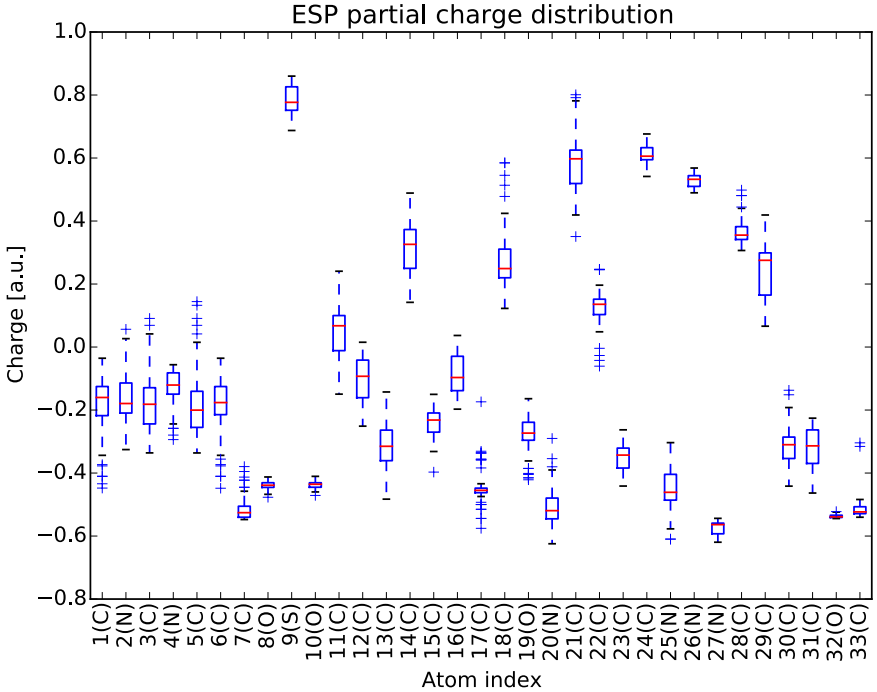

Fig. S16: Boxplot of ESP charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

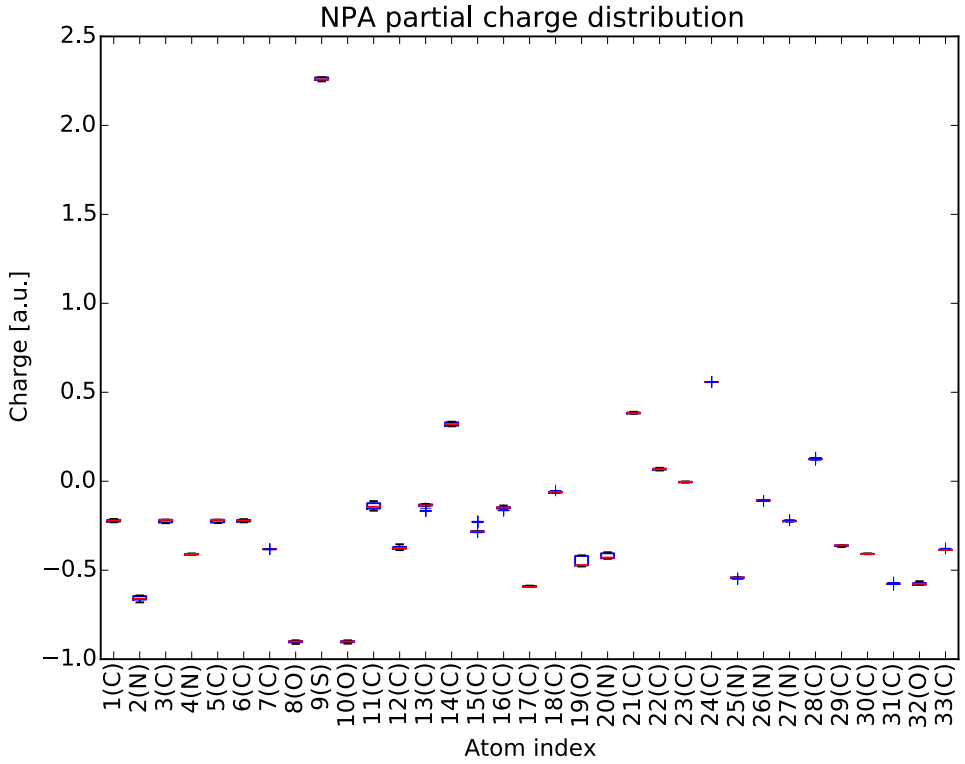

Fig. S17: Boxplot of NPA charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

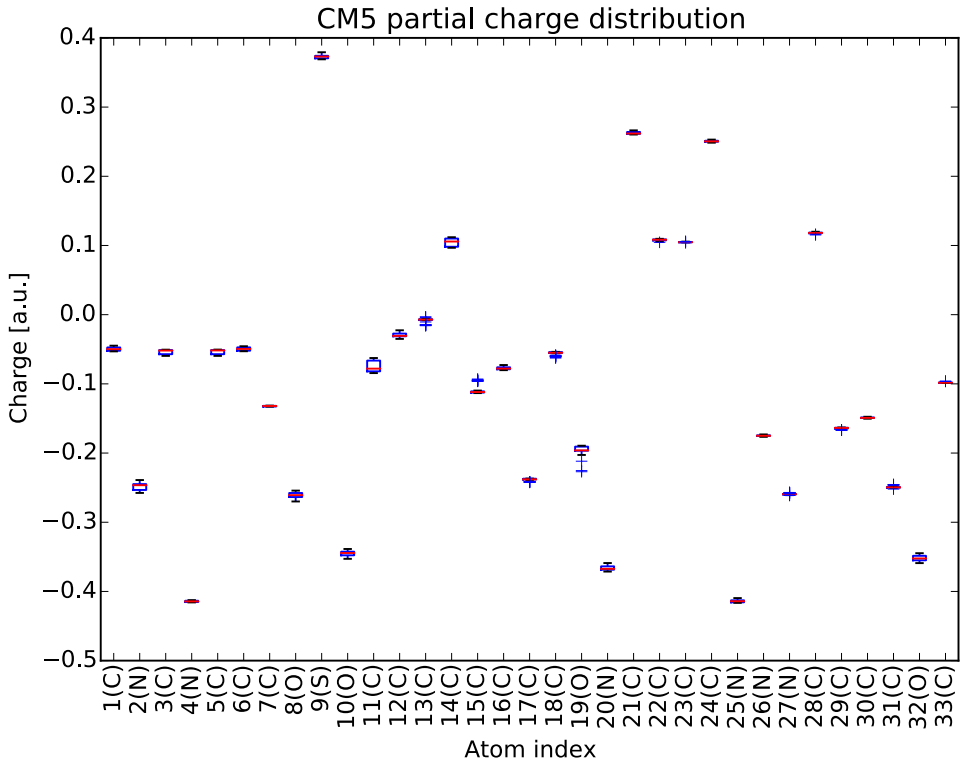

Fig. S18: Boxplot of CM5 charge distribution on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

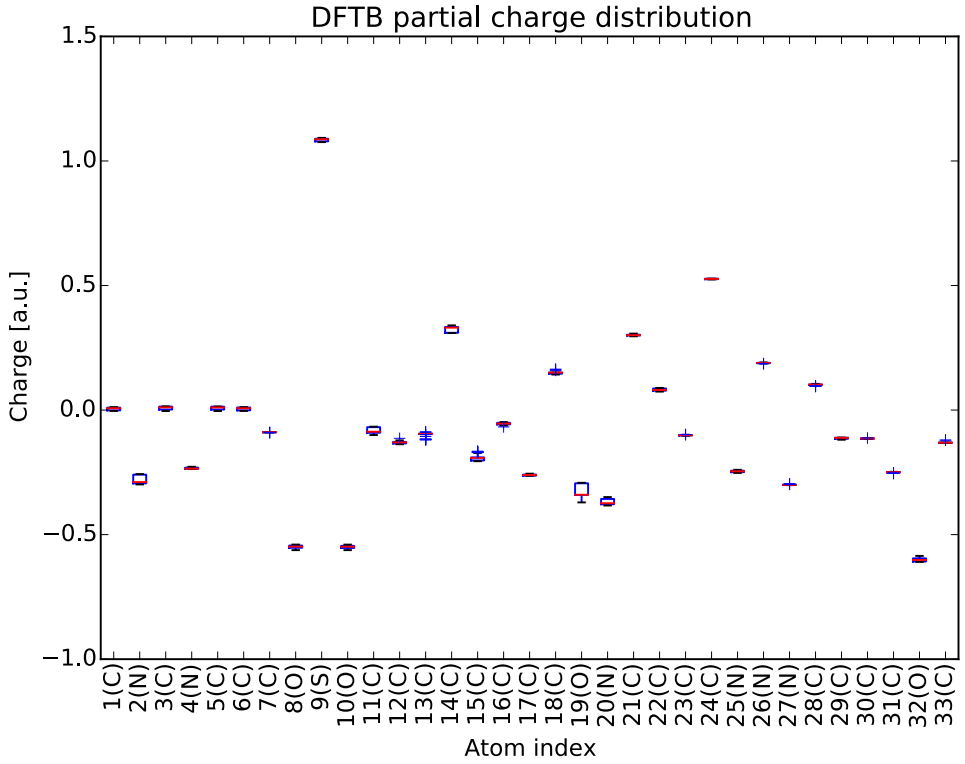

Fig. S19: Boxplot of DFTB charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

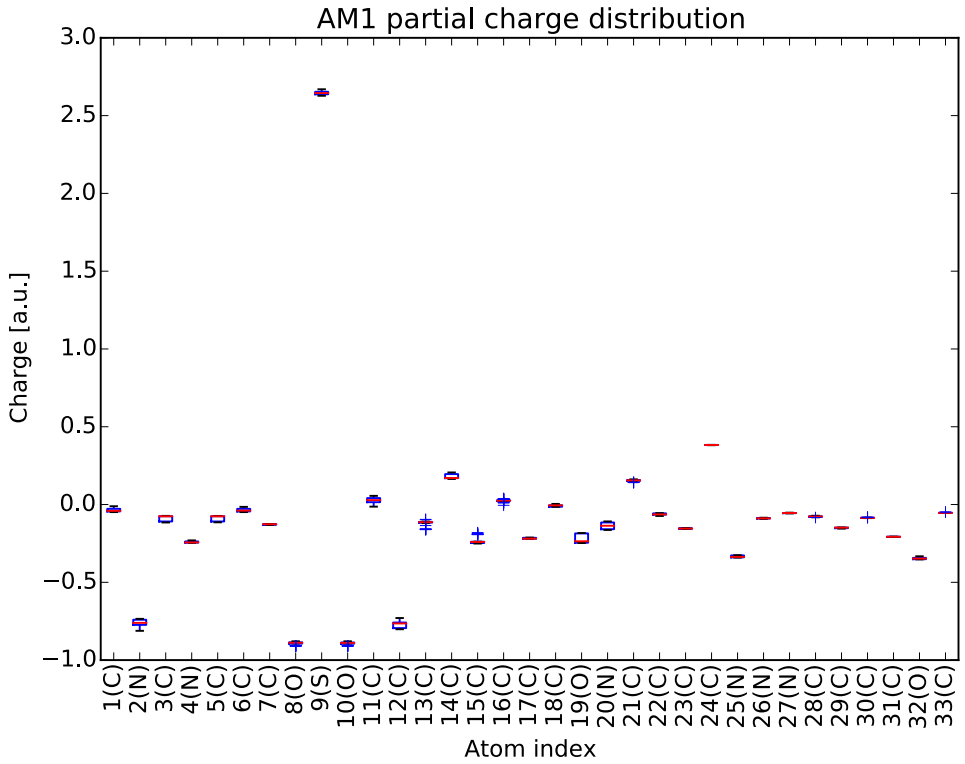

Fig. S20: Boxplot of AM1 charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

