Supporting Information: Outlier-Detection for Reactive Machine Learned Potential Energy Surfaces

Luis Itza Vazquez-Salazar,[*,†,‡] Silvan Käser,[*,†,‡] and Markus Meuwly[*,†]

†*Department of Chemistry, University of Basel, Klingelbergstrasse 80 , CH-4056 Basel, Switzerland.*

‡*These authors contributed equally*

E-mail: luisitza.vazquezsalazar@unibas.ch; silvan.kaeser@unibas.ch; m.meuwly@unibas.ch

February 26, 2024

1 Supplementary methods

1.1 Details for DER Multidimensional

For DER-M the outputs are constructed to be part of the covariance matrix $\mathbf{L}$ defined as

$$(\mathbf{L})_{ij} = \begin{cases} \text{SoftPlus}(\ell_i) + \epsilon & \text{If } i = j \\ l_{ij} + \epsilon & \text{If } i > j \\ 0 & \text{else} \end{cases}$$

Here, ℓ_{ij} are the outputs of the last layer $(E_{\text{pred}}, Q_{\text{pred}})$ of the modified PhysNet model. It must be mentioned that $\mathbf{L}$ is a lower triangular matrix. A difference between the original formulation of Meinert and Lavin[S1] and the one presented here is that the exponential function for the covariance matrix is replaced with the SoftPlus activation. Additionally, $\epsilon = 1 \times 10^{-6}$ is added to each of the outputs of the last layer as a regularizer. These modifications avoid numerical instabilities and/or singularities during training.

The parameter ν corresponds to the number of degrees of freedom of the distribution,[S2] and it is also an output of the PhysNet model. Meinert and Lavin[S1] relate ν to the number of virtual measurements of the variance. The value of ν is constrained to $\nu \in [3, 13]$; the lower boundary corresponds to the requirement that $\nu > n + 1$, where n is the number of predicted quantities. The upper boundary is $\nu < 13$ because it is empirically known that for $\nu \geq 13$ the resulting distribution is indistinguishable from a normal distribution.[S3] Then, the expression for ν is:

$$\nu = 10 \left(\frac{\tanh(x) + 1}{2} \right) + 3$$

The aleatoric (data) and epistemic (knowledge) uncertainty of the multidimensional model

are obtained from

$$\mathbb{E}[\sigma^2] = \frac{\nu}{\nu - 3}\mathbf{L}\mathbf{L}^\top \tag{S1}$$

$$Var[\mu] = \frac{\mathbb{E}[\sigma^2]}{\nu} \tag{S2}$$

1.2 Set up of the NN training

The neural network model used in this work is PhysNet.[S4] The original version in tensorflow was used for the ensemble method, while the Pytorch version was employed for DER. Five modules were used in both cases, each with two residual atomic modules and three residual interaction modules. The output of it was pooled into one residual output model. The number of radial basis functions was kept at 64, and the dimensionality of the feature space was 128. A batch size of 32 and a learning rate of 0.001 were used for training. An exponential learning rate scheduler with a decay factor of 0.1 every 1000 steps and the ADAM optimizer[S5] with a weight decay of 0.1 were employed. An exponential moving average for all the parameters was used to prevent overfitting. A validation step was performed every five epochs.

1.3 Classification

Following the methodology presented by Kahle and Zipoli,[S6] we classified the predictions obtained by the different models to determine if the predicted uncertainty can be used as a reliable estimation of the prediction error. In this case, the following classes were defined:

- True Positive (TP): $\varepsilon_i > \varepsilon^*$ and $\sigma_i > \sigma^*$.
- False Positive (FP): $\varepsilon_i < \varepsilon^*$ and $\sigma_i > \sigma^*$
- True Negative (TN): $\varepsilon_i < \varepsilon^*$ and $\sigma_i < \sigma^*$.
- False Negative (FN): $\varepsilon_i > \varepsilon^*$ and $\sigma_i < \sigma^*$.

