Supplementary Information (SI) for Digital Discovery.

Open-Source Generation of Sigma Profiles: Impact of Quantum Chemistry and Solvation Treatment on Machine Learning Performance

Fathya Salih[1], Dinis Abranches[2], Edward J. Maginn[1], Yamil J. Colón[1]

[1]Department of Chemical Engineering, University of Notre Dame, Notre Dame, IN 46556, United States

[2]CICECO – Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, Aveiro, 3810-193, Portugal

1 Additional ML Performance Results

1.1 Effect of Averaging Radius – For Additional Levels of Theory

Figure S1 through S5 below show the effect of averaging for all generated datasets, except HF/def2-SVP-YK which was presented in the main text. All figures illustrate the same point concluded in the main text, wherein, larger averaging radii lead to poor performance, and not averaging sigma profiles ensures consistently better testing performance for all properties.

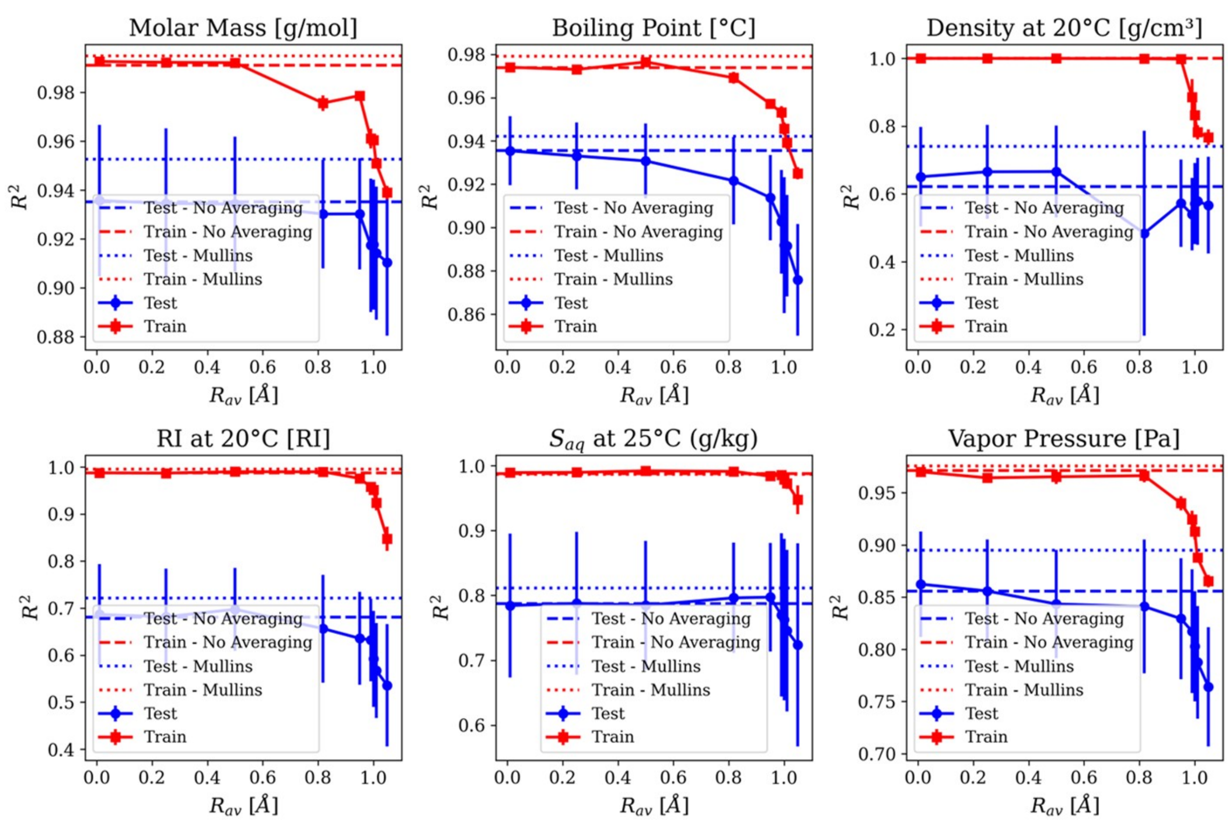

Figure S1: Effect of averaging radius on the GP performance at predicting different target properties using SPs generated using HF and the KS COSMO model. The dashed line indicates performance without any SP smoothing/averaging and the dotted line indicates the benchmark *performance using the Mullins SPs. Error bars indicate the variance in performance with cross validation using 10 folds.*

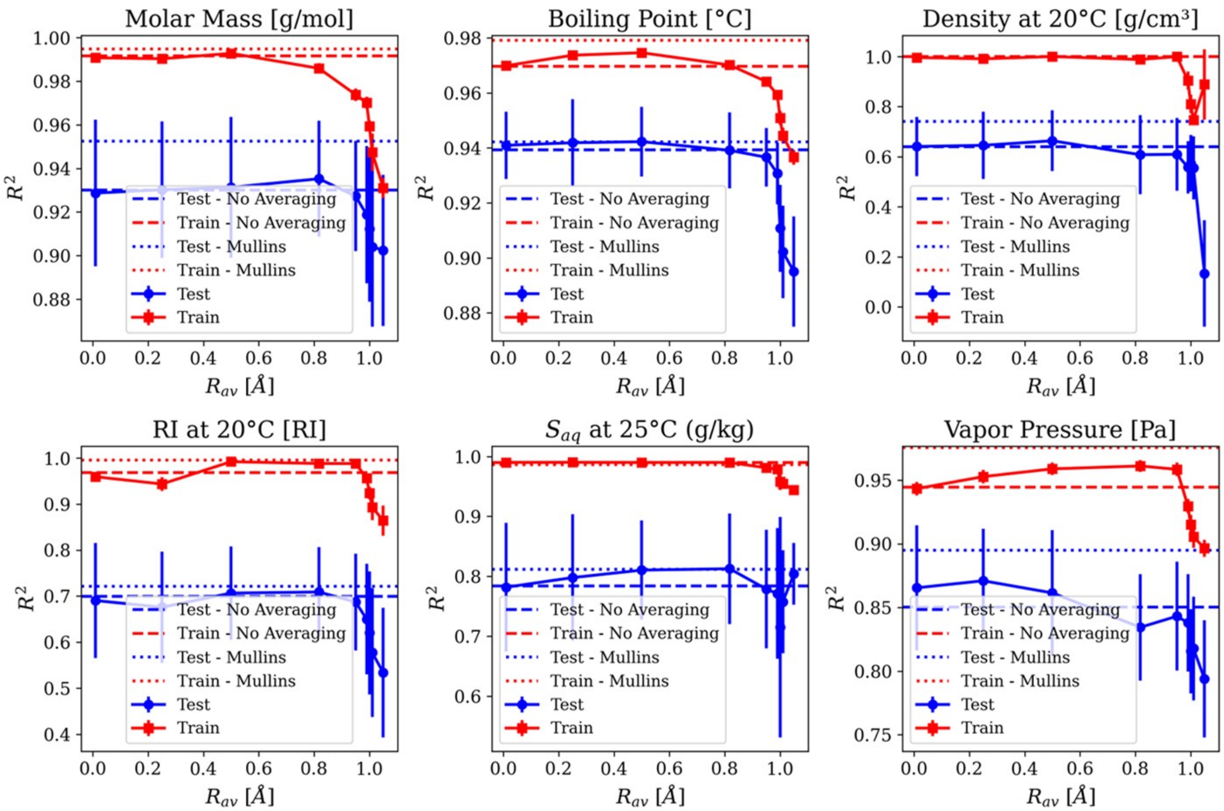

*Figure S2: Effect of averaging radius on the GP performance at predicting different target properties using SPs generated using 6-31G** and the KS COSMO model. The dashed line indicates performance without any SP smoothing/averaging and the dotted line indicates the benchmark performance using the Mullins SPs. Error bars indicate the variance in performance with cross validation using 10 folds.*