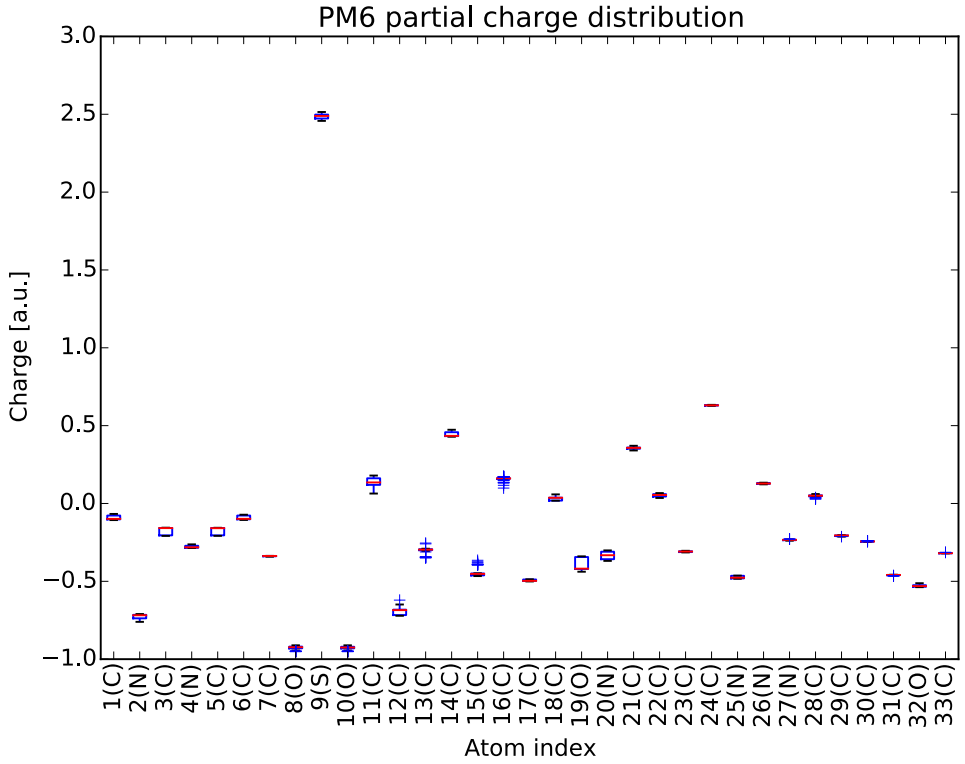

Fig. S21: Boxplot of PM6 charge distributions on heavy atoms of the Sildenafil conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

S.8.2 Conformational ensemble of Atorvastatin

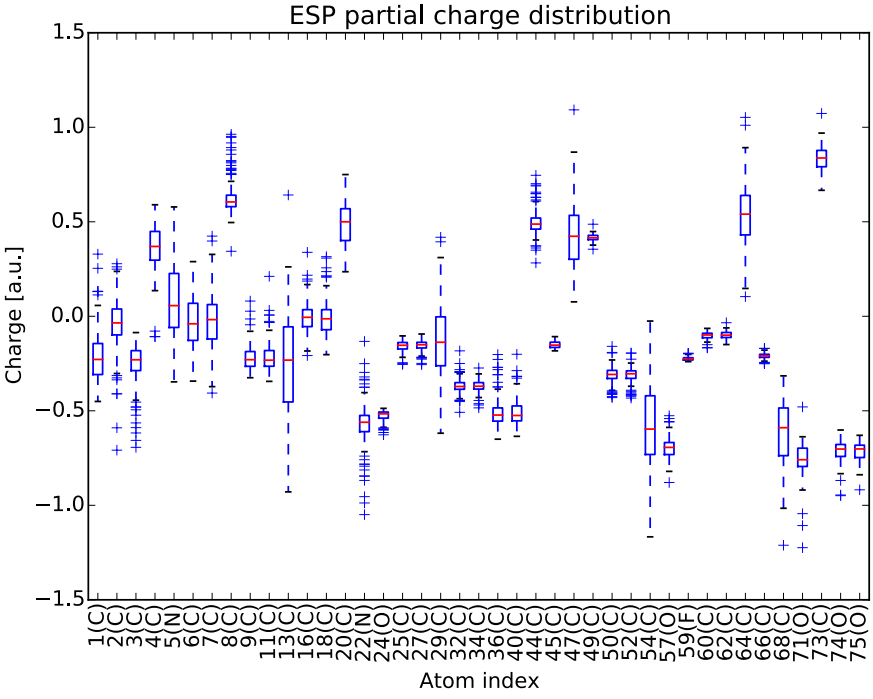

Fig. S22: Boxplot of ESP charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

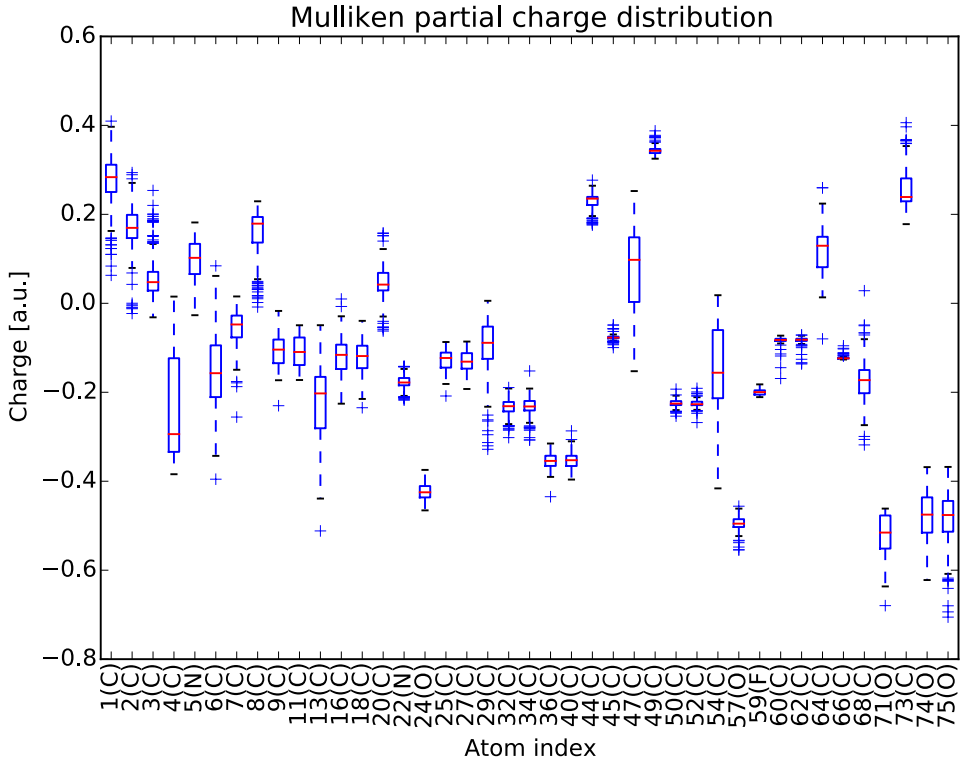

Fig. S23: Boxplot of Mulliken charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

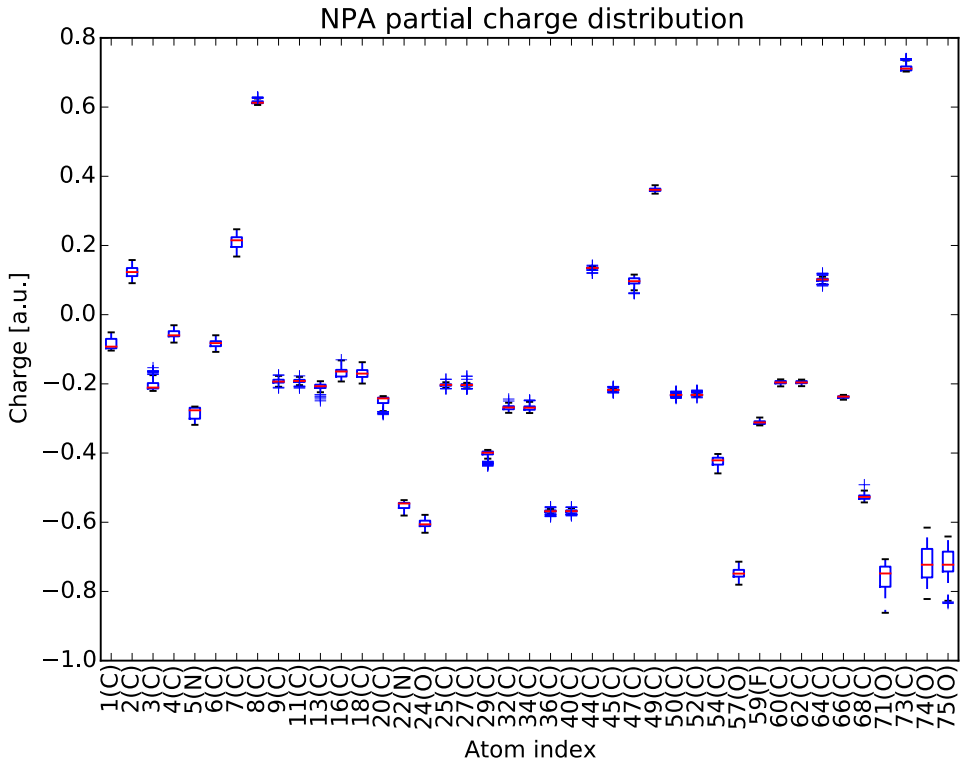

Fig. S24: Boxplot of NPA charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

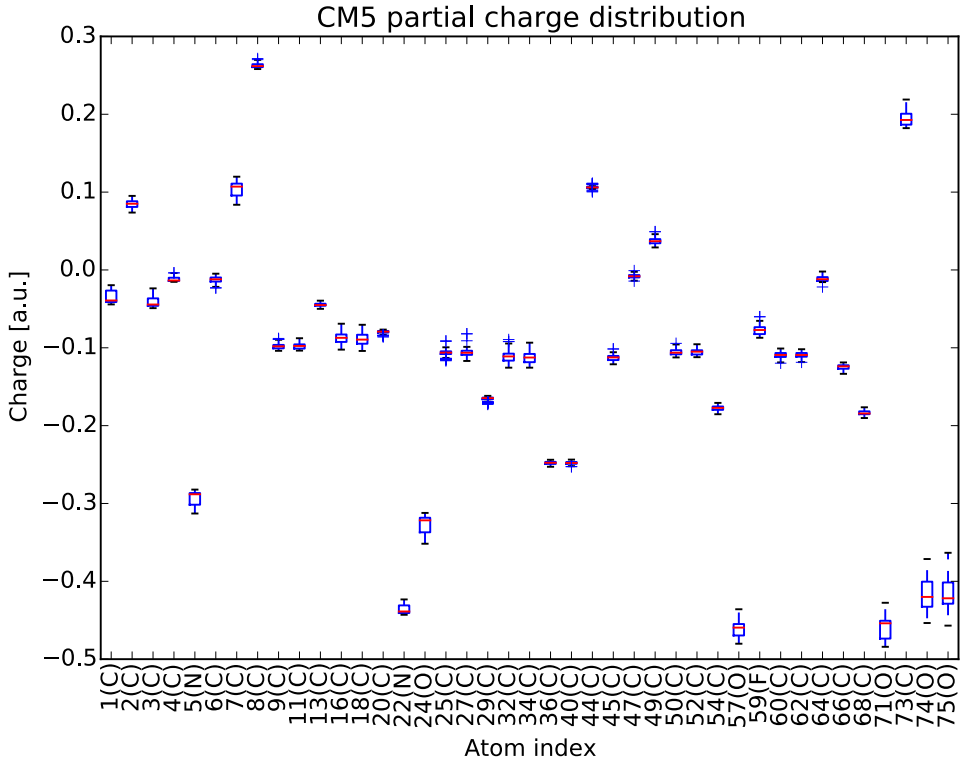

Fig. S25: Boxplot of CM5 charge distribution on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

