Supplementary Information: Machine learning of solvent effects on molecular spectra and reactions

Michael Gastegger,[1, 2, 3, a)] Kristof T. Schütt,[1, b)] and Klaus-Robert Müller[1, 4, 5]

[1)] *Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany*

[2)] *BASLEARN, BASF-TU joint Lab, Technische Universität Berlin, 10587 Berlin, Germany*

[3)] *DFG Cluster of Excellence "Unifying Systems in Catalysis" (UniSysCat), Technische Universität Berlin, 10623 Berlin, Germany*

[4)] *Department of Artificial Intelligence, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea*

[5)] *Max-Planck-Institut für Informatik, Saarbrücken, Germany*

[a)] Electronic mail: michael.gastegger@tu-berlin.de

[b)] Electronic mail: kristof.schuett@tu-berlin.de

Supplementary text 1: Response properties

The energy predicted by FieldSchNet is an analytic function of the coordinates $\mathbf{R}$, as well as the external fields and their associated atomic dipole moments. This makes it possible to access so-called response properties, which are partial derivatives of the potential energy[1]. Assuming the presence of an external electric $\boldsymbol{\epsilon}$ field and a magnetic field $\mathbf{B}$ with its corresponding nuclear magnetic moments $\{\mathbf{I}_i\}$, a general response property $\mathbf{\Pi}$ takes the form

$$\mathbf{\Pi}(n_{\mathbf{R}}, n_{\boldsymbol{\epsilon}}, n_{\mathbf{B}}, n_{\mathbf{I}_i}) = \frac{\partial^{n_{\mathbf{R}}+n_{\boldsymbol{\epsilon}}+n_{\mathbf{B}}+n_{\mathbf{I}_i}} E(\mathbf{R}, \boldsymbol{\epsilon}, \mathbf{B}, \mathbf{I}_i)}{\partial \mathbf{R}^{n_{\mathbf{R}}} \partial \boldsymbol{\epsilon}^{n_{\boldsymbol{\epsilon}}} \partial \mathbf{B}^{n_{\mathbf{B}}} \partial \mathbf{I}_i^{n_{\mathbf{I}_i}}}, \tag{1}$$

where the ns indicate the n-th order partial derivative w.r.t. the quantity in the subscript. A response property modeled by most machine learning potentials are the nuclear forces $\mathbf{F} = -\mathbf{\Pi}(1, 0, 0, 0)$, which are the negative first derivative of the energy with respect to the nuclear positions.

However, the expression above offers instructions on obtaining a wealth of other quantities, some of which are highly relevant for molecular spectroscopy and/or provide a direct connection to experiment. Infrared spectra can e.g. be simulated based on dipole moments $\boldsymbol{\mu} = -\mathbf{\Pi}(0, 1, 0, 0)$, while molecular polariziabilities $\boldsymbol{\alpha} = -\mathbf{\Pi}(0, 2, 0, 0)$ offer access to polarized and depolarized Raman spectra. A central response property of the magnetic field are nuclear magnetic shielding tensors $\boldsymbol{\sigma}_i = \mathbf{\Pi}(0, 0, 1, 1)$. These allow the computation of chemical shifts recorded in nuclear magnetic resonance spectroscopy NMR via their average trace $\sigma_i = \frac{1}{3}\mathrm{tr}[\boldsymbol{\sigma}_i]$.

The power of FieldSchNet (and field-based models in general) lies in the fact, that a single energy function provides access to a wide range of quantum chemical properties in a highly systematic manner. Moreover, the expression in Eq. 1 above guarantees the correct geometric transformations of the property tensors with respect to rotations and translations of the molecule in the external field without the need of explicitly encoding the corresponding symmetries. As is the practice with molecular forces, response properties can also be incorporated during training of the FieldSchNet model by including the appropriate squared errors into the loss function

$$\mathcal{L} = \eta_E(\tilde{E} - E)^2 + \frac{1}{3N}\sum_i^N |\tilde{\mathbf{F}}_i - \mathbf{F}_i|^2 + \eta_{\boldsymbol{\mu}}\frac{1}{3}|\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2 + \ldots \tag{2}$$

Here, the trade-offs η weight the importance of a property in the loss and N is the total number of atoms. The properties predicted by FieldSchNet according to Eq. 1 are indicated with a tilde.