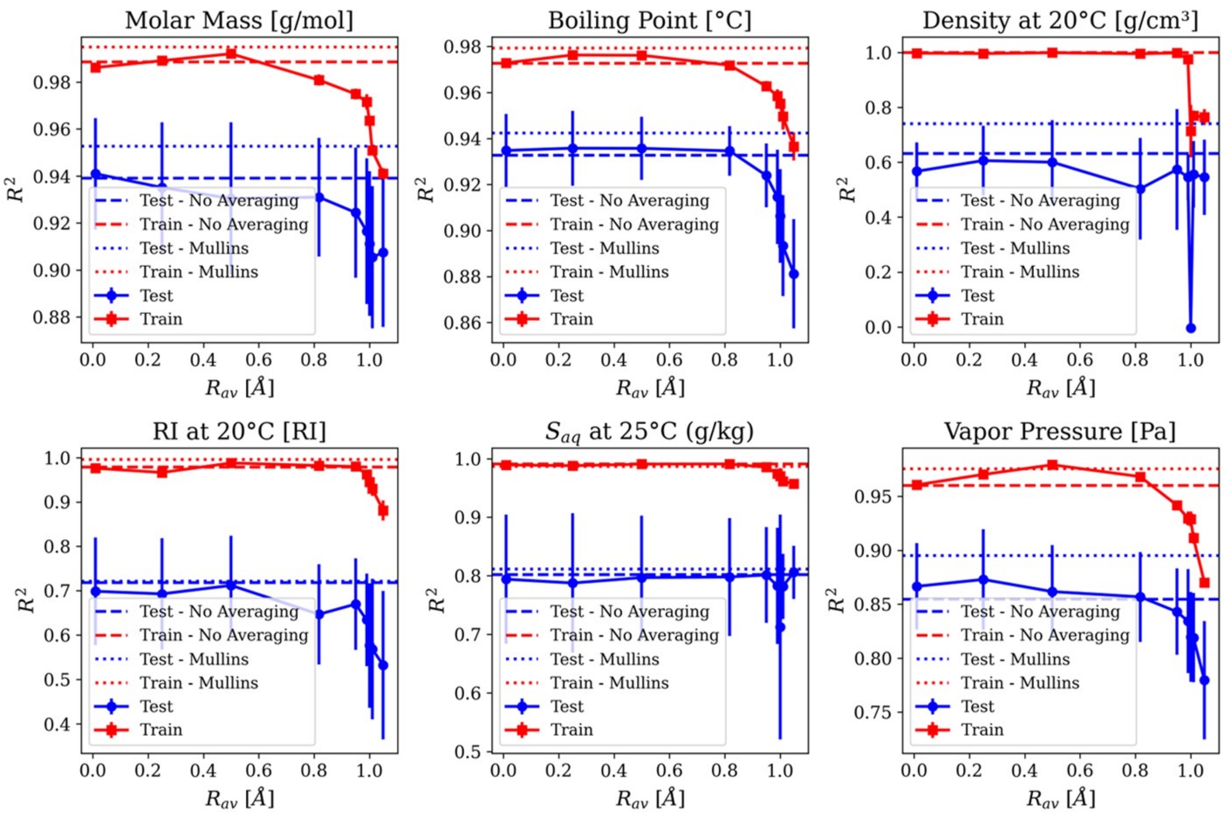

Figure S3: Effect of averaging radius on the GP performance at predicting different target properties using SPs generated using TZVP and the KS COSMO model. The dashed line indicates performance without any SP smoothing/averaging and the dotted line indicates the benchmark performance using the Mullins SPs. Error bars indicate the variance in performance with cross validation using 10 folds.

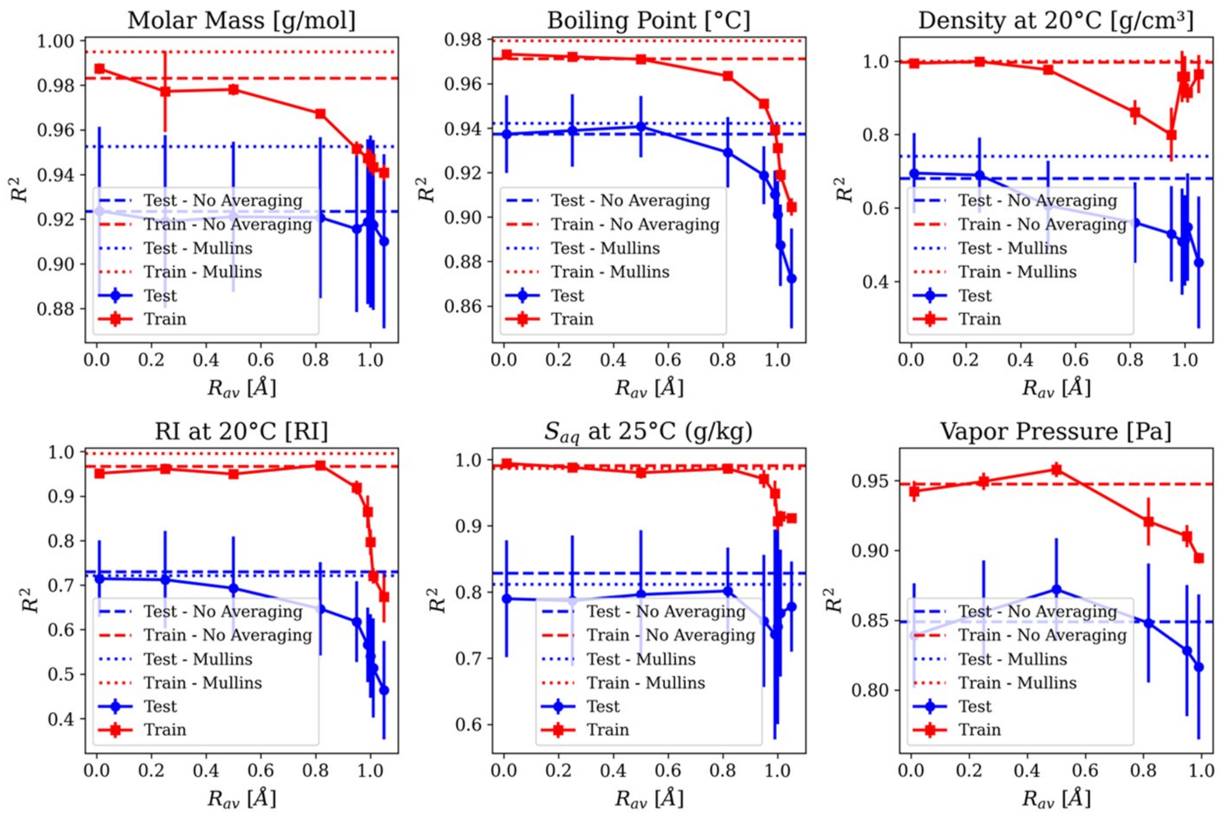

*Figure S4: Effect of averaging radius on the GP performance at predicting different target properties using SPs generated using 6-31G** and the YK COSMO model. The dashed line indicates performance without any SP smoothing/averaging and the dotted line indicates the benchmark performance using the Mullins SPs. Error bars indicate the variance in performance with cross validation using 10 folds.*