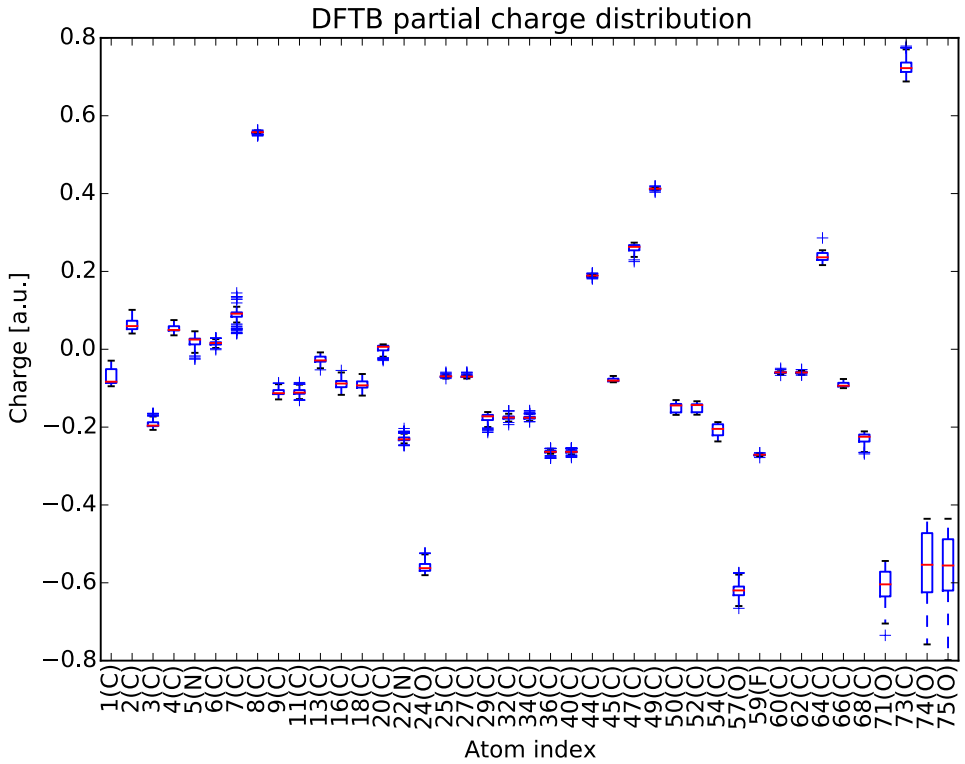

Fig. S26: Boxplot of DFTB charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

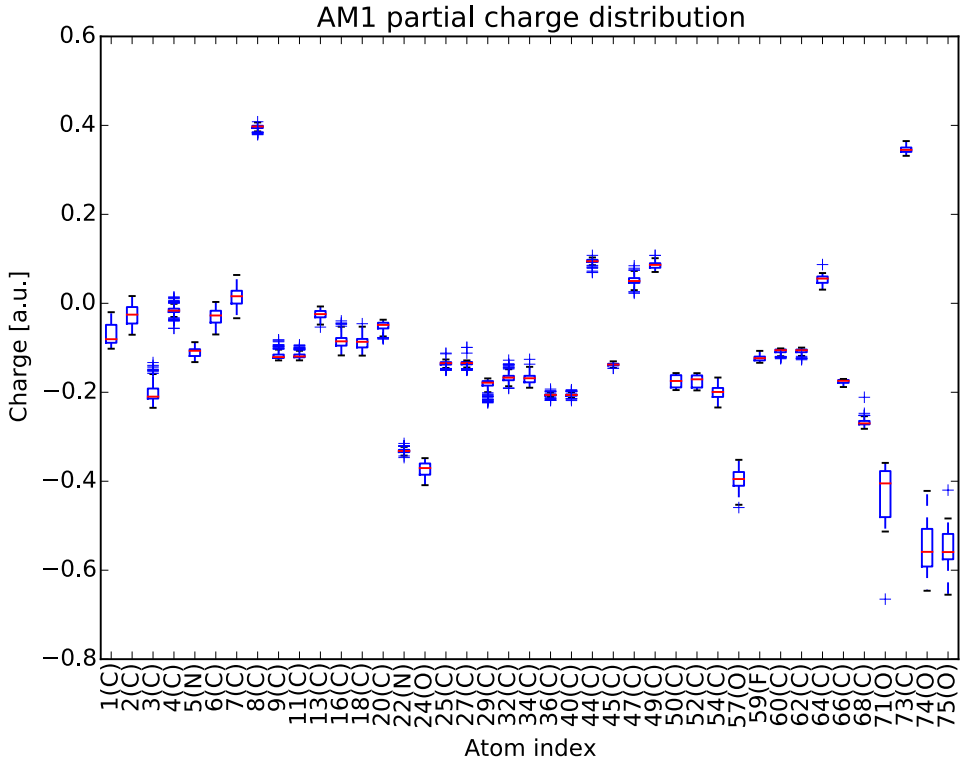

Fig. S27: Boxplot of AM1 charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

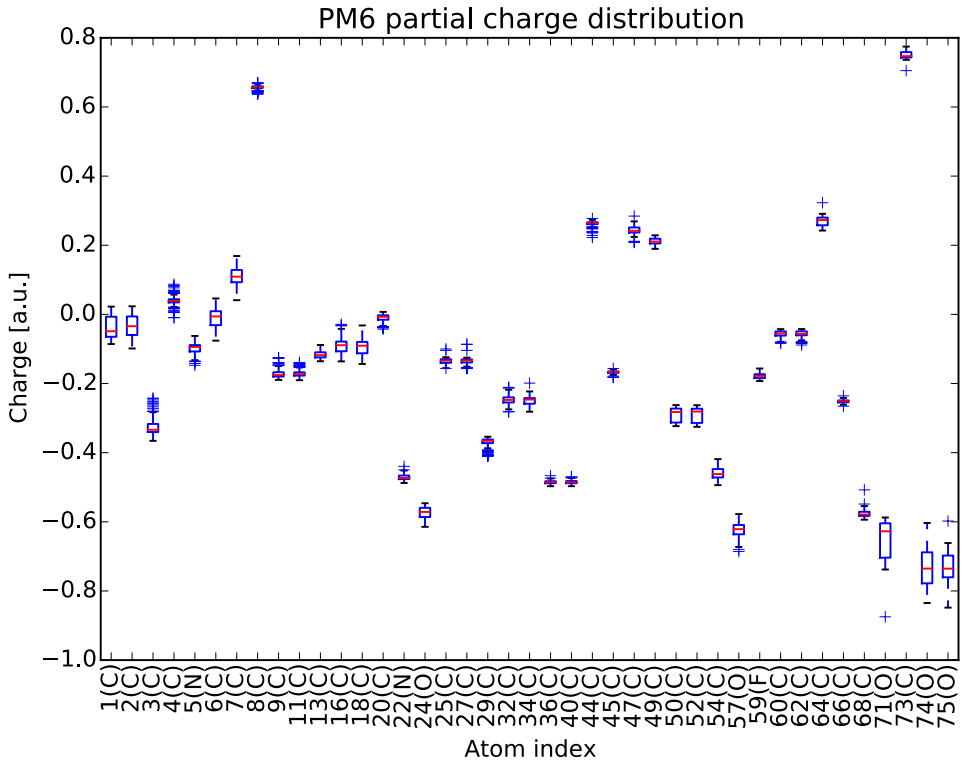

Fig. S28: Boxplot of PM6 charge distributions on heavy atoms of the Atorvastatin conformational ensemble. The atom numbering is given in Fig. 3 A) of the article.

S.8.3 Conformational ensemble of Sildenafil+

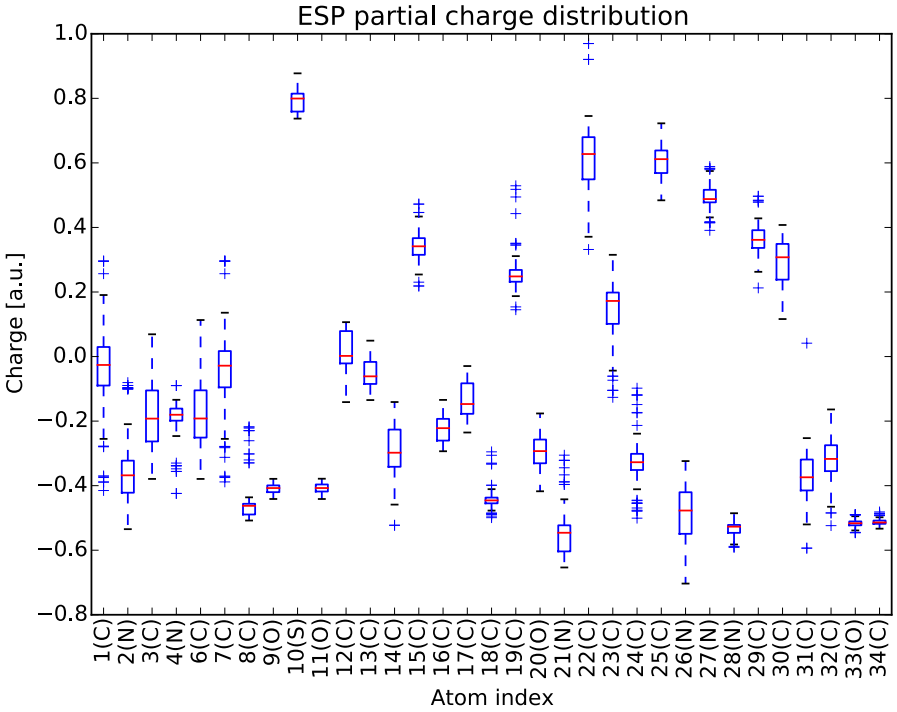

Fig. S29: Boxplot of ESP charge distributions on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

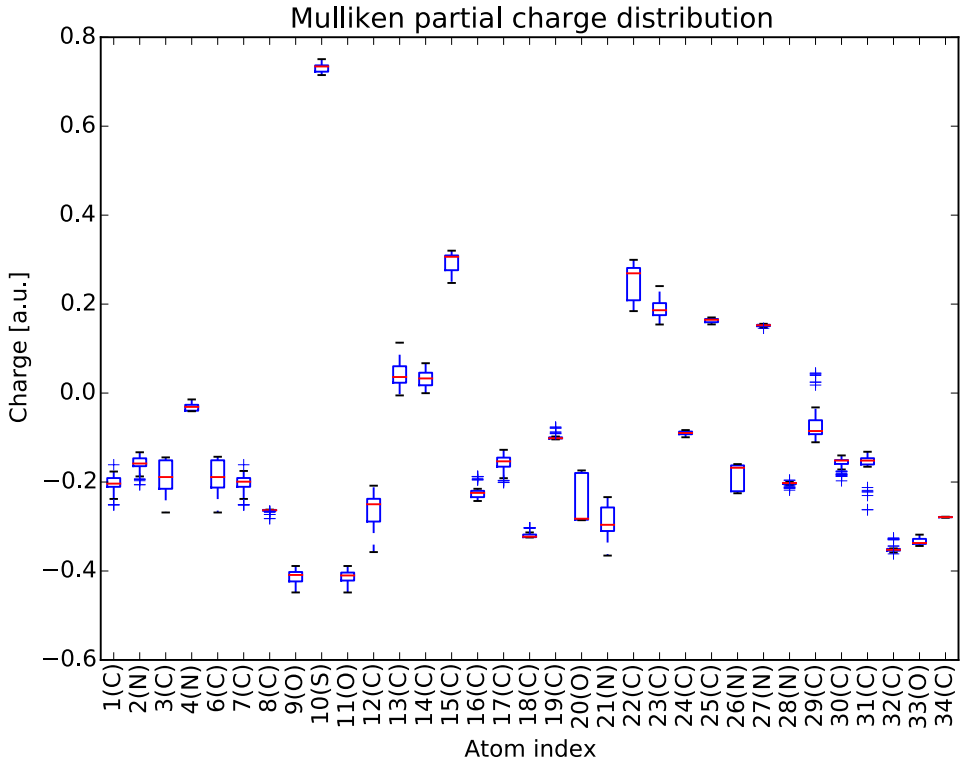

Fig. S30: Boxplot of Mulliken charge distributions on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