As a difference from our previous approach,[S7] we report the results when $\varepsilon^* = \text{MSE}$ (mean squared error) and $\sigma^* = \text{MV}$ (mean-variance) and also for different values of σ^* and ε^* to obtain decision boundaries for the relationship between variance and error. For the different values of σ^* and ε^*, common metrics of the overall performance were evaluated. In this work, we use the true positive rate (R_{TP}) or *sensitivity*. This quantity is defined as:[S8]

$$R_{\text{TP}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{S3}$$

Here, N_{TP} refers to the number of true positives and N_{FN} to the number of false negative samples. A large sensitivity value indicates that the model is unlikely to relate large variance values with small errors (c.f. false negatives).

Complementary to Equation S3 is the positive predictive value (P_{TP}) or *precision*:

$$P_{\text{TP}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}} \tag{S4}$$

where all the previous quantities keep their meaning and N_{FP} is the number of false positives. This quantity relates to how many of the samples predicted with high uncertainty correspond to a large error.

In addition it is desirable to quantify how often the model misclassifies a prediction. This can be measured by the False Positive Rate (FPR), which measures how many samples are classified with large uncertainty but low error. This is defined as:

$$R_{\text{FP}} = \frac{N_{\text{FP}}}{N_{\text{FP}} + N_{\text{TN}}} \tag{S5}$$

The opposite case can be quantified by the False Negative Rate (FNR) defined as:

$$R_{\mathrm{FN}} = \frac{N_{\mathrm{FN}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{S6}$$

1.4 Further Analysis of Structures / Outliers

Structures identified with a large variance present more considerable variations for error and variance than the corresponding structures with the biggest values of error; in the following, we will describe the error and variance for each structure following the enumeration of the samples. Test structure #3881 is related to the largest uncertainty for DER-L; this is only replicated by DER-M and GMM, which also assigns it a large uncertainty. However, the energy prediction is accurate for most of the models evaluated except for DER-S. Next, structure #3886 has the largest uncertainty for DER-M, while none of the other models associates it with a large uncertainty value. Nevertheless, this structure is hard to predict for all of them, with errors between 50 and 20 kcal/mol. Continuing with our analysis, molecule #11467 is discussed. This sample is identified with the largest uncertainty value for the GMM model. Nonetheless, all models, even GMM, perform well in predicting this sample. Structures #23550 and #24576 are identified with the largest variance for the models Ens-6 and Ens-3, respectively. Both structures are similar, with a difference in the orientation of the carbon atom attached to the O-O in the *(syn)*-Criegee complex. Both samples show problems to be predicted by the ensemble models; however, it looks like models based on DER show fewer difficulties. Regarding the predicted uncertainty for #23550 and #24576, GMM assigns it a large uncertainty while the DER models assign it a low uncertainty. Last but not least is sample #28980, which is identified with the largest variance for the DER-S model. This sample is hard to predict for all models, being the hardest for DER-L, which yields the largest error for it. Regarding the uncertainty, it is noticed that for most of the models, with the exception of Ens-6, the predicted uncertainty is low. This analysis clearly shows that the prediction error is comparable for most of the analyzed models. However, detecting this error is not easy, as none of the extreme uncertainty values predicted are re-

lated to the extreme error.

1.5 Evaluating the Multi-Reference Character of a Structure

Determining if a single reference method adequately describes a molecular system is challenging. Therefore, several diagnostic metrics have been proposed to evaluate multireference effects on the system. Among them is the T_1 diagnostic,[S9,S10] which is the Euclidean norm of the single substitution amplitudes vector (t_1) of the closed-shell Couple-Cluster Single Doubles (CCSD) wave function divided by the square root of the number of correlated electrons:

$$T_1 = \frac{\|t_1\|}{\sqrt{N_{\text{corr.elec.}}}} \tag{S7}$$

A single reference method will perform correctly if the value of the T_1 diagnostic[S11] is $T_1 < 0.02$. Complementary, the D_1 diagnostic[S12] is defined as the maximum Euclidean norm of the vectors formed by the product of the matrix $\boldsymbol{S}$ which elements s_1^2 are the single excitation amplitudes of the CCSD wavefunction. Then, the D_i diagnostic is defined as:

$$D_1 = ||\boldsymbol{S}||_2 = \max_{||x||_2=1} (||\boldsymbol{S}\vec{x}||_2) \tag{S8}$$