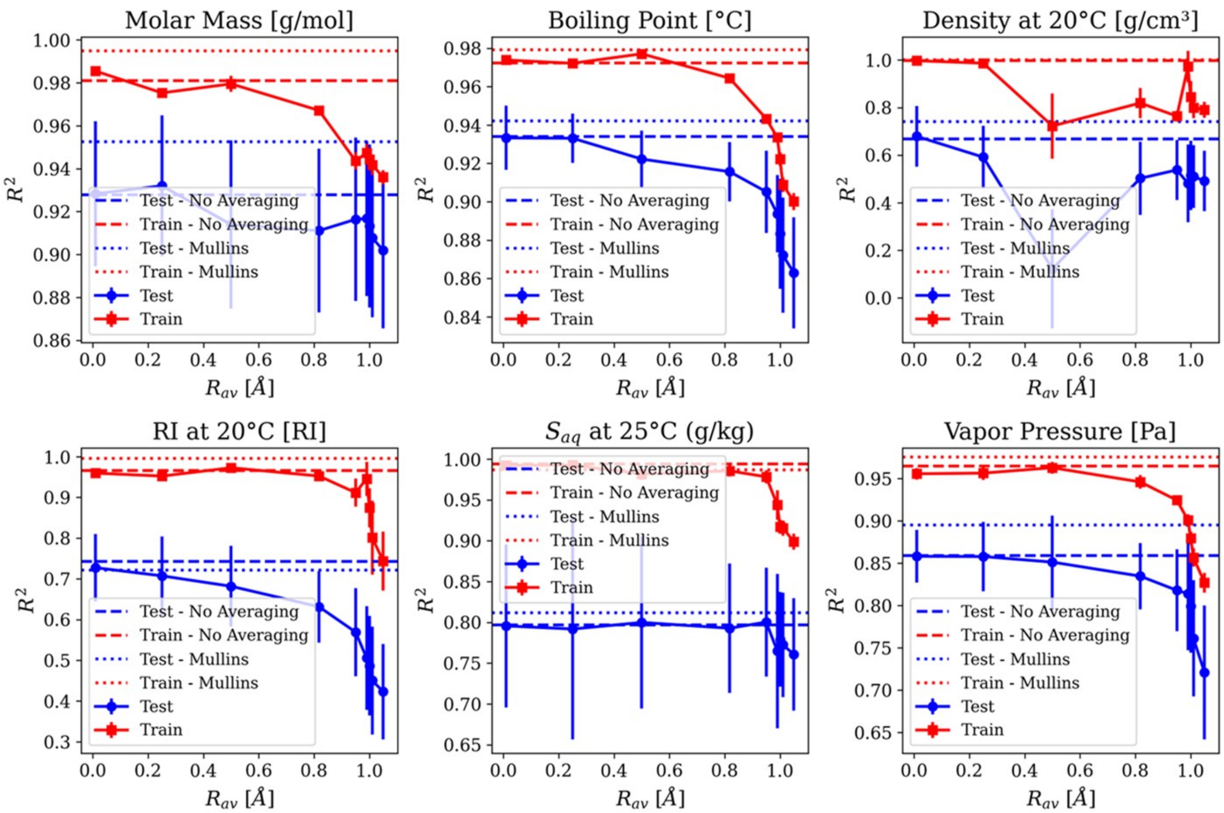

Figure S5: Effect of averaging radius on the GP performance at predicting different target properties using SPs generated using TZVP and the YK COSMO model. The dashed line indicates performance without any SP smoothing/averaging and the dotted line indicates the benchmark performance using the Mullins SPs. Error bars indicate the variance in performance with cross validation using 10 folds.

1.2 Effect of Level of Theory and COSMO Model – at Selected R_{av}

This section shows additional figures for the ML performance of different unaveraged SP datasets at selected averaging radii. Figure S6 and Figure S7 show ML performances for 0.5 Å (the averaging radius recommended by Klamt)[1] and 0.81742 Å (the averaging radius recommended by Mullins,[2] respectively.

The effect of COSMO model seems to shift, wherein, the KS model either matches or outperforms the YK model for most properties when averaging radius is increased (especially for molar mass). However, considering that even the benchmark performance drops when averaging is applied, this observation does not change the recommendations made in the main script of this article.

As for level of theory, the trend of slightly increasing performance with increasing level of theory established in the no-averaging case presented in the main text (see Figure 6) is broken. As the performance of the highest level of theory (B3LYP/def2-TZVP) significantly drops upon averaging. That effect is, also, non-monotonic. As the drop in performance for B3LYP/def2-TZVP is more significant for $R_{av} = 0.5$ Å than it is for $R_{av} = 0.82$ Å.

Finally, Figure S8 shows the effect of level of theory for the no averaging case, but only for the YK model – to show the trends explained in Figure 6 of the main script more clearly.