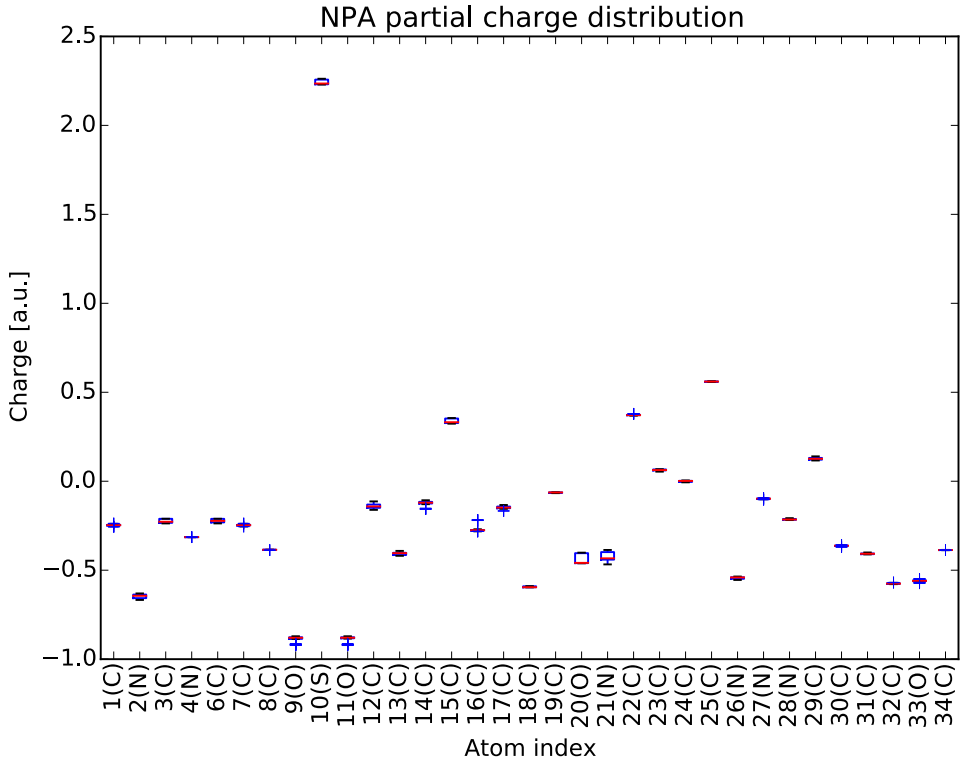

Fig. S31: Boxplot of NPA charge distributions on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

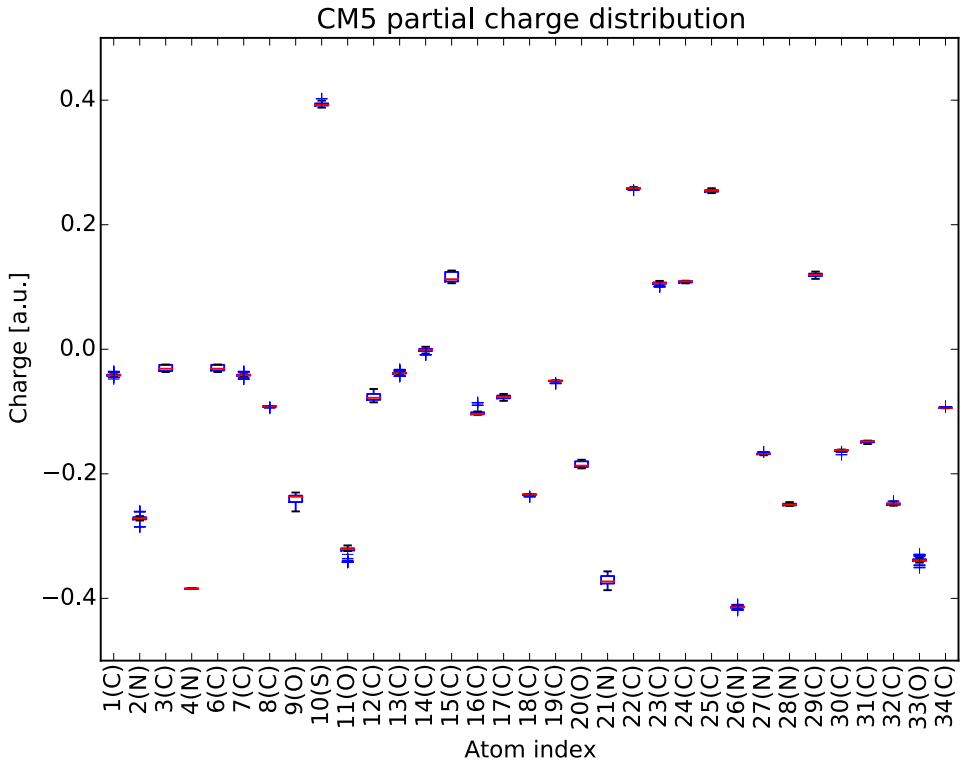

Fig. S32: Boxplot of CM5 charge distribution on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

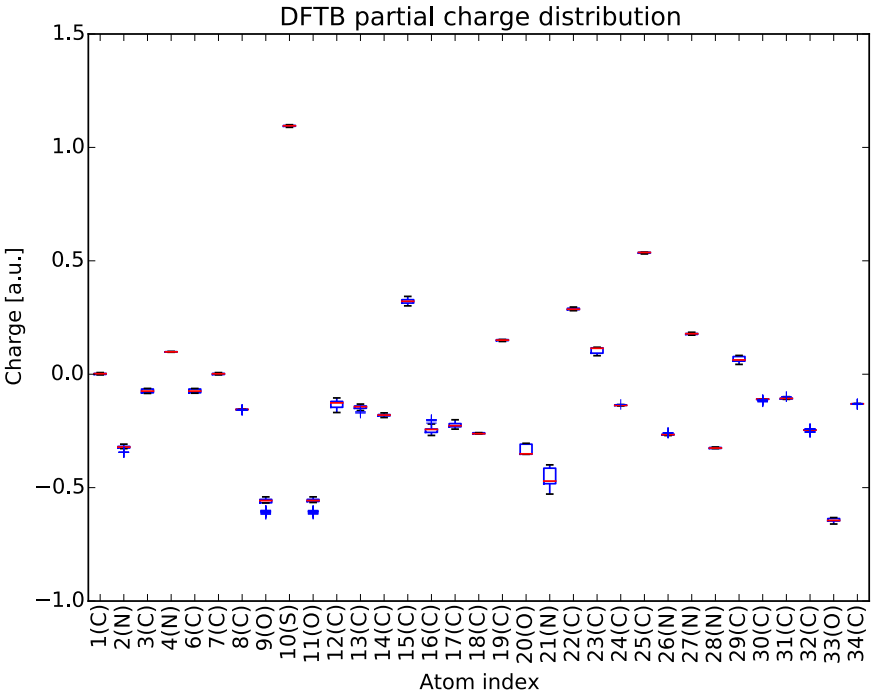

Fig. S33: Boxplot of DFTB charge distribution on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

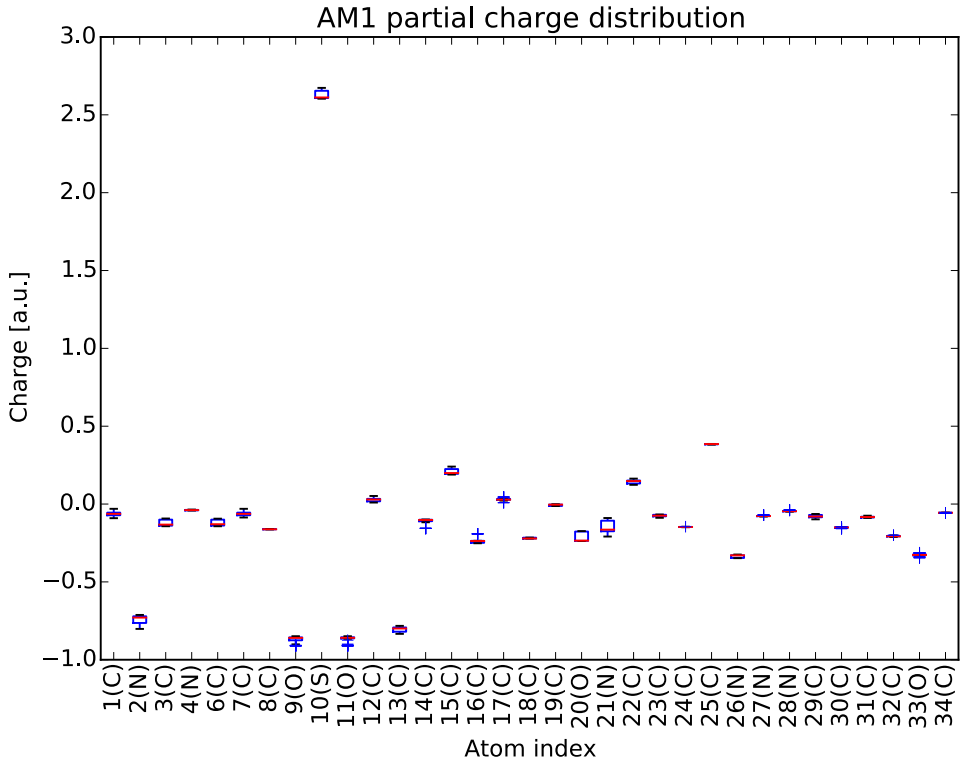

Fig. S34: Boxplot of AM1 charge distribution on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

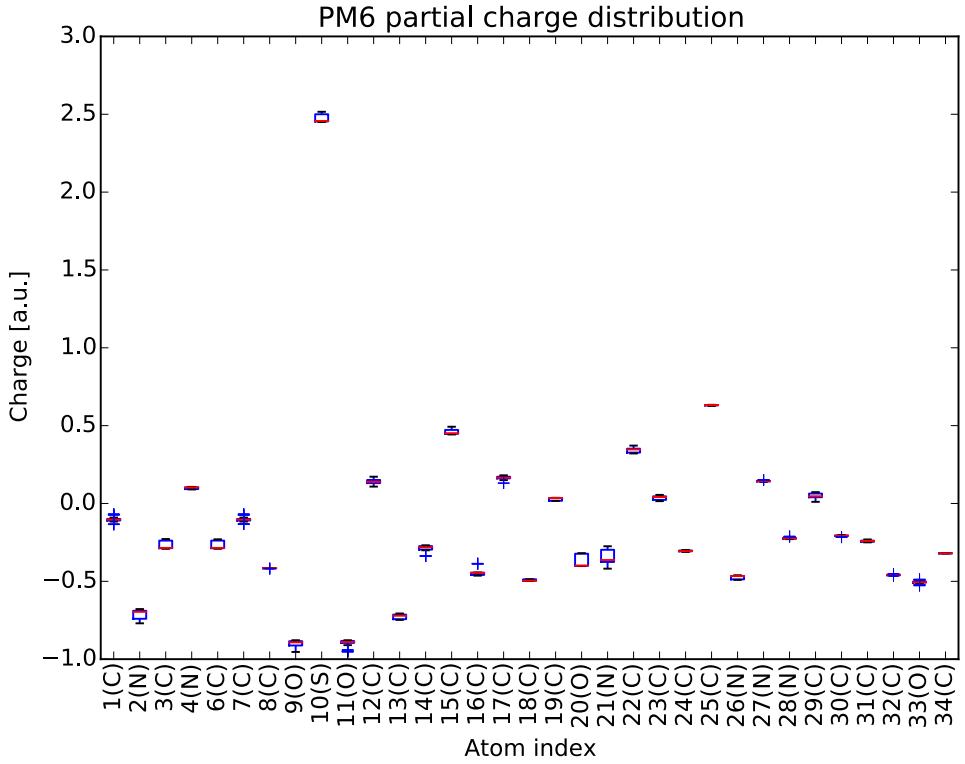

Fig. S35: Boxplot of PM6 charge distribution on heavy atoms of the Sildenafil+ conformational ensemble. The atom numbering is given in Fig. S1.