Here $\boldsymbol{S} \in \mathbb{R}^{o \times v}$ with o and v denoting the number of active occupied and active virtual orbitals. For $D_1 > 0.05$ the molecule is dominated by dynamic correlation.[S11] T_1 and D_1 are suggested to be used together because T_1 represents an average value for the complete molecule, which might fail to indicate problems for small regions of the molecule. In those cases, D_1 can be used as evidence if the molecule has regions that single reference methods can not adequately describe.

In this work, T_1 and D_1 diagnostics were determined for the structures identified with the largest error for each model tested and those with the largest uncertainty value. Then, each

molecule was computed at the CCSD(T)-F12 level of theory with the aug-cc-pVTZ basis function with the MOLPRO suite.[S13] Then, the values of T_1 and D_1 are reported on Table S5 for the molecules with large errors and Table S6 for those with large variance.

1.6 Energy Conservation Simulations

The energy conservation of the models was estimated by running molecular dynamics simulations over the generated potentials using the Atomic Simulation Environment (ASE).[S14] NVE simulations were run using Verlet dynamics. The initial velocities were assigned to follow a Maxwell-Boltzmann distribution at 300 K. The simulation was run from the *(syn)*-Criegee intermediate for 0.5 ns using a time step of 0.1 fs. The energies were saved for every 1000 steps.

2 Supplementary Tables

Table S1: Summary of the statistical metrics of the predictions of energy and forces for the models tested in this work. The first two columns correspond to the values for energies, while the last two columns are the values for forces. Units are kcal/mol for energies and (kcal/mol)·Å^{-1} for forces.

Model	MAE(E)	RMSE(E)	MAE(F)	RMSE(F)
Ens-3	0.44	1.80	1.54	11.98
Ens-6	0.43	1.79	1.48	11.47
DER-S	1.03	2.61	32.06	90.60
DER-L	0.69	2.35	31.79	90.09
DER-M	2.19	5.17	33.55	91.54
GMM	0.47	1.83	1.68	9.73

Table S2: Harmonic frequencies of *(syn)*-Criegee: *Ab initio* MP2 reference values are compared to the frequencies determined on the different PESs.

s-Cri.	MP2 Ref.	Ens-3	Ens-6	DER-S	DER-L	DER-M	GMM
1	224.2	225.9	222.8	251.7	170.8	218.2	223.8
2	304.0	298.8	297.2	337.0	273.3	479.0	300.0
3	481.5	476.2	475.8	440.0	460.6	518.8	475.7
4	698.5	691.6	691.2	679.2	687.9	686.6	691.2
5	745.3	738.2	738.1	710.9	761.6	750.6	738.8
6	939.6	928.3	928.6	919.0	924.2	951.9	927.9
7	996.4	998.7	998.3	1000.7	993.0	1010.8	998.8
8	1031.1	1035.1	1034.8	1018.6	1019.7	1067.4	1035.3
9	1130.3	1132.2	1132.0	1112.8	1118.6	1237.4	1132.4
10	1295.6	1286.9	1287.4	1305.4	1300.1	1328.2	1288.6
11	1397.6	1397.4	1397.4	1379.0	1390.9	1387.2	1397.1
12	1456.6	1451.3	1451.2	1403.4	1450.5	1441.0	1451.1
13	1474.2	1471.3	1471.2	1484.4	1486.9	1494.1	1471.2
14	1514.3	1513.1	1513.5	1541.3	1525.9	1540.5	1514.6
15	3047.8	3044.2	3045.1	3060.4	3030.3	2818.8	3046.7
16	3101.5	3088.9	3090.2	3148.6	3069.4	3085.0	3091.0
17	3207.3	3206.2	3206.5	3171.7	3198.7	3126.4	3210.1
18	3253.2	3255.9	3255.9	3186.7	3301.3	3253.7	3256.9
MAE	-	4.7	4.5	27.3	17.9	46.5	4.4

Table S3: Harmonic frequencies of transition state: *Ab initio* MP2 reference values are compared to the frequencies determined on the different PESs.