Supplementary text 2: Reference data generation

All electronic structure reference computations were carried out at the PBE0/def2-TZVP[2,3] level of theory using the ORCA quantum chemistry package[4]. SCF convergence was set to tight and integration grid levels of 4 and 5 were employed during SCF iterations and the final computation of properties, respectively. In the case of allyl-p-tolyl ether, the RIJK approximation was used to accelerate computations[5]. Nuclear shielding tensors were computed with the Gauge Including Atomic Orbitals approach[6] implemented in ORCA, while continuum solvents calculations were performed with the in package conductor-like polarizable continuum model[7].

The reference data for ethanol was generated by selecting 10 000 random configurations from the MD17 database[8] and recomputing them at the above level of theory. In addition, continuum solvent calculations for the four studied continuum solvents (toluene, ethanol, methanol, water) were carried out for the structures selected in this manner. A training set for continuum models containing 30 000 ethanol configurations were constructed by merging the vacuum, ethanol and water data. The data was then filtered for configurations showing artificially high forces due to numerical problems in the surface cavity generation, yielding a final number of 27 990 structures. Reference data for the ML/MM simulations was generated in a two-step approach. Initially, a periodic box of 1250 ethanol molecules was equilibrated with the NAMD molecular dynamics package[9] using the CHARMM General Force Field[10] for 1 μs. Using the native NAMD interface to ORCA, electrostatic embedding QM/MM simulations were carried out, where one of the ethanols was described at the PBE/def2-SVP[3,11] level of theory. CHELPG charges[12] were used as partial charges for the quantum regions and simulations were run for 50 ps using 0.5 fs time steps. For all simulations, temperatures were kept at 300 K using a Langevin thermostat[13] and pressures at 1 atm using a Langevin piston barostat[14]. From this trajectory, 30 000 QM ethanol configurations and the associated charge distributions of the environment were sampled at random and recomputed at the PBE0/def2-TZVP level.

Reference data for the allyl-p-tolyl ether Claisen rearrangement reaction in vacuum was obtained via metadynamics[15] at the PBE/def2-SVP level of theory. The two bonds involved in the reaction were selected as collective coordinates and Gaussians with a height of 1 kcal/mol and a width of 0.529 Å were deposited each 100 simulation steps. The system was simulated for a total of 50 ps using 0.5 fs time steps. Temperature was kept constant at 500 K by means of a Nose-Hoover chain thermostat[16]. We then selected 61 000 configurations from this and recalculated them with the reference level of theory. Data for MM/ML simulations was generated by suspending 20334 configurations sampled during metadynamics in periodic solvation boxes with 9260 TIP3P waters[17]. Keeping the allyl-p-tolyl ether coodinates frozen, the water box was then optimized and simulated for 50 ps with NAMD. For temperature and pressure control, the same setup as in the ethanol box was used. From each of these boxes, 3 ether configurations and associated charge distributions were drawn and recomputed at the PBE0/def2-TZVP level, yielding 61 002 reference data points.

Supplementary text 3: Model training

The training settings for each data set are provided in Supplementary Tab. 1. The initial learning rates were decayed by a factor of 0.8 after t_{patience} epochs of no improvement. Training was stopped the learning rate reached a value of 1e-6. The dipole cutoffs were chosen to be the same as the cutoffs for the interactions. Supplementary Tab. 2 provides the tradeoffs *eta* used for the different response properties in the composited loss function minimized during training. As the ^{1}H, ^{13}C and ^{17}O chemical shifts lie on completely different scales, the shielding tensor loss terms for each atom were weighted by an element dependent factor in order to achieve equal relative accuracy between all contributions. We used factors of $\omega_H = 1.0$, $\omega_C = 0.167$ and $\omega_O = 0.022$, which were determined based on the reference data. The FieldSchNet model and training procedures were implemented using PyTorch[18] and the SchNetPack code package for machine learning in atomistic systems[19].

Supplementary text 4: Molecular dynamics and spectra computation

Unless stated otherwise, the velocity Verlet algorithm and a time step of 0.5 fs were used to integrate the equations of motion. All simulations not using NAMD[9] were carried out with the molecular dynamics module implemented in SchNetPack[19].

Supplementary Table 1: Training parameters for all models.