S.9 Scatter plots of partial charges on heavy atoms of conformational ensembles of Sildenafil, Atorvastatin and Sildenafil+ colored according to conformer energy

S.9.1 Sildenafil conformational ensemble

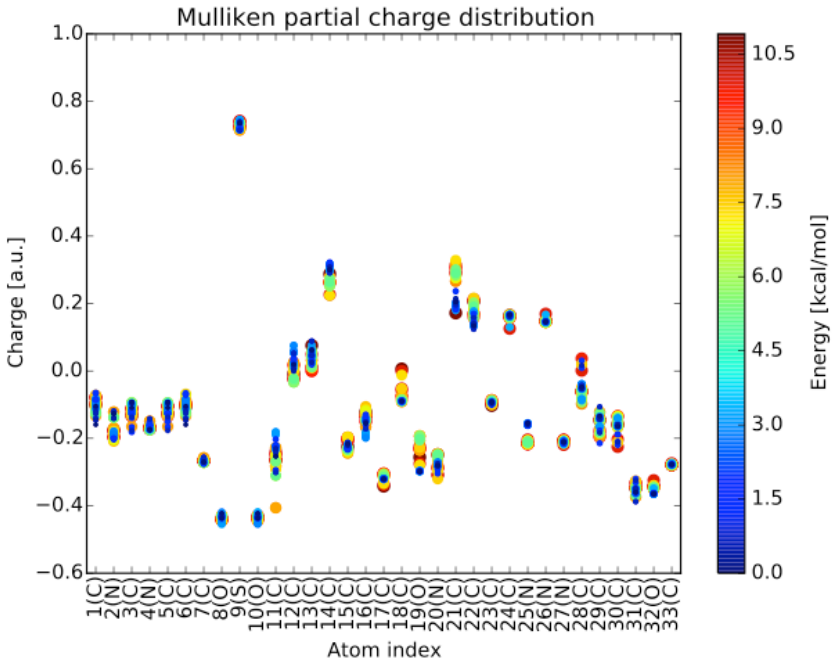

Fig. S36: Mulliken charge distributions on heavy atoms of the Sildenafil conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

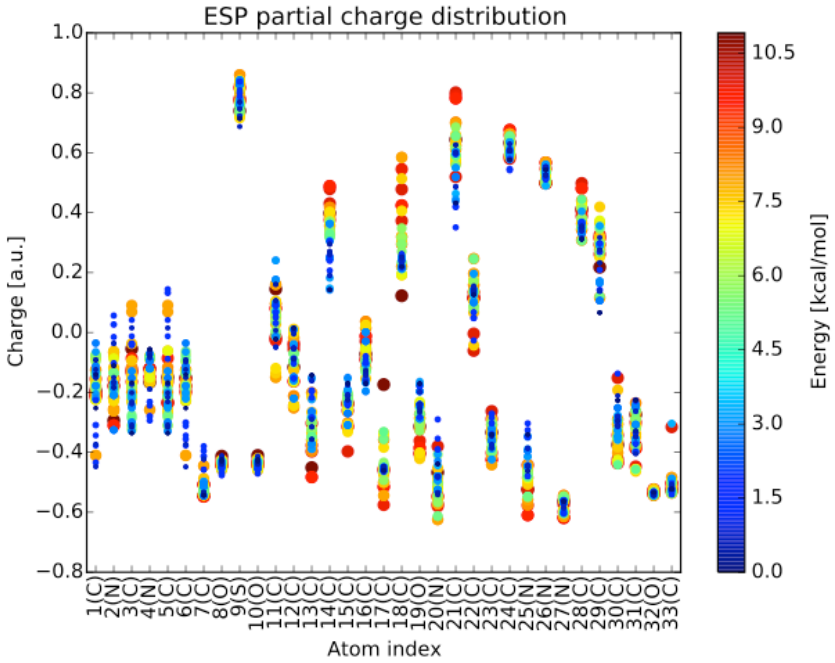

Fig. S37: ESP charge distributions on heavy atoms of the Sildenafil conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

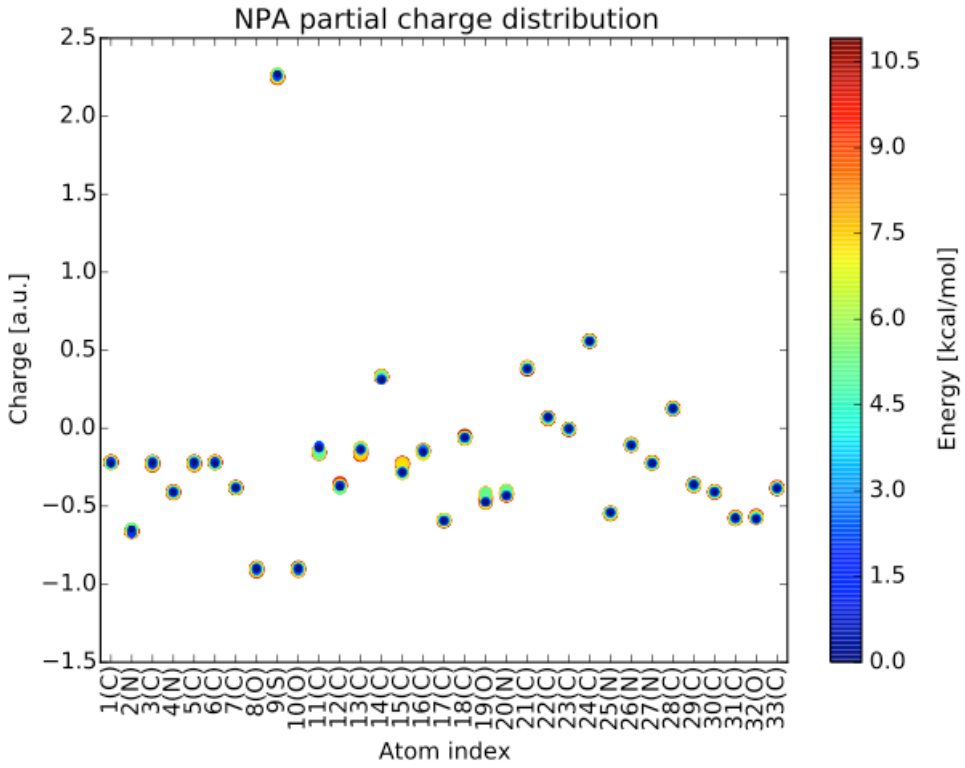

Fig. S38: NPA charge distributions on heavy atoms of the Sildenafil conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

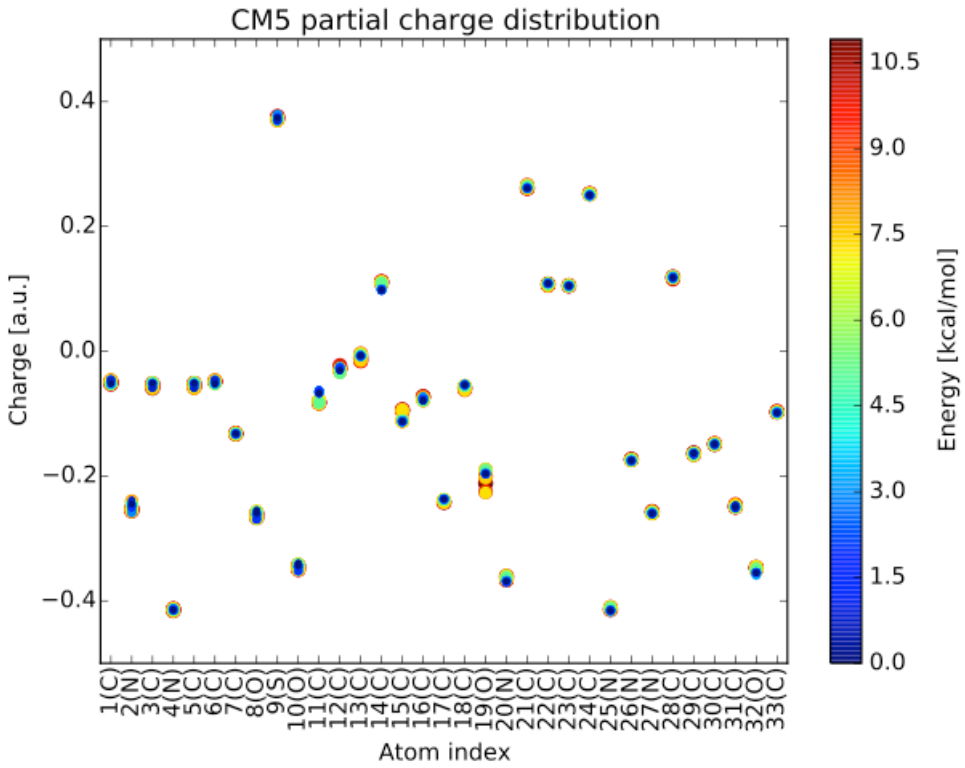

Fig. S39: CM5 charge distributions on atoms of the Sildenafil conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

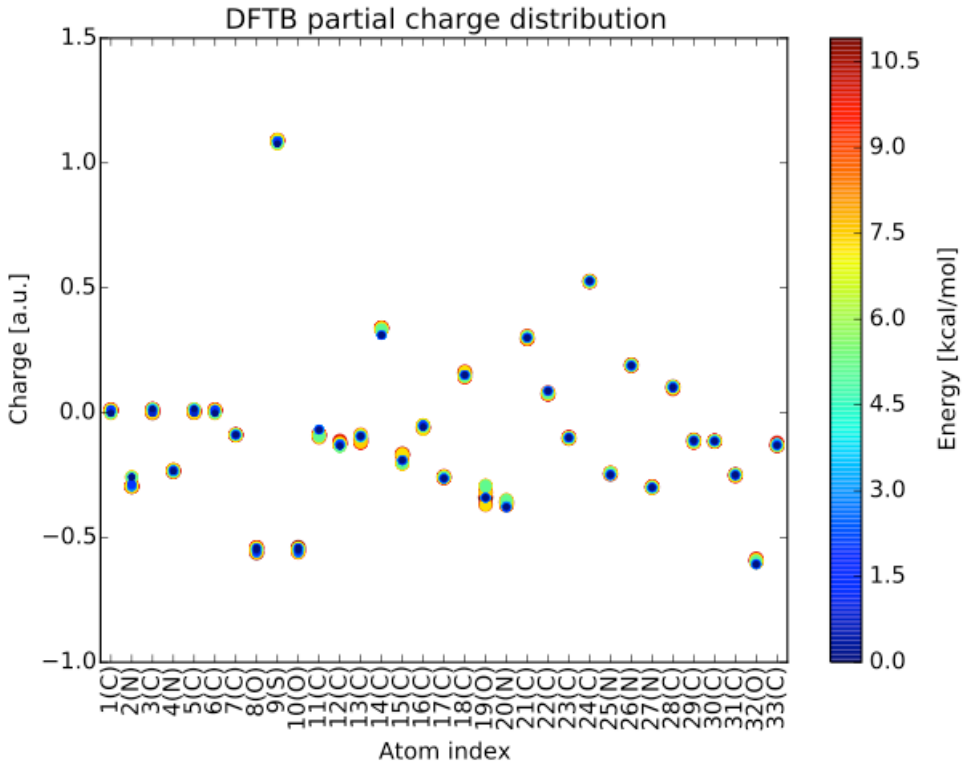

Fig. S40: DFTB charge distributions on heavy atoms of the Sildenafil conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

S.9.2 Atorvastatin conformational ensemble

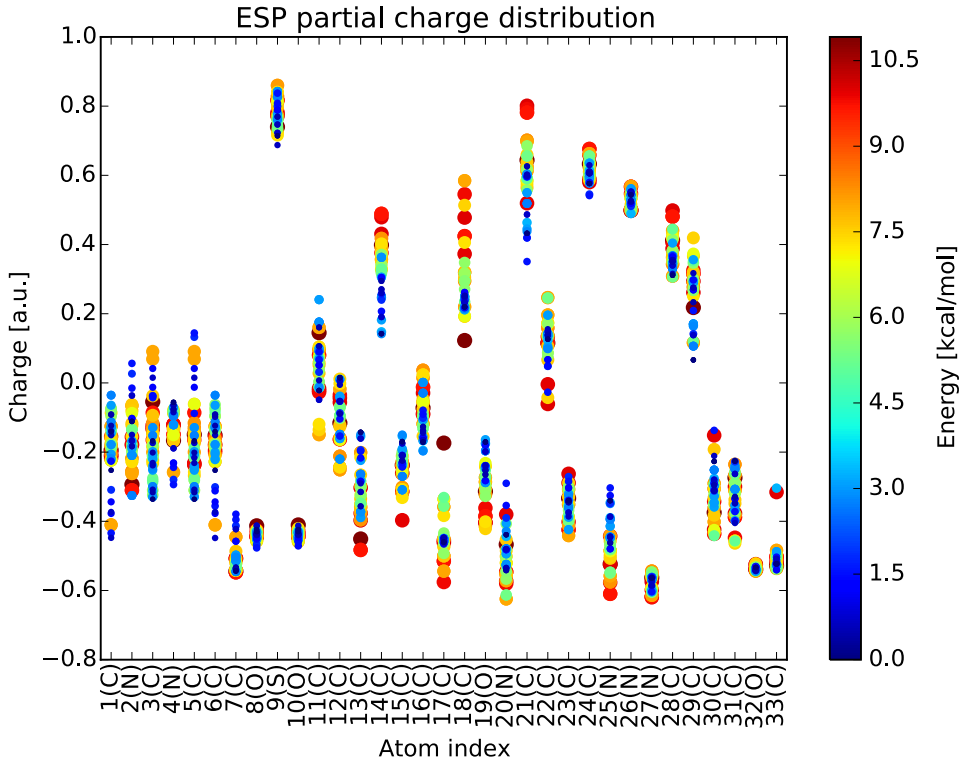

Fig. S41: ESP charge distributions on heavy atoms of the Atorvastatin conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