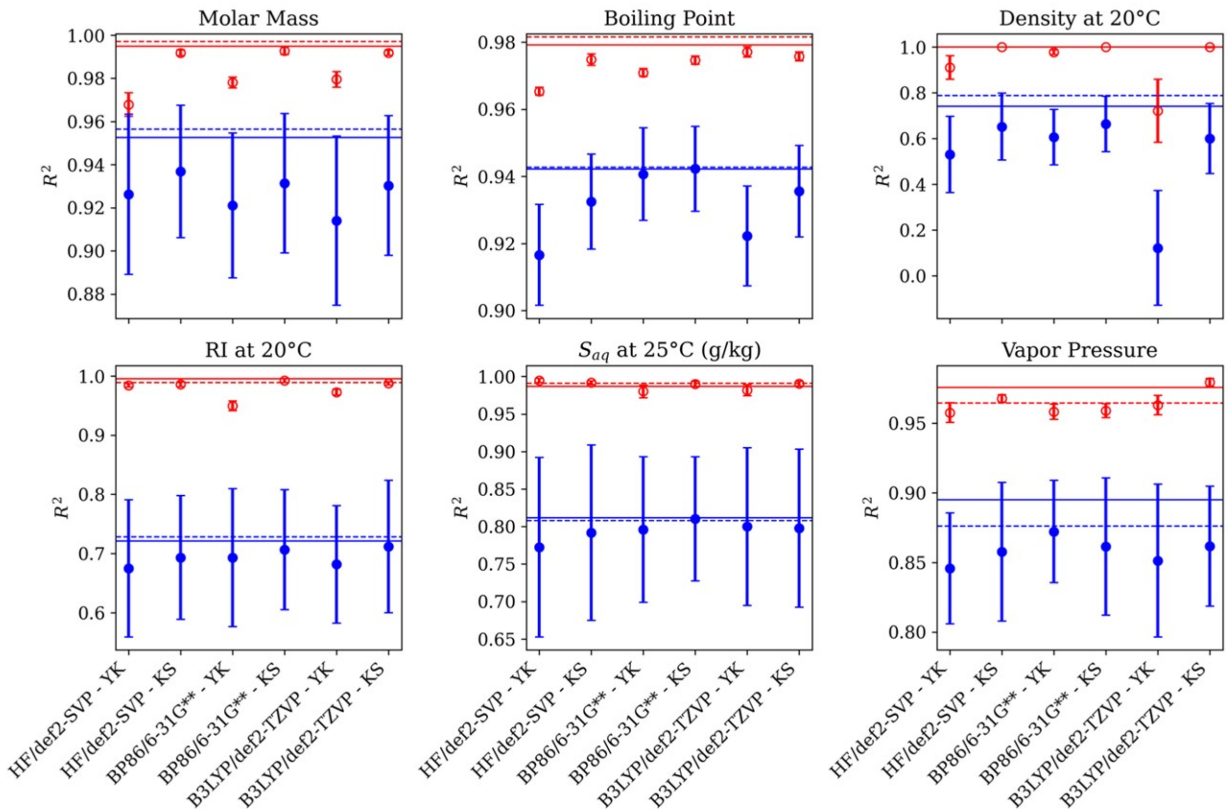

Figure S6: Effect of level of theory and COSMO model on GP performance for different target properties. Blue dots indicate performance on the testing set, while hollow red dots indicate performance on the training set. Error bars are the standard deviation in R^2 values for 10 stratified data splits (cross-validation folds). The solid lines indicate the benchmark performance from the Mullins dataset, while dashed lines indicate performance from the unaveraged Mullins dataset. All datasets made up of sigma profiles averaged with the Klamt-recommended $R_{av} = 0.5$ *Å*

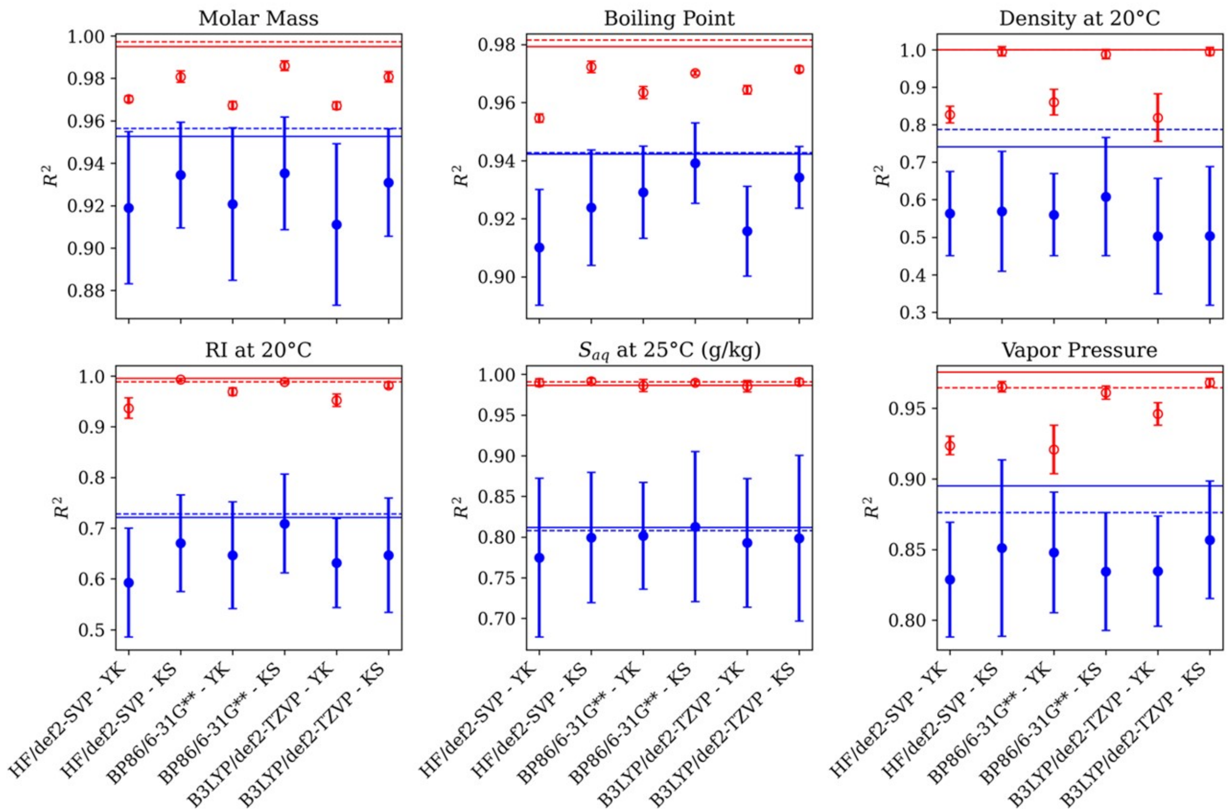

Figure S7: Effect of level of theory and COSMO model on GP performance for different target properties. Blue dots indicate performance on the testing set, while hollow red dots indicate performance on the training set. Error bars are the standard deviation in R^2 *values for 10 stratified data splits (cross-validation folds). The solid lines indicate the benchmark performance from the Mullins dataset, while dashed lines indicate performance from the unaveraged Mullins dataset. All datasets made up of sigma profiles averaged with Mullins-recommended* $R_{av} = 0.82$ *Å*