TS	MP2 Ref.	Ens-3	Ens-6	DER-S	DER-L	DER-M	GMM
1	518.0	517.4	517.4	494.4	506.0	453.4	517.4
2	533.0	528.5	528.5	541.3	524.2	502.2	528.4
3	745.3	744.9	744.9	715.7	721.0	686.0	744.9
4	770.9	768.6	768.6	765.7	748.2	766.5	768.6
5	857.7	853.7	853.7	845.5	846.0	833.7	853.8
6	969.9	964.0	964.0	929.0	932.1	973.2	964.0
7	1010.3	1007.4	1007.4	992.2	1000.4	1011.2	1007.4
8	1036.7	1033.2	1033.2	1030.7	1042.4	1063.8	1033.3
9	1223.2	1221.3	1221.3	1201.0	1220.8	1184.8	1221.3
10	1281.6	1281.2	1281.2	1272.0	1296.0	1250.9	1281.2
11	1360.3	1360.0	1360.1	1329.5	1382.1	1412.2	1360.0
12	1504.5	1503.3	1503.3	1466.9	1510.7	1555.6	1503.2
13	1557.9	1554.2	1554.2	1542.5	1564.2	1572.7	1554.2
14	1875.3	1866.3	1866.4	1795.1	1805.2	2021.0	1866.1
15	3116.3	3118.9	3118.8	3095.8	3071.0	3124.2	3118.7
16	3237.2	3236.3	3236.3	3215.1	3130.6	3235.1	3236.0
17	3251.9	3252.9	3252.9	3230.5	3159.4	3264.3	3252.8
i	1523.0	1518.3	1518.2	1574.3	1544.7	1331.7	1518.5
MAE	-	2.8	2.8	25.3	28.9	42.3	2.8

Table S4: Harmonic frequencies of VHP: *Ab initio* MP2 reference values are compared to the frequencies determined on the different PESs.

VHP	MP2 Ref.	Ens-3	Ens-6	DER-S	DER-L	DER-M	GMM
1	149.1	178.3	178.7	194.1	176.2	209.8	178.7
2	253.1	254.4	254.4	258	240.6	250.9	254.5
3	332.5	331.8	331.8	338.5	338.5	376.4	331.8
4	612.4	613.0	613.0	622.4	595.6	562.0	613.2
5	711.2	708.7	708.7	668.5	626.0	602.0	708.8
6	843.8	840.7	840.6	797.6	783.9	796.6	840.8
7	878.3	876.1	876.0	878.5	859.2	839.3	876.2
8	972.2	968.2	968.3	909.0	878.1	890.1	968.4
9	975.0	971.7	971.7	994.6	988.4	1030.6	971.9
10	1158.8	1156.3	1156.2	1152.8	1153.6	1130.2	1156.4
11	1319.1	1319.1	1319.1	1340.6	1372.8	1270.9	1319.1
12	1374.2	1372.6	1372.6	1350.4	1388.3	1325.9	1372.7
13	1428.7	1425.4	1425.4	1449.6	1417.4	1464.7	1425.4
14	1693.6	1691.6	1691.6	1704.9	1711.8	1689.2	1691.6
15	3216.3	3222.5	3222.4	3144.9	3178.9	3191.5	3222.3
16	3236.0	3235.3	3235.2	3178.8	3237.1	3299.5	3235.1
17	3330.0	3333.7	3333.8	3313.4	3289.2	3393.3	3333.4
18	3762.9	3759.0	3758.9	3716.0	3765.8	3821.4	3758.8
MAE	-	3.9	4.0	28.5	28.8	48.1	3.9

Table S5: Diagnostics for assessing the multireference character of the structures identified with the largest error in the test dataset. These quantities are unitless. A value of $T_1 > 0.02$ indicates a multireference character, and $D_1 > 0.05$ points to dynamical multireference effects.[S11]

Molecule	T_1	D_1
3429	0.09	0.45
3986	0.05	0.23
28980	0.05	0.24

Table S6: Diagnostic metrics for the multireference character of the structures identified with the largest uncertainty in the test dataset. A value of $T_1 > 0.02$ indicates a multireference character. Complementary, $D_1 > 0.05$ indicates the presence of dynamical multireference effects.