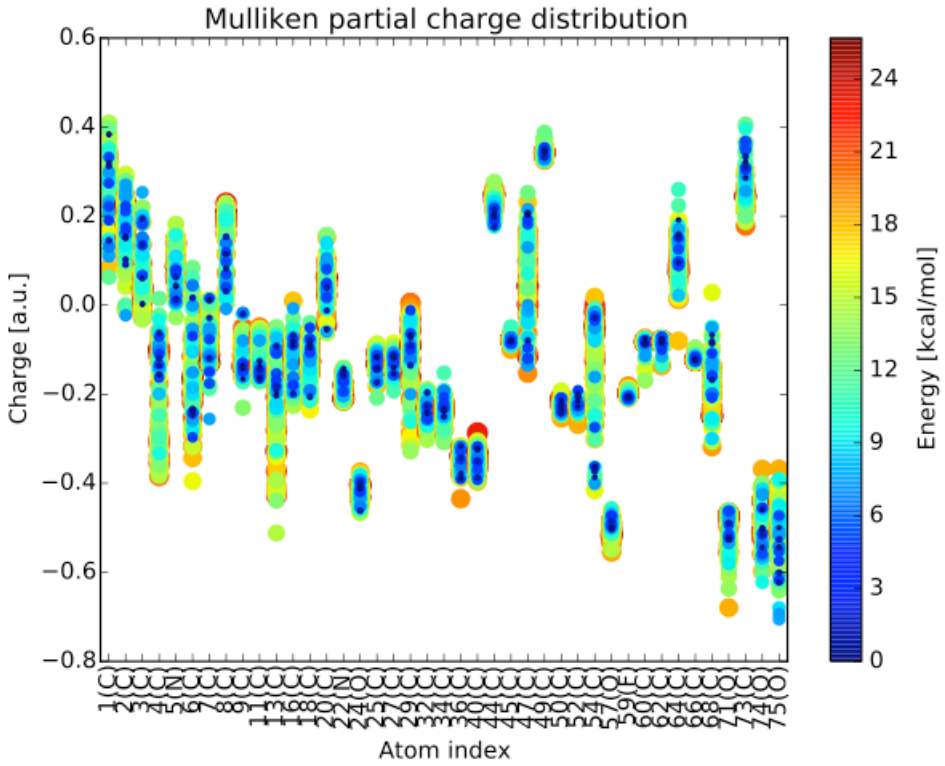

Fig. S42: Mulliken charge distributions on heavy atoms of the Atorvastatin conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

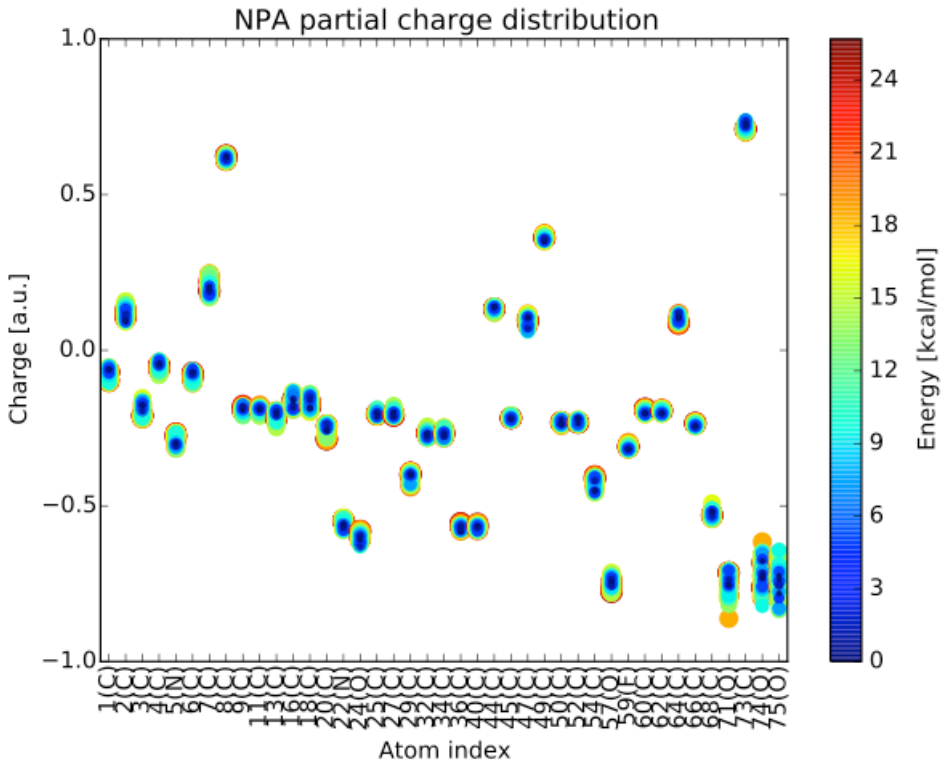

Fig. S43: NPA charge distributions on heavy atoms of the Atorvastatin conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

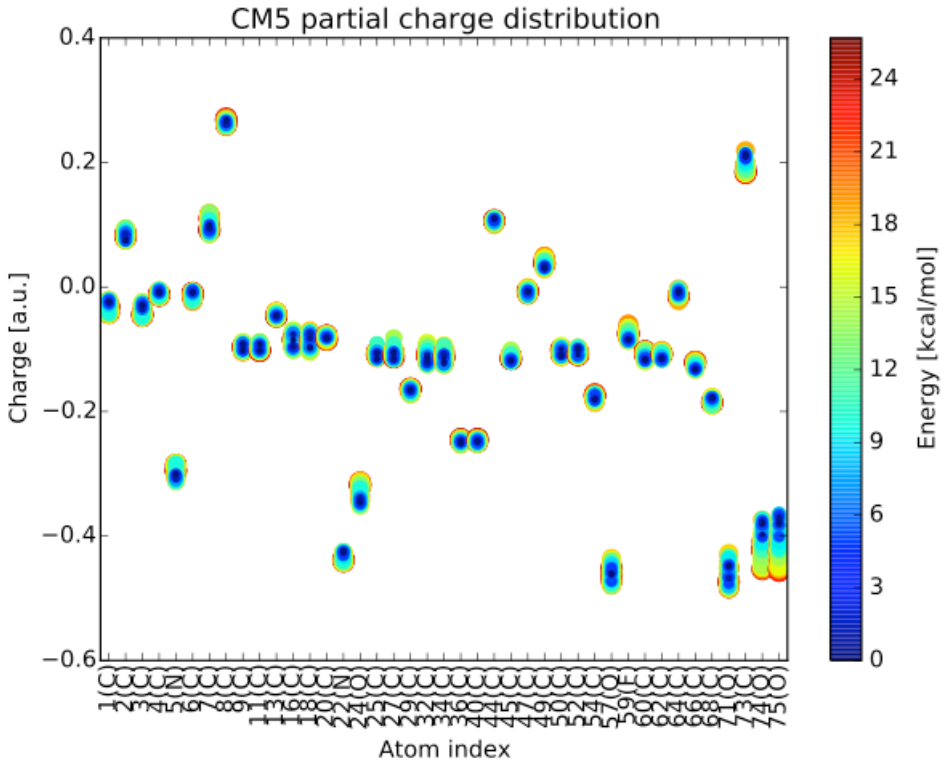

Fig. S44: CM5 charge distributions on heavy atoms of the Atorvastatin conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

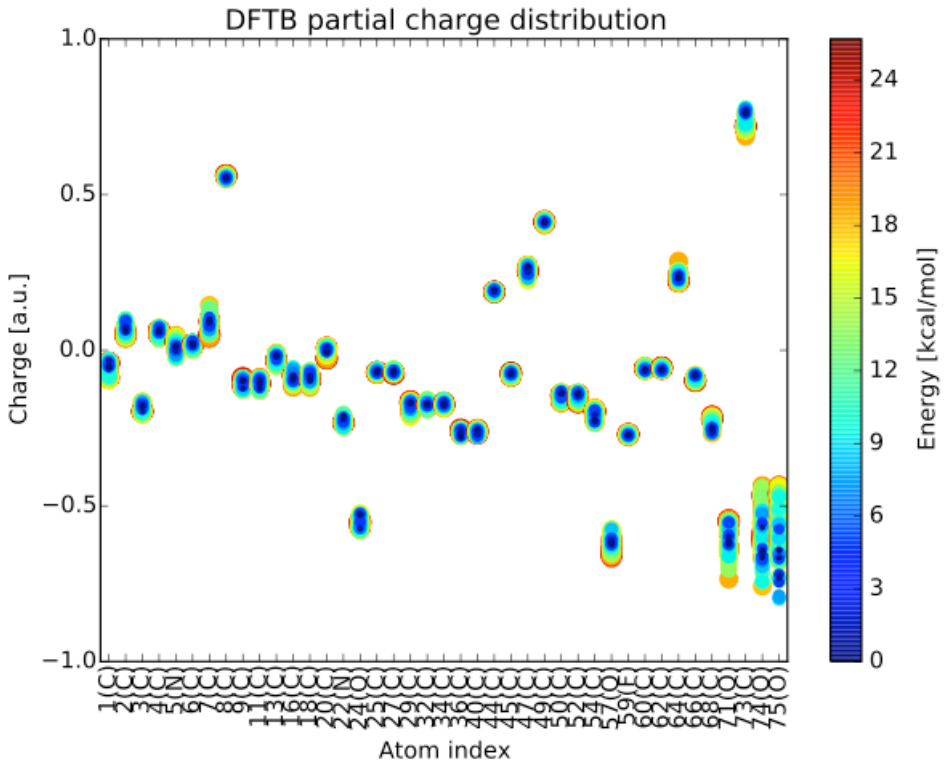

Fig. S45: DFTB charge distributions on heavy atoms of the Atorvastatin conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. 3 A).

S.9.3 Sildenafil+ conformational ensemble

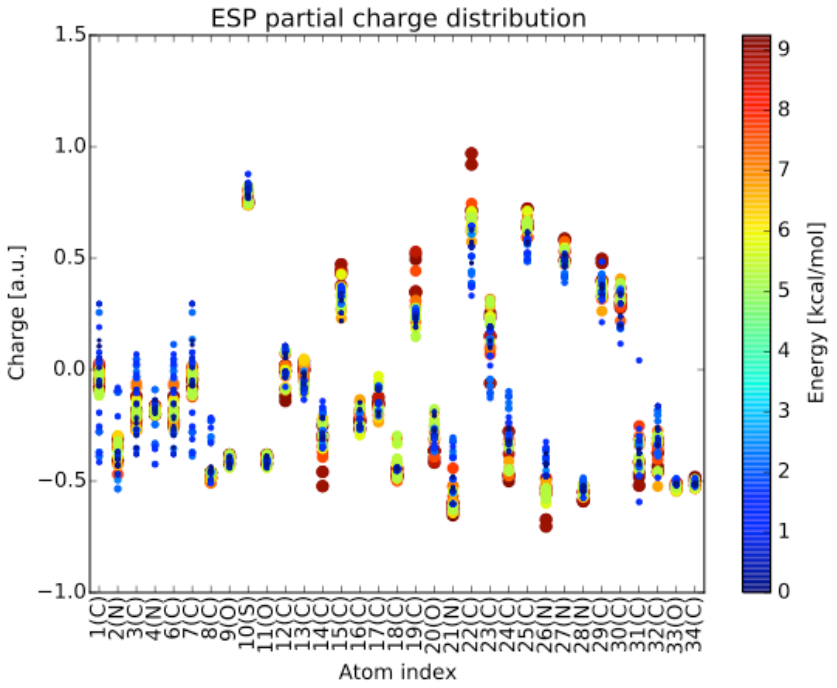

Fig. S46: ESP charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. S1.