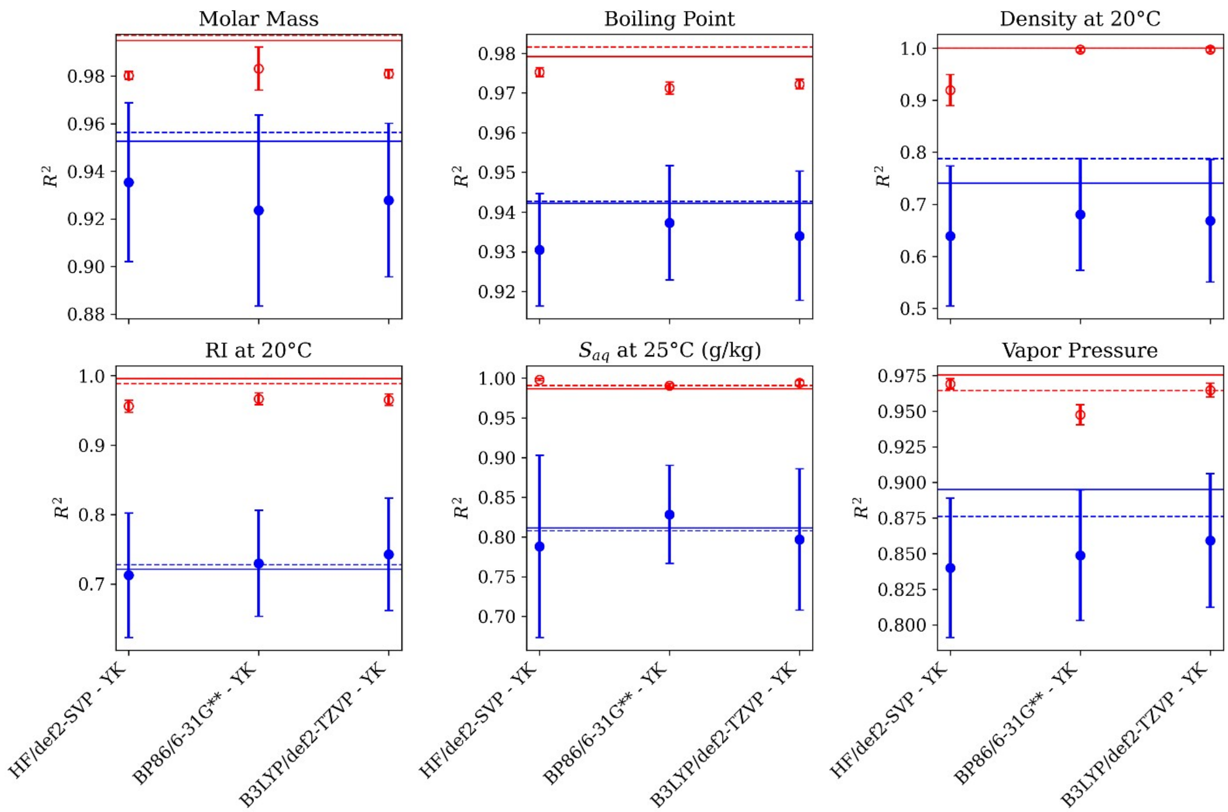

Figure S8: Effect of level of theory on GP performance for different target properties with the YK COSMO model fixed. Blue dots indicate performance on the testing set, while hollow red dots indicate performance on the training set. Error bars are the standard deviation in R^2 values for 10 stratified data splits (cross-validation folds). The solid lines indicate the benchmark performance from the Mullins dataset, while dashed lines indicate performance from the unaveraged Mullins dataset. All datasets made up of unaveraged sigma profiles

1.3 Correlations Between Performance and Information Metrics

This section contains the trends between the area under SP curve and sigma-range information metrics and ML performance. Figure S9 and Figure S10 plot the ML performance of each dataset against the dataset's average information metric for area under SP and sigma-range, respectively. The figures show that area under SP and charge range do not exhibit consistent trends with ML performance (i.e. slopes of linear trends change between positive to negative for different properties and some properties exhibit a flat linear trend or no dependence on the information metrics).

Additionally, Table S1 lists the average and standard deviation of each information metric for all considered SP datasets.

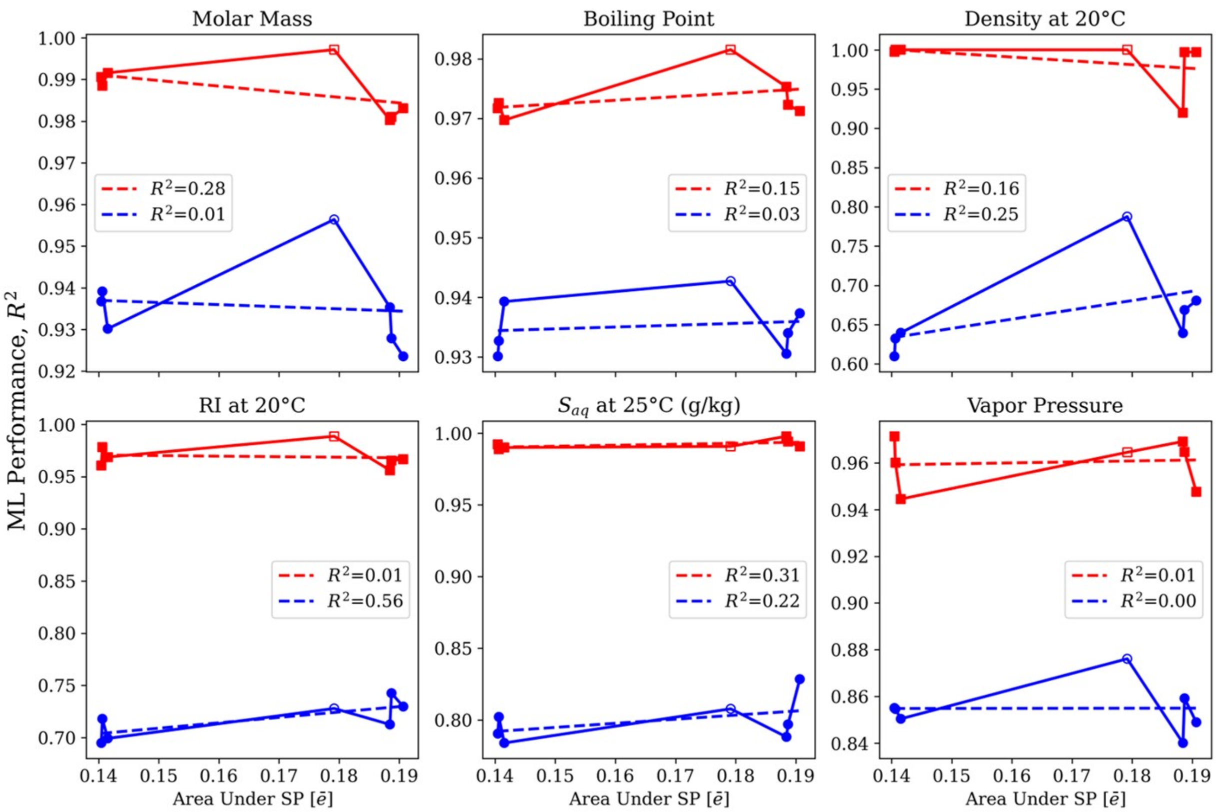

Figure S9: Correlation between the area under SP and GP performance. Blue lines with circle markers indicate testing set performance, while the red lines with square markers are for training set performance. Hollow markers are for the Mullins or reference datasets, while filled-in markers are for the generated datasets. Dashed lines indicate the linear fit between information metrics and ML performance, with the goodness of fit shown in terms of R^2 in the legend. Error bars are omitted for clarity, and all datasets considered are for un-averaged SPs (including the Mullins dataset).