Dataset	n_{train}	n_{valid}	n_{test}	n_{batch}	lr_{init}	t_{patience}	n_{features}	$n_{\text{interactions}}$	r_{cutoff} [Å]
ethanol (vacuum)	8000	1000	1000	20	1e-4	15	256	6	5.0
ethanol (continuum)	16 000	2000	9990	20	1e-4	15	256	6	5.0
ethanol (ML/MM)	18 000	2000	10 000	20	1e-4	15	256	6	5.0
ethanol (ML/MM, reduced)	1800	200	28 000	20	1e-4	15	256	6	5.0
ethanol + methanol (ML/MM)	1890	105	105	20	1e-4	15	256	6	5.0
ether (vacuum)	50 000	5000	6000	10	1e-4	25	256	5	5.0
ether (ML/MM)	50 000	5000	6002	10	1e-4	25	256	5	5.0

Supplementary Table 2: Tradeoffs η used for training the different properties, assuming all quantities use atomic units.

	ethanol	allyl-p-tolyl ether	
Property	all	vacuum	ML/MM
$\mathbf{E}$	1.0	1.0	1.0
$\mathbf{F}$	10.0	5.0	5.0
$\boldsymbol{\mu}$	0.01	0.05	0.01
$\boldsymbol{\alpha}$	0.01	0.001	0.0001
$\boldsymbol{\sigma}_{\text{all}}$	0.05	10.0	0.1

Classical molecular dynamics simulations for ethanol in vacuum and continuum solvents were carried for 50 ps at a temperature of 300 K controlled via Nose-Hoover chain[16] thermostat with a chain length of 3 and time constant of 100 fs. The first 10 ps of these trajectories were then discarded. Ring polymer molecular dynamics were performed for 20 ps, using a time step of 0.2 fs and a specially adapted global Nose-Hoover chain as introduced in Ref.[20] to keep the temperature at 300 K. Once again, a chain length of 3 and time constant of 100 fs were chosen for the thermostat.

Simulations for ethanol ML/MM models were carried out using a custom interface between NAMD and our machine learning code. First, a periodic box of 1250 ethanol molecules

Supplementary Table 3: **Test set performance of ethanol models.** Mean absolute errors of FieldSchNet trained on ethanol in vaccuum and pc-FieldSchNet trained with vaccum, ethanol and water as solvents. Solvents marked by * have not been used to train the continuum model.

Property	**Unit**	**Vacuum**	**Continuum**					**ML/MM**
			vacuum	*toluene**	*ethanol*	*methanol**	*water*	
E	$\mathrm{kcal\,mol^{-1}}$	0.017	0.035	0.137	0.052	0.056	0.062	0.557
$\mathbf{F}$	$\mathrm{kcal\,mol^{-1}}\,\AA^{-1}$	0.128	0.145	0.174	0.139	0.140	0.142	0.683
$\boldsymbol{\mu}$	D	0.004	0.004	0.006	0.005	0.005	0.005	0.007
$\boldsymbol{\alpha}$	Bohr^3	0.008	0.007	0.243	0.007	0.007	0.008	0.010
$\boldsymbol{\sigma}_{\mathrm{all}}$	ppm	0.169	0.157	0.149	0.140	0.140	0.141	0.154
σ_H	ppm	0.123	0.122	0.116	0.113	0.113	0.114	0.094
σ_C	ppm	0.194	0.186	0.175	0.166	0.166	0.167	0.182
σ_O	ppm	0.401	0.312	0.298	0.248	0.248	0.250	0.453

Supplementary Table 4: Test set errors obtained for the allyl-p-toly Claisen rearrangement datasets.

Property	**Unit**	**Vacuum**	**ML/MM**
$\mathbf{E}$	$\mathrm{kcal\,mol^{-1}}$	0.084	0.400
$\mathbf{F}$	$\mathrm{kcal\,mol^{-1}}\,\AA^{-1}$	0.141	0.454
$\boldsymbol{\mu}$	D	0.003	0.026
$\boldsymbol{\alpha}$	Bohr^3	0.039	0.157
$\boldsymbol{\sigma}_{\mathrm{all}}$	ppm	0.273	1.144
σ_H	ppm	0.045	0.154
σ_C	ppm	0.301	1.331
σ_O	ppm	2.732	11.144

was equilibrated with NAMD for 1 μ, using the CHARMM General Force Field[10]. Bonds to hydrogens were kept frozen with the SHAKE algorithm[21] and a time step of 1 fs was used. One ethanol was then selected for modeling via the FieldSchNet ML/MM model and ML/MM simulations were carried out using a custom interface between NAMD and our machine learning code for a total of 50 ps. For both simulations, temperatures were kept at 300 K with a Langevin thermostat[13] and pressures at 1 atm using Langevin piston barostat[14]. The first 10 ps of the trajectory were discarded.

Simulations for methanol ML/MM models were carried out with the same protocol, using a periodic box of 1860 methanol molecules.