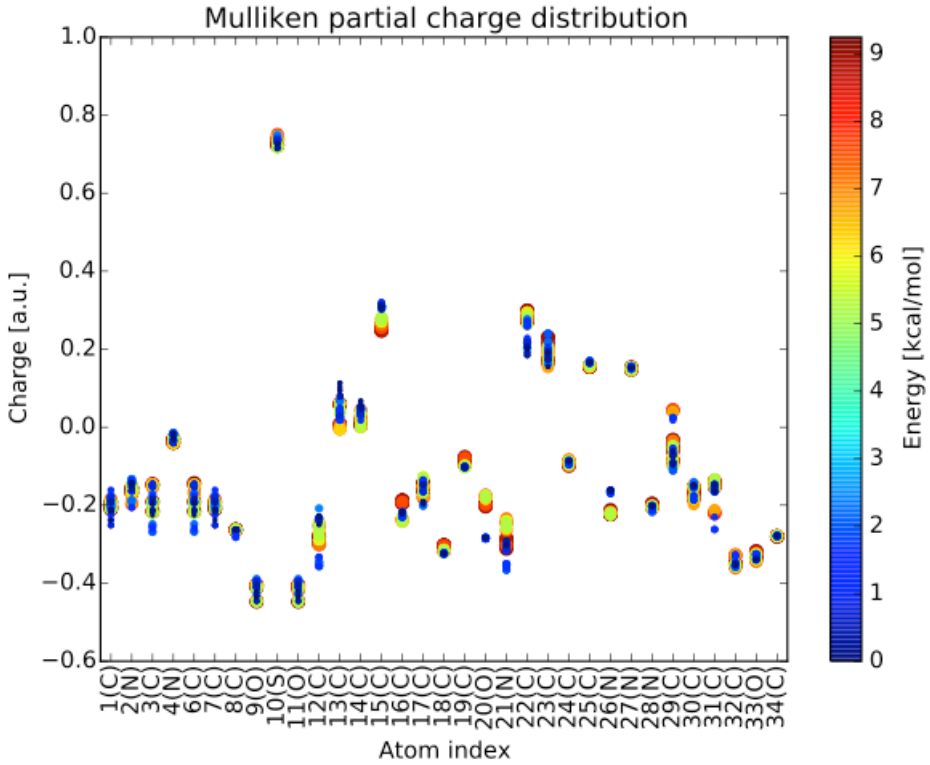

Fig. S47: Mulliken charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. S1.

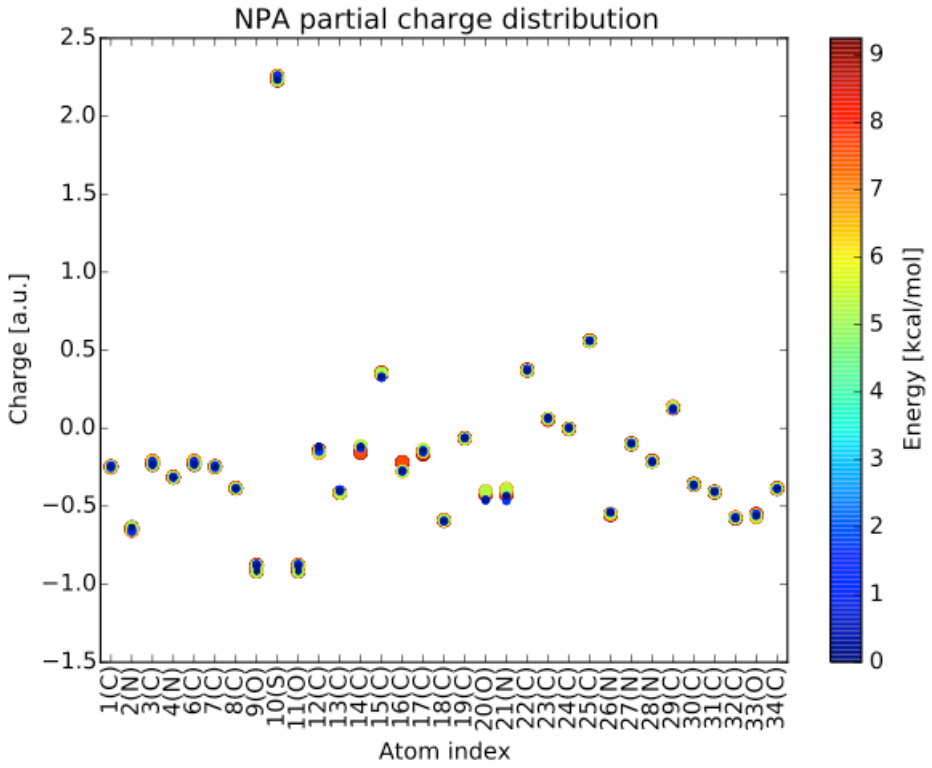

Fig. S48: NPA charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. S1.

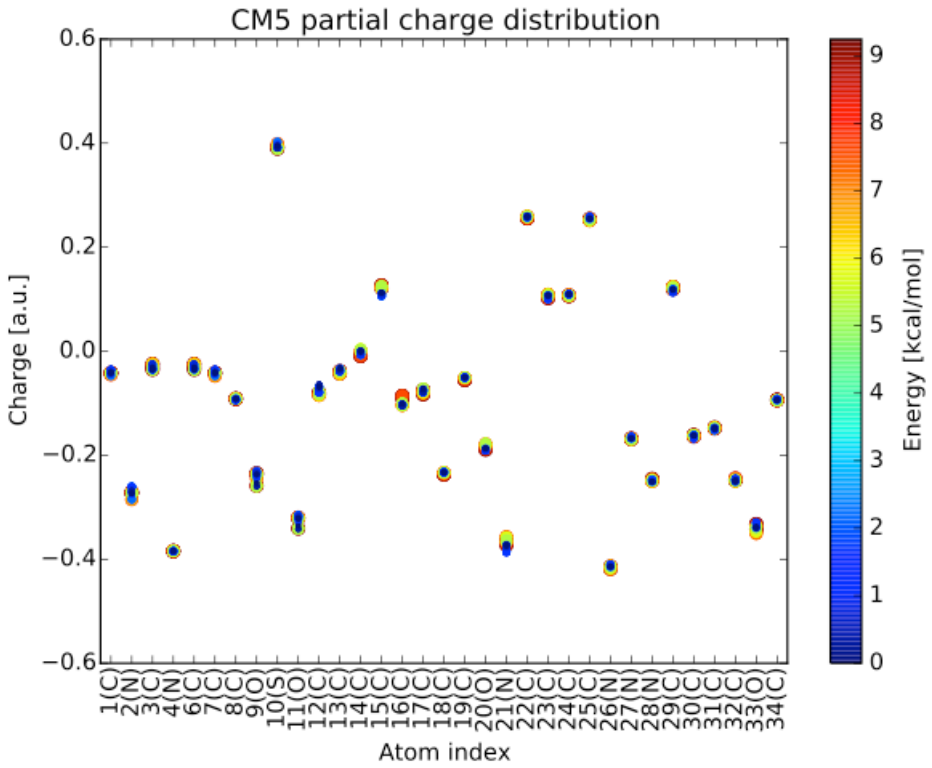

Fig. S49: CM5 charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. S1.

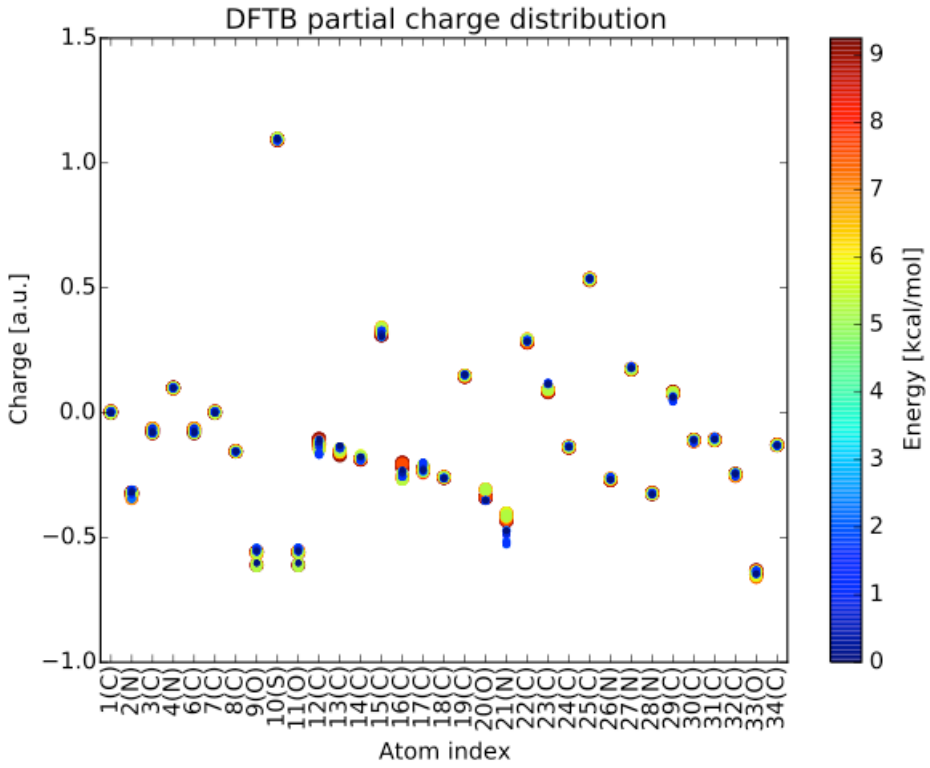

Fig. S50: DFTB charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. The atom numbering is given in Fig. S1.

S.10 Partial charge distributions of Sildenafil obtained with semi-empirical methods when structures are optimized with the respective method

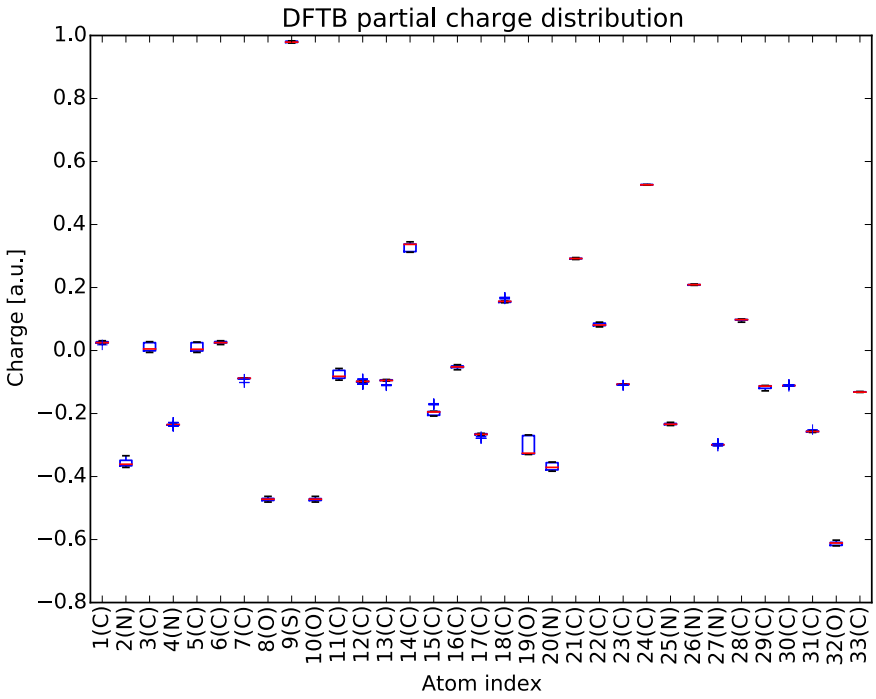

Fig. S51: Boxplot of DFTB charge distribution on heavy atoms of the Sildenafil conformational ensemble. Structures were optimized with the DFTB method. The atom numbering is given in Fig. 3 A) of the article.