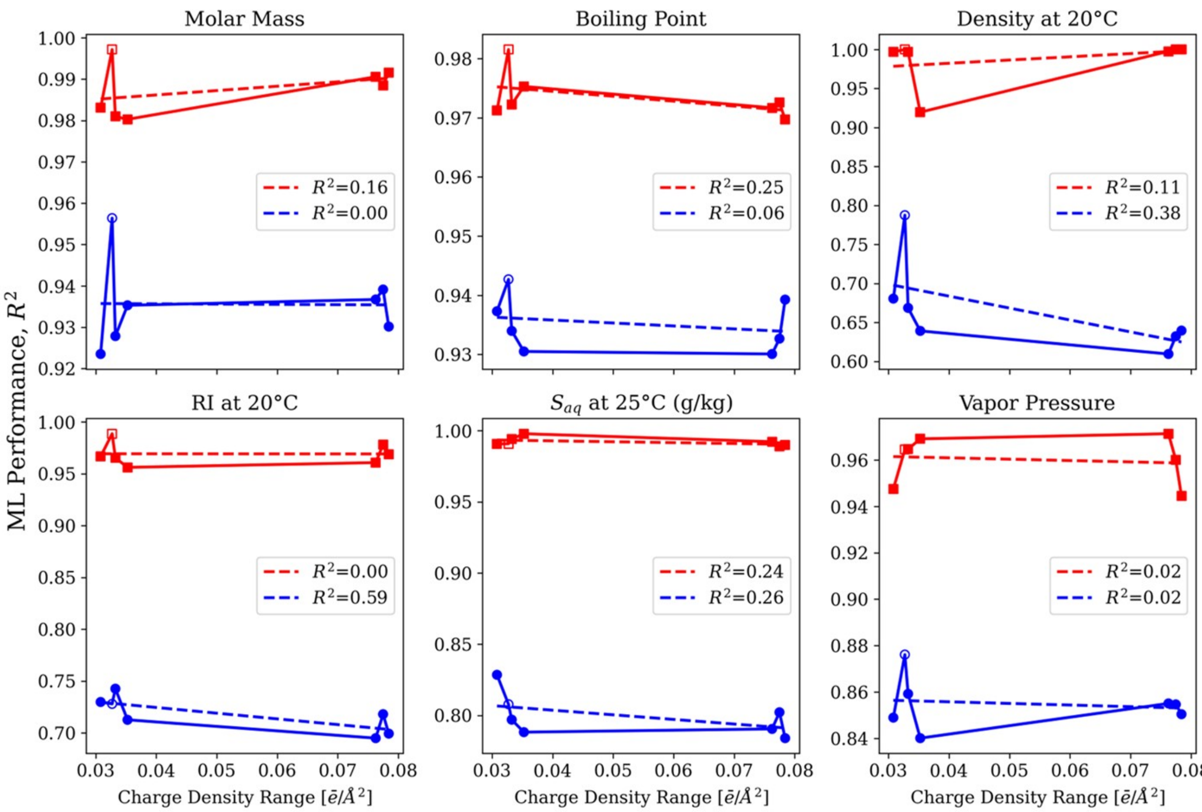

Figure S10: Correlation between the $\sigma -$ range and GP performance. Blue lines with circle markers indicate testing set performance, while the red lines with square markers are for training set performance. Hollow markers are for the Mullins or reference datasets, while filled-in markers are for the generated datasets. Dashed lines indicate the linear fit between information metrics and ML performance, with the goodness of fit shown in terms of R^2 in the legend. Error bars are omitted for clarity, and all datasets considered are for un-averaged SPs (including the Mullins dataset).

Table S1: Information metrics for each unaveraged SP dataset with information metric standard deviations in parenthesis

SP Dataset	Information Metric			
	Area Under SP $[e]$	**σ – Range** $[e/Å^2]$	**Hellinger Distance to** $\delta(\sigma)$	**Non-Uniformity** $[Å^2/e]$
HF/def2-SVP – YK	0.19 (0.08)	0.04 (0.02)	11.63 (2.09)	1512 (1683)
HF/def2-SVP – KS	0.14 (0.05)	0.08 (0.04)	10.37 (1.75)	1376 (1311)
BP86/6-31G** – YK	0.19 (0.08)	0.03 (0.01)	11.83 (2.29)	1829 (2261)
BP86/6-31G** – KS	0.14 (0.05)	0.08 (0.05)	10.49 (1.85)	1676 (1494)
B3LYP/def2-TZVP – YK	0.19 (0.08)	0.03 (0.01)	11.77 (2.20)	1722 (1784)
B3LYP/def2-TZVP – KS	0.14 (0.05)	0.08 (0.05)	10.47 (1.81)	1472 (1250)
Mullins (no averaging)	0.18 (0.07)	0.03 (0.01)	13.14 (2.51)	607 (296)

1.4 Effect of Segment Size (Reduced Datasets)

In the main text, the effect of segment size on ML performance was shown for 3 and 4 tessellation passes, but the effect of reducing the number of tessellation passes to 2 was omitted due to convergence issues. Here, the 90 un-converged molecules from the 2-tessellation dataset are listed in Table S2 and ML performance results excluding those molecules is shown in Figure S11. Additionally, to help with GP fitting the reduced datasets, a Matern52 kernel was used (see equation 1) as opposed to the RBF kernel used for all other results presented in the main text and in other sections of this document (equation 2).

$$k(r) = \sigma^2\left(1 + \sqrt{5}\, r + 5/3r^2\right)\exp\left(-\sqrt{5}\, r\right) \quad 1$$

$$k(r) = \sigma^2\exp\left(-\tfrac{1}{2}\, r\right) \quad 2$$

Where $'r'$ is the Euclidean distance between the input points, scaled by the length scale parameter l. σ^2 is the variance parameter. And both l and σ^2 are trainable hyperparameters.[3]

Figure S11 shows that increasing tessellation passes improves testing performance and reduces overfitting. The difference in ML performance between 3 and 4 passes is less severe than between 2 and 3 tessellation passes. It is, also, clear that the NWChem default of 2 tessellation passes does not suffice for ML applications.