Umbrella sampling simulations for the Claisen rearrangement set up according to the following protocol. Using the difference between the bonds formed and broken as the reaction coordinate, we determined the centers for the harmonic bias potentials by choosing 50 equidistant points along the reaction coordinate. The centers ranged from values of -4.15 Å to 5.18 Å with an increment of 0.19 Å. For each center, we selected the closest lying structure in the metadynamics trajectory used for generating the reference data as a starting configuration for the umbrella sampling run. All simulations used a force constant of 112.04 kcal/mol/Å^2. Umbrella sampling in vacuum was carried out for each window by first equilibrating the system for 25 ps using a Berendsen thermostat[22] at 300 K (time constant of 100 fs) followed by 25 ps production simulation at the same temperature with a Nose-Hoover chain (chain length of 3 and time constant of 100 fs). For the ML/MM model umbrella simulations, the starting configurations were first solvated in a periodic box of 9260 water molecules treated with the TIP3P force field. Keeping the allyl-p-tolyl ether structures frozen, the water box was first minimized and the equilibrated for 200 ps to a temperature of 300 K and pressure of 1 atm with a Langevin thermostat and Langevin piston barostat using NAMD. Bonds involving water hydrogens were kept frozen with the SHAKE algorithm and a time step of 1 fs was used. Starting from the systems prepared in this manner, ML/MM simulations were performed for 25 ps using the same pressure and temperature control as above.

Free energy profiles were constructed from the umbrella sampling data using the WHAM code with convergence set to 1e-9 and a temperature of 300 K[23,24].

Infrared and polarized as well as depolarized Raman spectra were computed from the time-autocorrelation functions of the dipole moment and polarizability time derivatives according to the relations given in Ref.[25]. Autocorrelation functions were computed using the Wiener-Khinchin theorem[26] and a autocorrelation depth of 2048 fs. In order to enhance the quality of the spectra, a Hann window function[27] and zero-padding were applied to the autocorrelation functions before computing the spectra. A laser frequency of 514 nm and temperature of 300 K were used for calculating the Raman spectra.

NMR chemical shifts were computed as the average trace of the nuclear shielding tensor $\sigma_i = \frac{1}{3}\mathrm{tr}[\boldsymbol{\sigma}_i]$. These chemical shifts were then referenced to the shifts computed for a tetramethylsilane molecule via

$$\sigma_i = \sigma_{\mathrm{ref}}^{(Z)} - \sigma_i \tag{3}$$

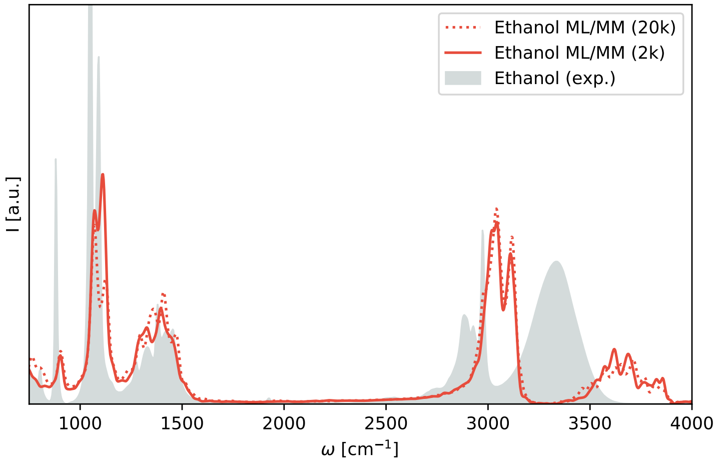
The reference shifts computed with the PBE0/def2-TZVP were $\sigma_{\mathrm{H}} = 31.77$ ppm and $\sigma_{\mathrm{H}} = 188.53$ ppm.

Supplementary text 5: Adaptive sampling

In order to obtain representative structures from the original ethanol ML/MM data, we applied adaptive sampling in a postprocessing manner. Starting from a randomly selected set of 100 structures, we trained three separate FieldSchNet models using the settings and tradeoffs reported in supplementary Tabs 1 and 2, omitting the nuclear shielding tensors. We then used a weighted sum of the variances of the network predictions as an uncertainty measure:

$$\nu = \sum_{\pi} \frac{w_\pi}{M-1} \max\left[\sum_{m}^{M}\left(\mathbf{\Pi}_{\pi}^{(m)} - \frac{1}{M}\sum_{m}^{M}\mathbf{\Pi}_{\pi}^{(m)}\right)^2\right], \tag{4}$$

where π is the index for a particular property, w_π the associated tradeoff weight, M the total number of ensemble models and $\mathbf{\Pi}_{\pi}^{(m)}$ the prediction of model m for property π. In case of a vectorial or tensorial property (e.g. dipole moment), the element with the maximal value was used. As weights, we used the same tradeoffs as reported in supplementary Tab. 2. The 100 molecules with the highest uncertainty were then selected and added to the initial dataset. A new model ensemble was trained on this new data and the procedure was repeated until satisfactory accuracy was achieved. The final dataset contained 2000 ethanol configurations, resulting in a 10-fold reduction in training data. The ML/MM spectrum simulated with a FieldSchNet model trained on the reduced dataset (using the settings described in supplementary text 4) shows excellent agreement with the original spectrum