Molecule	T_1	D_1
3881	0.07	0.25
3886	0.08	0.35
23366	0.04	0.19
23550	0.05	0.24
24576	0.07	0.36

3 Supplementary Figures

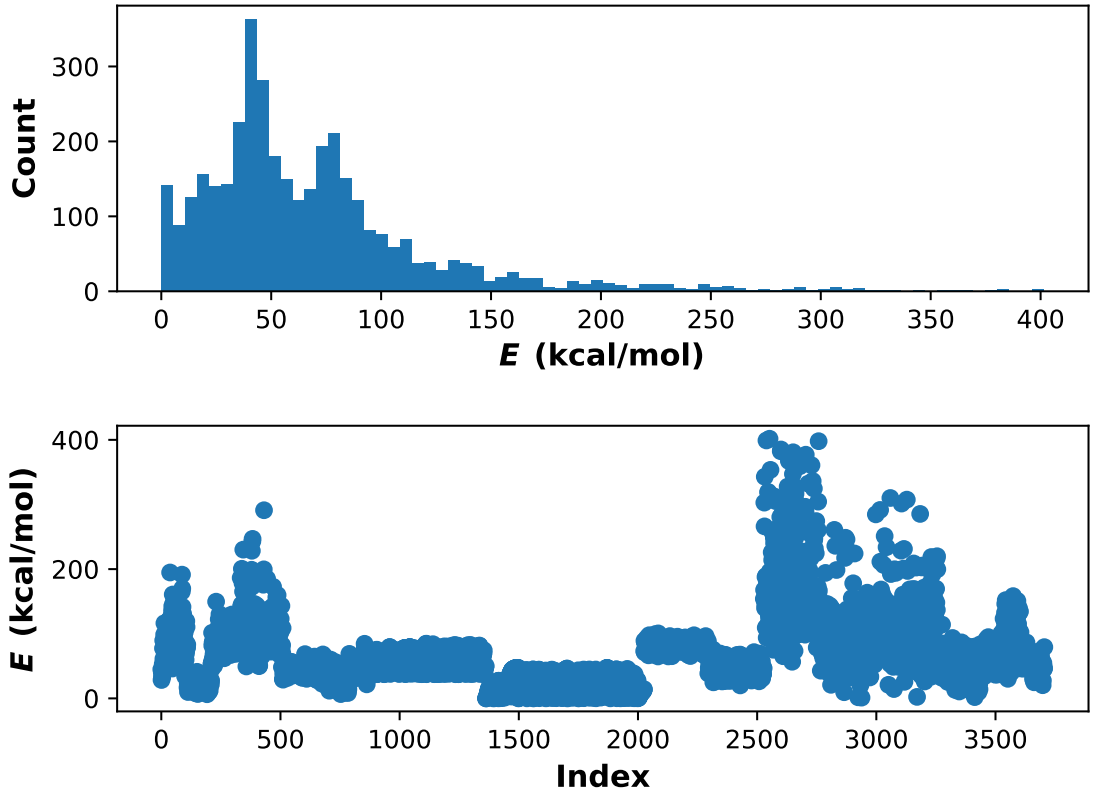

Figure S1: Energy distribution of the data set employed to train the first generation ML-PES.