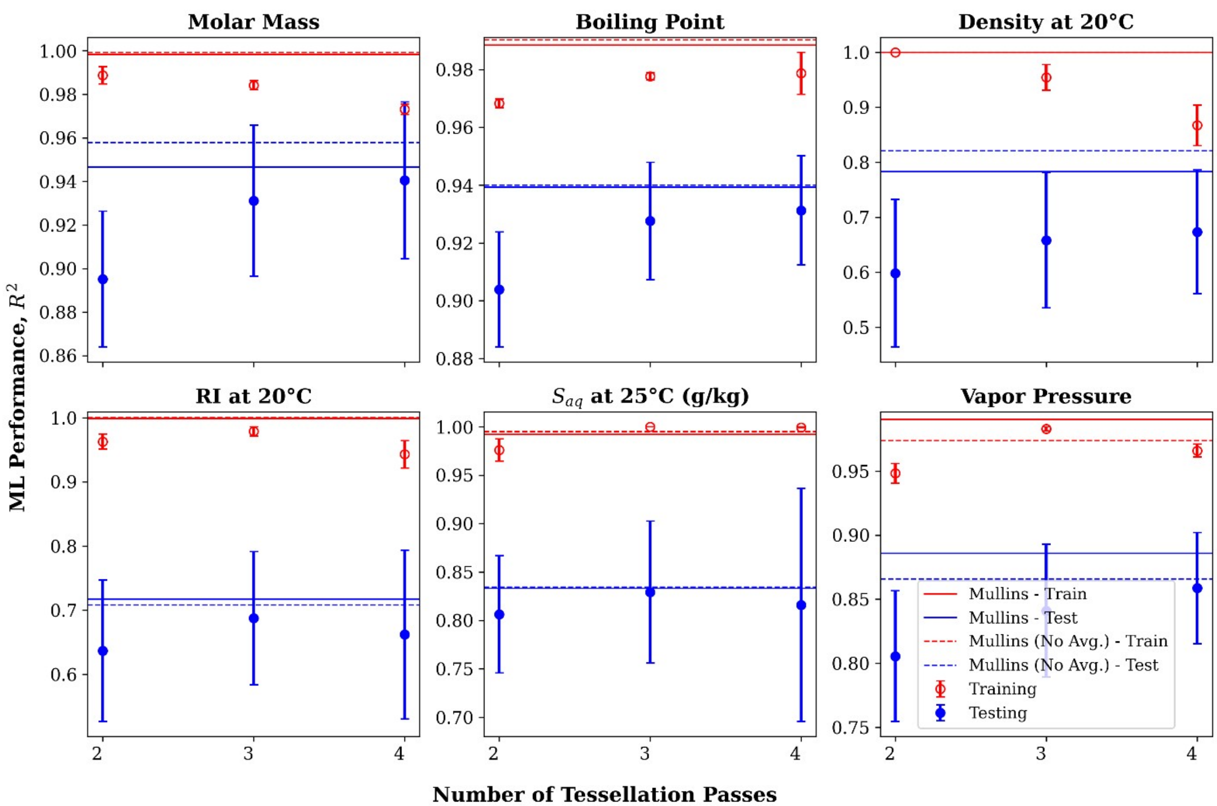

Figure S11: Effect of segment size on GP performance for different target properties. Blue dots indicate performance on the testing set, while hollow red dots indicate performance on the training set. Error bars are the standard deviation in R^2 values for 10 stratified data splits (cross-validation folds). The solid lines indicate the benchmark performance from the Mullins dataset, while dashed lines indicate performance from the unaveraged Mullins dataset. All generated datasets are made up of unaveraged sigma profiles.

Table S2: List of molecules excluded from the effect of segment size comparison

Index	Name	CAS Number	Error Message
50	SQUALANE	111-01-3	Sigma values outside of range...
57	N-TETRADECANE	629-59-4	Sigma values outside of range...

Index	Name	CAS Number	Error Message
62	N-NONADECANE	629-92-5	Sigma values outside of range...
65	N-HENEICOSANE	629-94-7	Sigma values outside of range...
66	N-DOCOSANE	629-97-0	Sigma values outside of range...
67	N-TRICOSANE	638-67-5	Sigma values outside of range...
69	N-PENTACOSANE	629-99-2	Sigma values outside of range...
71	N-HEPTACOSANE	593-49-7	Sigma values outside of range...
220	2,5-DIMETHYL-2,4-HEXADIENE	764-13-6	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
234	1-OCTYNE	629-05-0	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
294	N-OCTYLBENZENE	2189-60-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
332	1-ETHYLNAPHTHALENE	1127-76-0	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
334	2,6-DIMETHYLNAPHTHALENE	581-42-0	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
340	2,7-DIMETHYLNAPHTHALENE	582-16-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
351	CIS-STILBENE	645-49-8	Sigma values outside of range...
412	1-HEXANAL	66-25-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
414	1-NONANAL	124-19-6	Sigma values outside of range...
421	1-DECANAL	112-31-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
423	1-DODECANAL	112-54-9	Sigma values outside of range...
441	METHYL-ISOBUTYL-KETONE	108-10-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
490	1-HEXANOL	111-27-3	Sigma values outside of range...
500	1-HEPTANOL	111-70-6	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
504	1-OCTANOL	111-87-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
505	2-OCTANOL	123-96-6	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
509	1-UNDECANOL	112-42-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
511	DODECANOL	112-53-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.

Index	Name	CAS Number	Error Message
512	1-TRIDECANOL	112-70-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
516	HEPTADECANOL	1454-85-9	Sigma values outside of range...
517	1-OCTADECANOL	112-92-5	Sigma values outside of range...
519	1-EICOSANOL	629-96-9	Sigma values outside of range...
520	1-NONADECANOL	1454-84-8	Sigma values outside of range...
574	1,6-HEXANEDIOL	629-11-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
600	N-TRIDECANOIC-ACID	638-53-9	Sigma values outside of range...
601	N-TETRADECANOIC-ACID	544-63-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
605	STEARIC-ACID	57-11-4	Sigma values outside of range...
649	ETHYL-PROPIONATE	105-37-3	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
651	N-BUTYL-PROPIONATE	590-01-2	Sigma values outside of range...
662	N-BUTYL-VALERATE	591-68-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
665	ETHYL-METHACRYLATE	97-63-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
666	N-PROPYL-METHACRYLATE	2210-28-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
680	N-OCTYL-ACETATE	112-14-1	Sigma values outside of range...
682	N-DECYL-ACETATE	112-17-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
695	N-BUTYL-STEARATE	123-95-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
726	DINONYL-ETHER	2456-27-1	Sigma values outside of range...
760	VINYL-METHYL-ETHER	107-25-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
762	CUMENE-HYDROPEROXIDE	80-15-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
782	1,4-DICHLORO-TRANS-2-BUTENE	110-57-6	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.

Index	Name	CAS Number	Error Message
877	1-BROMONAPHTHALENE	90-11-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
890	N-PENTYLAMINE	110-58-7	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
893	DIISOBUTYLAMINE	110-96-3	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
895	TETRADECYLAMINE	2016-42-4	Sigma values outside of range...
905	HEXAMETHYLENEDIAMINE	124-09-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
922	QUINOLINE	91-22-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
933	QUINALDINE	91-63-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
936	1-NITROPROPANE	108-03-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
943	1-NITROBUTANE	627-05-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
976	SEC-BUTYL-MERCAPTAN	513-53-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1000	DI-N-OCTYL-SULFIDE	8/6/2690	Sigma values outside of range...
1015	CHLOROACETYL-CHLORIDE	79-04-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1185	PENTADECANOIC-ACID	1002-84-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1191	N-HEPTADECANOIC-ACID	506-12-7	Sigma values outside of range...
1193	N-EICOSANIC-ACID	506-30-9	Sigma values outside of range...
1220	CETYL-METHACRYLATE	2495-27-4	Sigma values outside of range...
1230	DIETHYL-SUCCINATE	123-25-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1231	DIHEXYL-ADIPATE	110-33-8	Sigma values outside of range...
1232	ETHYLIDENE-DIACETATE	542-10-9	Sigma values outside of range...
1233	DIALLYL-MALEATE	999-21-3	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1249	ETHYLAL	462-95-3	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1309	DODECYLAMINE	124-22-1	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.

Index	Name	CAS Number	Error Message
1311	PENTAETHYLENE-HEXAMINE	4067-16-7	Sigma values outside of range...
1315	TETRAETHYLENEPENTAMINE	112-57-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1319	N-BUTYL-ISOCYANATE	111-36-4	Sigma values outside of range...
1344	2,4,6-TRINITROTOLUENE	118-96-7	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1346	3,5-DINITROTOLUENE	618-85-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1356	2,2,2-NITRILOTRIS-ACETONITRILE	7327-60-8	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1364	ISOQUINOLINE	119-65-3	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1376	METHYLGLUTARONITRILE	4553-62-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1389	N-METHYLFORMAMIDE	123-39-7	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1417	4-HYDROXYBUTYRALDEHYDE	25714-71-0	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1424	2,2,4-TM-1,3-PD-MONOISOBUTYNATE	77-68-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1509	DI-2-ETHYLHEXYL-ADIPATE	103-23-1	'/tmp/1041648.1291.hpc/tmp.Q16QwEwFlt/Python/../SP-GivenInitXYZ-Mol_1509/SP-GivenInitXYZ-Mol_1509_0/SP-GivenInitXYZ-Mol_1509_0.cosmo.xyz'
1520	DI-N-OCTYLAMINE	1120-48-5	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1546	2-2-HEXOXYETHOXY-ETHANOL	112-59-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1556	DI-N-BUTYL-SULFATE	625-22-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1609	MALIC-ACID	6915-15-7	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1681	ETHYL-CYANOACETATE	105-56-6	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1693	2-MERCAPTOETHANOL	60-24-2	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.