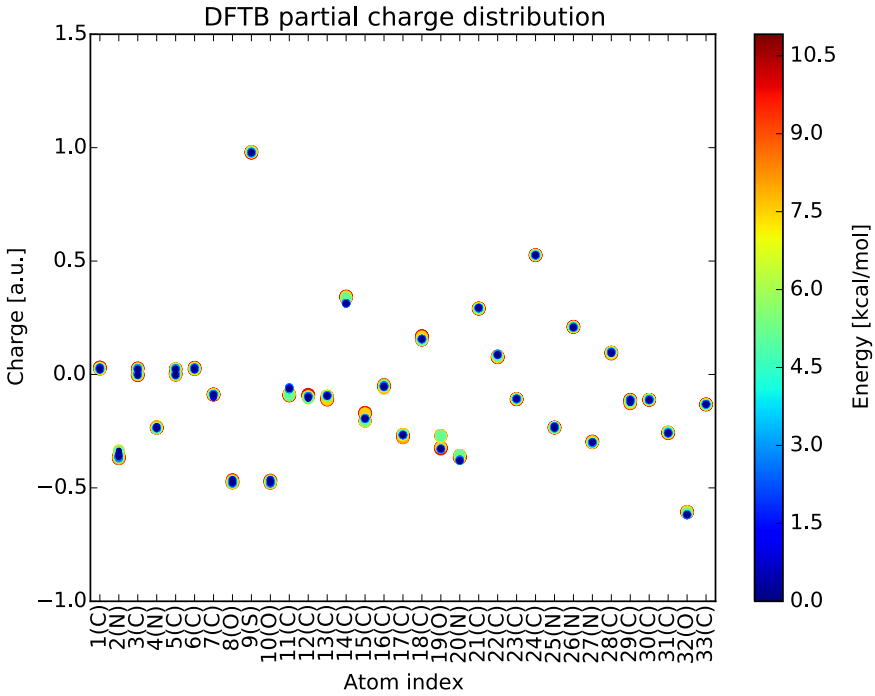

Fig. S52: DFTB charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. Structures were optimized with the DFTB method. The atom numbering is given in Fig. 3 A) of the article.

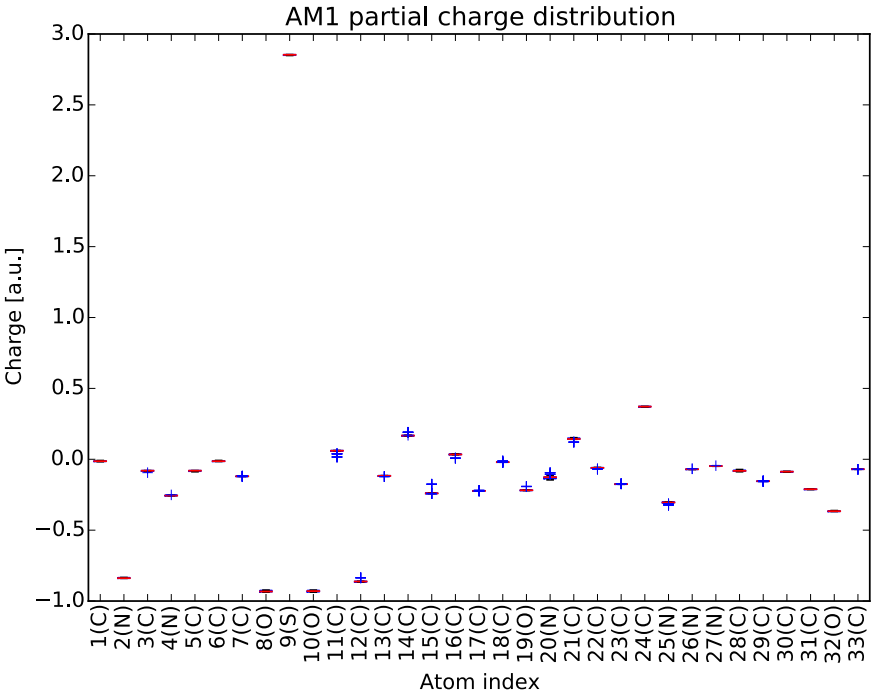

Fig. S53: Boxplot of AM1 charge distribution on heavy atoms of the Sildenafil conformational ensemble. Structures were optimized with the AM1 method. The atom numbering is given in Fig. 3 A) of the article.

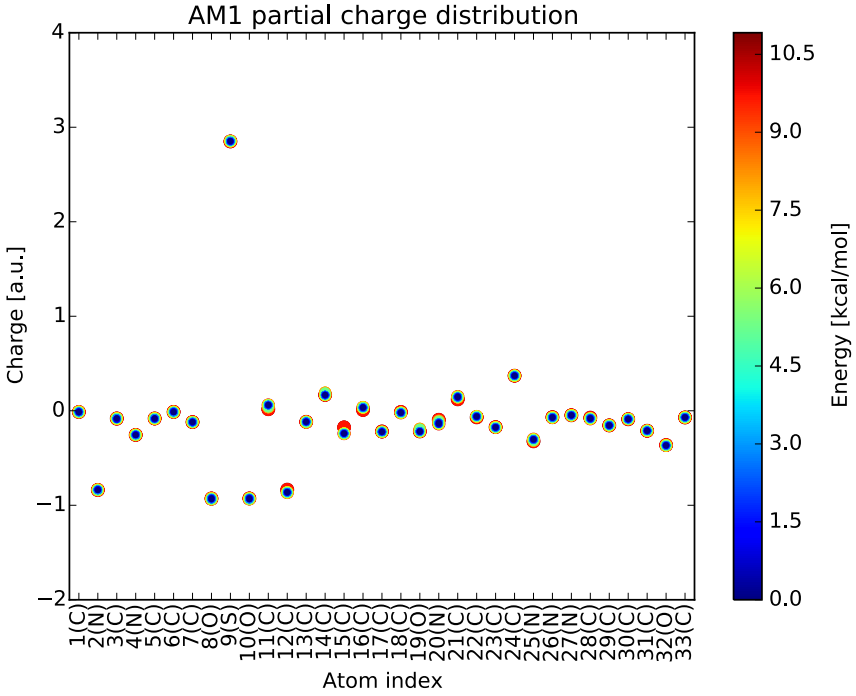

Fig. S54: AM1 charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. Structures were optimized with the AM1 method. The atom numbering is given in Fig. 3 A) of the article.

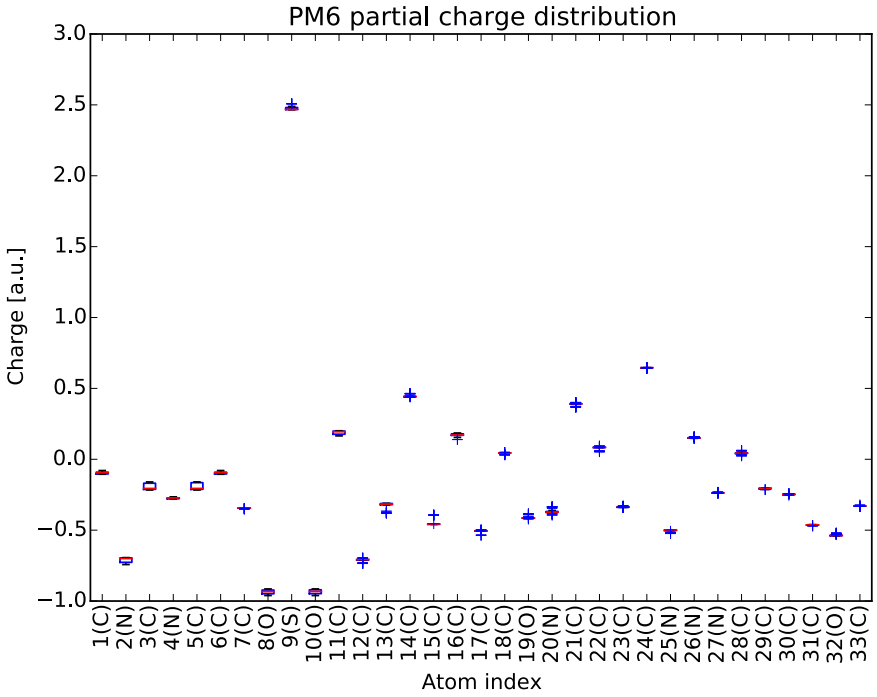

Fig. S55: Boxplot of PM6 charge distribution on heavy atoms of the Sildenafil conformational ensemble. Structures were optimized with the PM6 method. The atom numbering is given in Fig. 3 A) of the article.

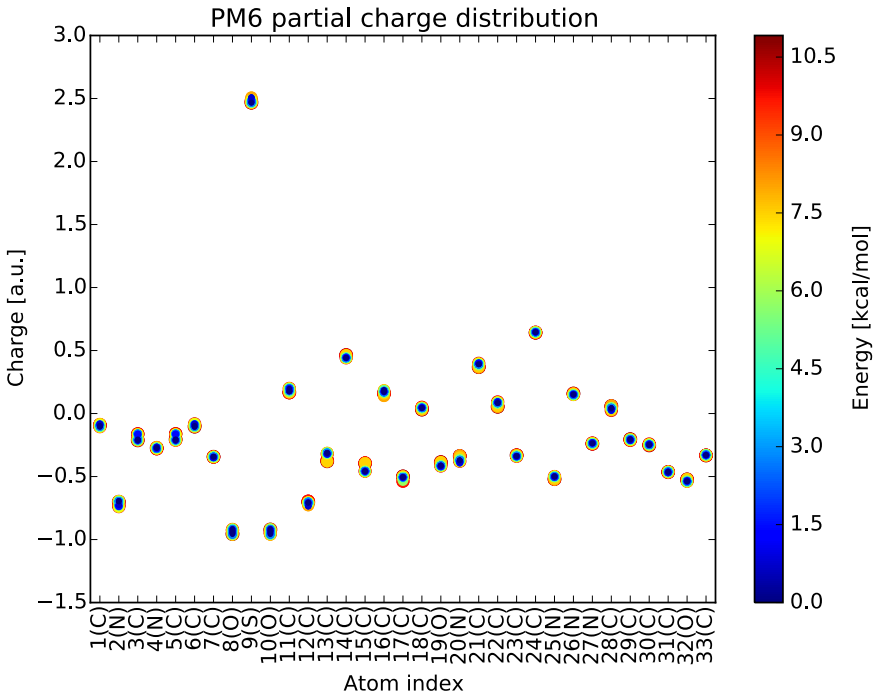

Fig. S56: PM6 charge distributions on heavy atoms of the Sildenafil+ conformational ensemble colored according to conformer energy. Structures were optimized with the PM6 method. The atom numbering is given in Fig. 3 A) of the article.

[1] J. Sadowski, J. Gasteiger and G. Klebe, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1000-1008.
[2] *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, **2015**.
[3] P. Labute, *J. Chem. Inf. Model.*, 2010, **50**, 792–800.
[4] Gaussian 09, Revision **D.01**, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.
[5] B. I. Dunlap, *J. Mol. Struc.-Theochem*, 2000, **529**, 37–40.
[6] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
[7] S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, 32, 1456-1465.
[8] H. Hu, Z. Lu and W. Yang, *J. Chem. Theory Comput.*, 2007, **3**, 1004–1013.
[9] J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.
[10] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
[11] J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396–1396.
[12] A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
[13] C. Lee, W. Yang and R. G. Parr, *Phys. Rev., B*, 1988, **37**, 785–789.
[14] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
[15] W. J. Hehre, R. F. Stewart and J. A. Pople, *J. Chem. Phys.*, 1969, **51**, 2657–2664.
[16] J. B. Collins, P. von R Schleyer, J. S. Binkley and J. A. Pople, *J. Chem. Phys.*, 1976, **64**, 5142–5151.
[17] A. Schaefer, H. Horn and R. Ahlrichs, *J. Chem. Phys.*, 1992, **97**, 2571–2577.
[18] F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
[19] F. Weigend, *Phys. Chem. Chem. Phys.*, 2006, **8**, 1057–1065.
[20] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, *J. Am. Chem. Soc.*, 1985, **107**, 3902–3909.
[21] J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173-1213.
[22] B. Aradi, B. Hourahine and T. Frauenheim, *J. Phys. Chem. A*, 2007, **111**, 5678–5684.
[23] M. Gaus, A. Goez and M. Elstner, *J. Chem. Theory Comput.*, 2013, **9**, 338-354.
[24] M. Gaus, X. Lu, M. Elstner and Q. Cui, *J. Chem. Theory Comput.*, 2014, **10**, 1518-1537.
[25] M. Kubillus, T. Kubar, M. Gaus, J. Řezáč and M. Elstner, *J. Chem. Theory Comput.*, 2015, **11**, 332-342.
[26] S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
[27] S. van der Walt, S. C. Colbert and G. Varoquaux, *Comput. Sci. Eng.*, 2011, **13**, 22–30.
[28] J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90–95.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825-2830.
[30] K. C. Gross, P. G. Seybold and C. M. Hadad, *Int. J. Quantum Chem.*, 2002, **90**, 445-458.