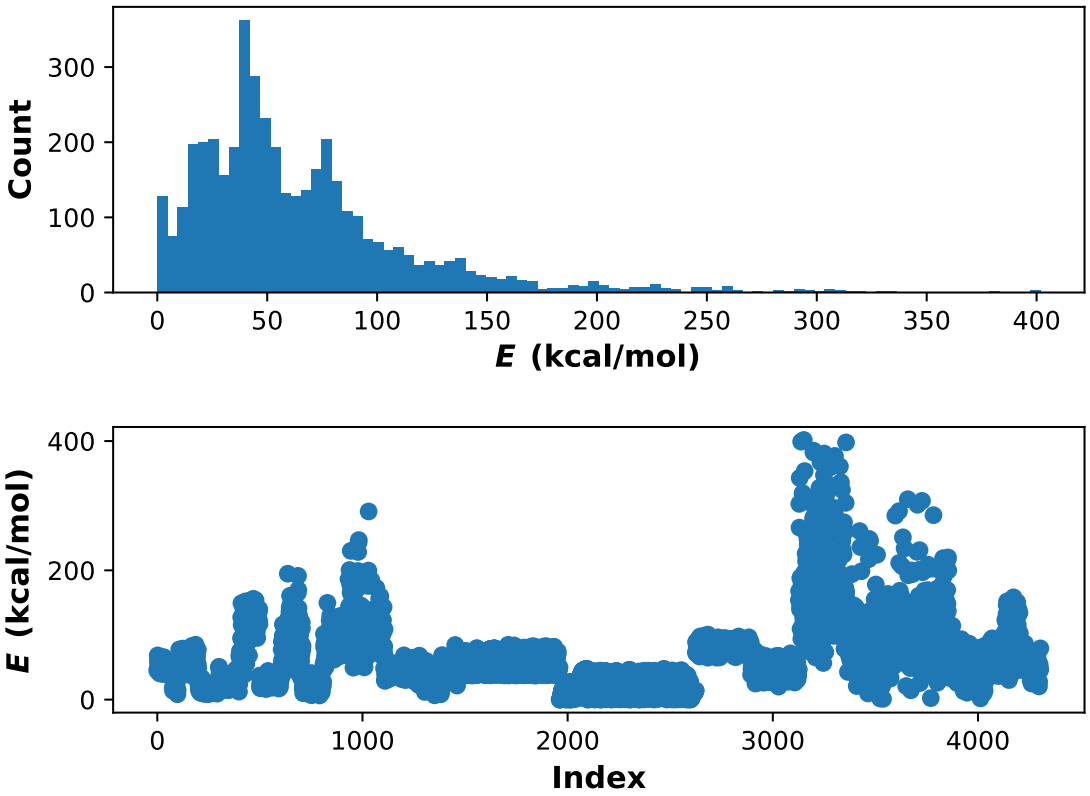

Figure S2: Energy distribution of the data set employed to train the final generation ML-PES.