Supplementary Figure 1: **Ethanol ML/MM spectrum of the reduced dataset:** Comparison of the ML/MM spectra obtained with FieldSchNet models trained on the full (20 000 data points) and reduced datasets (2000 data points). The experimental spectrum is shown in gray.

(see supplementary Fig 1).

A modified adaptive sampling procedure was used to extend the FieldSchNet ML/MM model to liquid ethanol. Starting with an ensemble of three FieldSchNet models trained on the reduced dataset generated above, we performed 10 ps of ML/MM simulations on an equilibrated box of 1860 methanol molecules (see supplementary text 4). During simulation, uncertainties were computed for each timestep according to supplementary Eq. 4, yielding the model uncertainty as a function of time. After the simulation, we determined the maxima of this function and selected those with the 100 highest associated uncertainties. The corresponding structures were then recomputed with the electronic structure reference used to generated the original data (supplementary text 2). Finally, new ML models were trained on the dataset expanded in this manner (2000 ethanol structures + 100 methanol structures) and used in ML/MM simulations to compute infrared spectra.

REFERENCES

[1] Jensen, F. *Introduction to Computational Chemistry* (Wiley, 2007).

[2] Adamo, C. & Barone, V. Toward reliable density functional methods without adjustable parameters: The pbe0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).

[3] Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).

[4] Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2**, 73–78 (2012).

[5] Weigend, F. A fully direct ri-hf algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **4**, 4285–4291 (2002).

[6] Helgaker, T., Jaszuński, M. & Ruud, K. Ab initio methods for the calculation of nmr shielding and indirect spin- spin coupling constants. *Chem. Rev.* **99**, 293–352 (1999).

[7] Barone, V. & Cossi, M. Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **102**, 1995–2001 (1998).

[8] Chmiela, S. *et al.* Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).

[9] Phillips, J. C. *et al.* Scalable molecular dynamics with namd. *J. Comput. Chem.* **26**, 1781–1802 (2005).

[10] Vanommeslaeghe, K. *et al.* Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).

[11] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).

[12] Breneman, C. M. & Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. the need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **11**, 361–373 (1990).

[13] Bussi, G. & Parrinello, M. Accurate sampling using langevin dynamics. *Phys. Rev. E* **75**, 056707 (2007).

[14] Feller, S. E., Zhang, Y., Pastor, R. W. & Brooks, B. R. Constant pressure molecular dynamics simulation: The langevin piston method. *J. Chem. Phys.* **103**, 4613–4621 (1995).

[15] Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *WIREs Comput. Mol. Sci.* **1**, 826–843 (2011).

[16] Martyna, G. J., Klein, M. L. & Tuckerman, M. Nos-hoover chains: The canonical ensemble via continuous dynamics. *J. Chem. Phys.* **97**, 2635–2643 (1992).

[17] MacKerell Jr, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[18] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

[19] Schütt, K. T. *et al.* SchNetPack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2018).

[20] Ceriotti, M., Parrinello, M., Markland, T. E. & Manolopoulos, D. E. Efficient stochastic thermostatting of path integral molecular dynamics. *J. Chem. Phys.* **133**, 124104 (2010).

[21] Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327 – 341 (1977).

[22] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

[23] Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* **13**, 1011–1021 (1992).

[24] Grossfield, A. Wham: the weighted histogram analysis method. `http://membrane.urmc.rochester.edu/wordpress/?page_id=126` (2.0.9.1).

[25] Thomas, M., Brehm, M., Fligg, R., Vöhringer, P. & Kirchner, B. Computing vibrational spectra from ab initio molecular dynamics. *Phys. Chem. Chem. Phys.* **15**, 6608–6622 (2013).

[26] Wiener, N. Generalized harmonic analysis. *Acta Math.* **55**, 117–258 (1930).

[27] Blackman, R. B. & Tukey, J. W. The measurement of power spectra from the point of view of communications engineering – Part I. *Bell Syst. Tech. J.* **37**, 185–282 (1958).