Index	Name	CAS Number	Error Message
1707	ETHYL-3-ETHOXYPROPIONATE	763-69-9	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1722	2,2,3-TRIMETHYLHEXANE	16747-25-4	NWChem job failed to converge in COSMO solvation medium, but converged in vacuum.
1725	CAPRINITRILE	1975-78-6	Sigma values outside of range...

2 Comments on Reproducibility

Previous works[4] have shown that it is not expected that the same sigma profile is generated by different quantum chemistry software even if the same settings are used. As such, the standard is not set for the presented tool to match other published sigma profiles, however, it should be expected that the tool is internally consistent. This brief section shows the differences between 5 sigma profiles generated with the developed tool for the same molecule with the same starting geometry file, and the effect of those differences on the performance of pre-trained GP models.

Figure S12 below shows 5 sigma profiles of n-butanol generated from the same starting structure and the difference between them. The difference between the different runs is not visible when the entire SP is plotted, so the per-bin differences between SPs are plotted just below the SPs. The difference in SP per-bin appears to be on the scale of 10^{-5} $Å^2$, but the significance of these differences should be judged based on their effect on ML performance. And they seem to have no effect on ML performance, as all predictions for all properties seem to only depend on the GP model or the fold/data split it was trained on.

From this brief illustration one can conclude that the sigma profiles obtained using this tool are internally consistent.

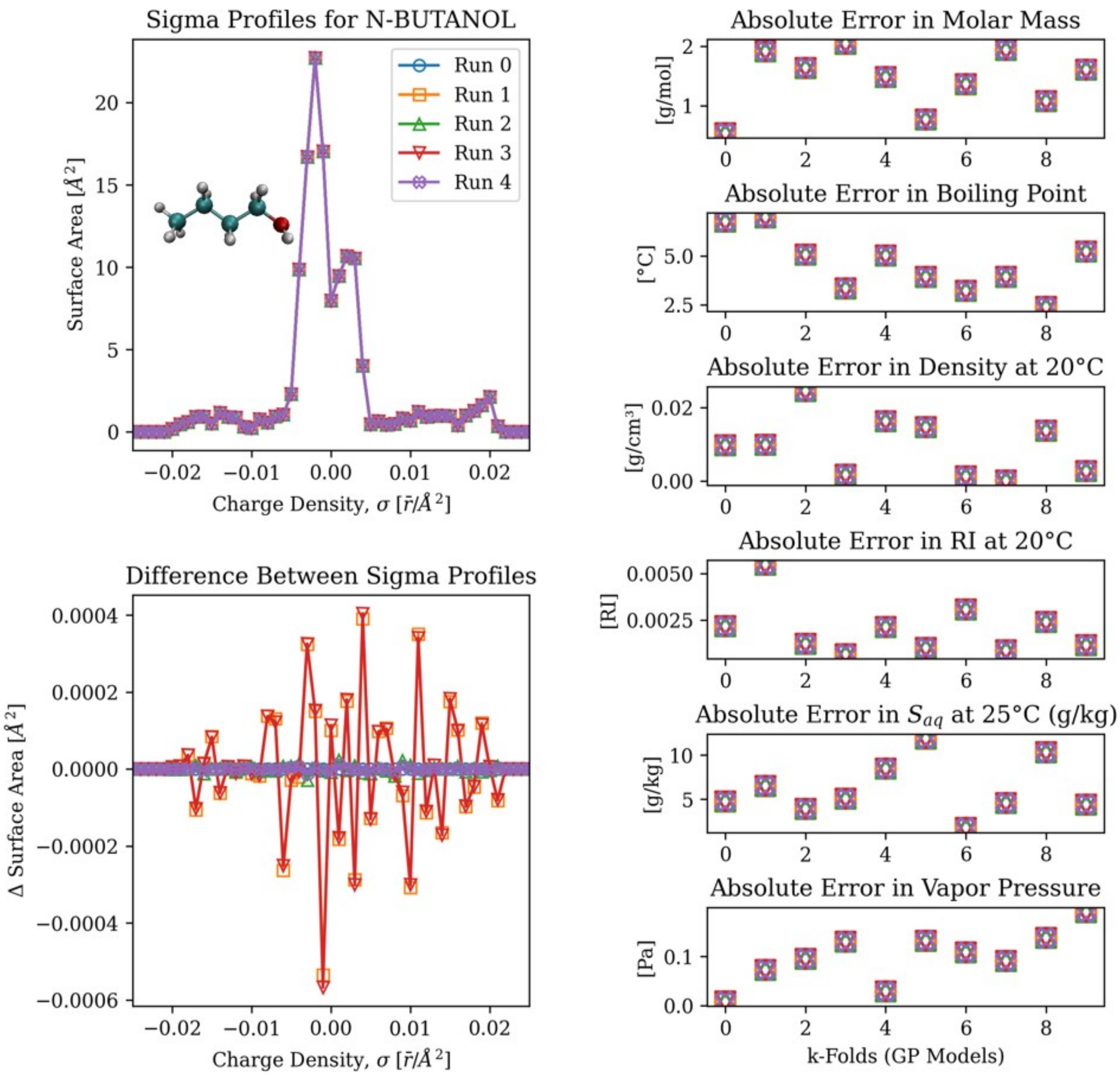

Figure S12: Reproducibility of sigma profiles from the same starting geometry file. Top left shows 5 sigma profiles for N-butanol generated from the same starting geometry (using HF-YK level of theory) with the difference between additional runs and run 0 below it (bottom left). Figures on the right show the absolute error in different target properties when the 5 generated sigma profiles are used as input to the pre-trained GP models for the HF-YK dataset. The colors and shapes of the markers match those of the 5 generated sigma profiles.

References

(1) Klamt, A.; Schuurmann, G. *COSMO: A New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient*; 1993.

(2) Mullins, E.; Liu, Y. A.; Ghaderi, A.; Fast, S. D. Sigma Profile Database for Predicting Solid Solubility in Pure and Mixed Solvent Mixtures for Organic Pharmacological Compounds

with COSMO-Based Thermodynamic Methods. *Ind Eng Chem Res* **2008**, *47* (5), 1707–1725. https://doi.org/10.1021/ie0711022.

(3) Matthews, A. G. de G.; van der Wilk, M.; Nickson, T.; Fujii, K.; Boukouvalas, A.; León-Villagrá, P.; Ghahramani, Z.; Hensman, J. GPflow: A Gaussian Process Library Using TensorFlow. *Journal of Machine Learning Research* **2017**, *18* (40), 1–6.

(4) Bell, I. H.; Mickoleit, E.; Hsieh, C. M.; Lin, S. T.; Vrabec, J.; Breitkopf, C.; Jäger, A. A Benchmark Open-Source Implementation of COSMO-SAC. *J Chem Theory Comput* **2020**, *16* (4), 2635–2646. https://doi.org/10.1021/acs.jctc.9b01016.