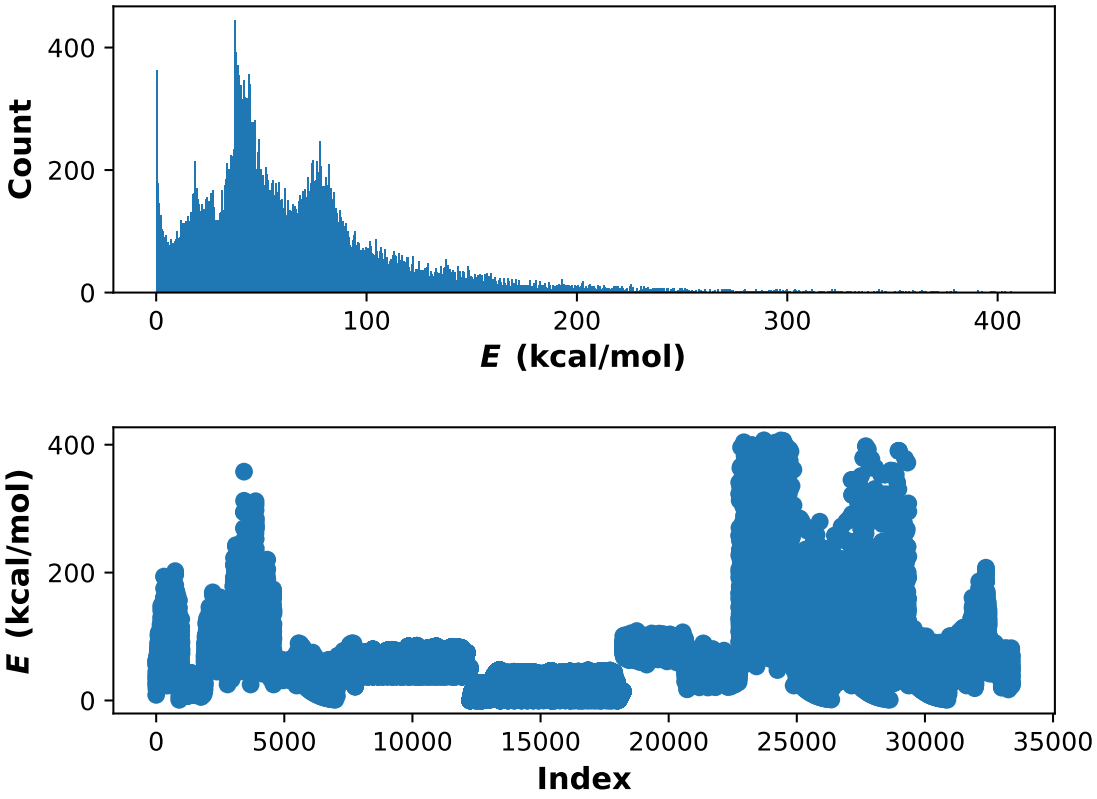

Figure S3: Energy distribution of the test data set.

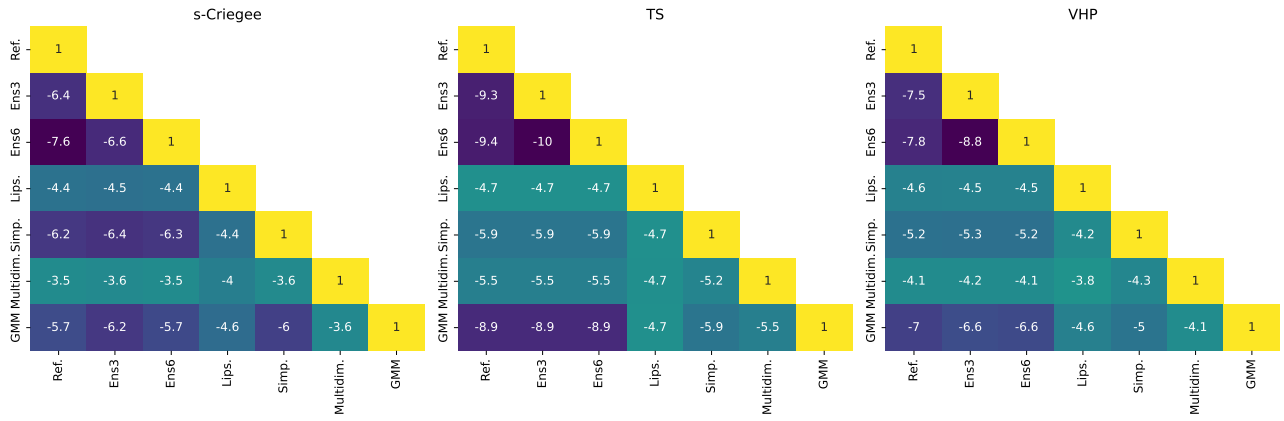

Figure S4: Root Mean Square Displacement of the stationary points (VHP, Transition State and S-Criegee) of the potential energy surface with respect to the *ab-initio* reference structure and between the different obtained geometries. Notice that the logarithm of the value of RMSD is reported to exemplify the differences between the values better.

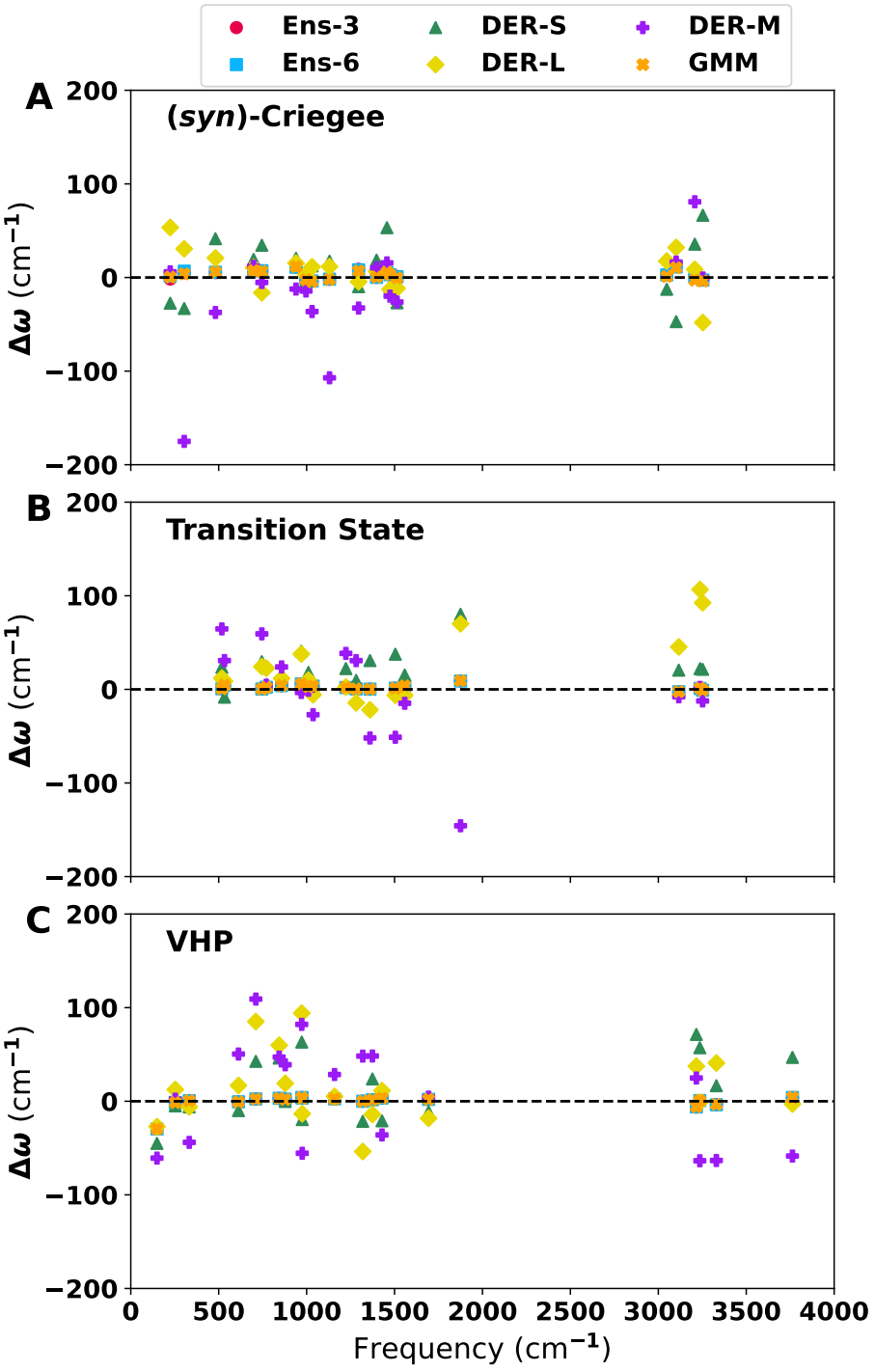

Figure S5: Error per predicted harmonic frequency ($\Delta\omega = \omega_{\text{ref}} - \omega_{\text{pred}}$) of the *(syn)*-Criegee (A), transition state (B) and VHP (C) for all the UQ methods evaluated in this work. The values of the frequencies are reported in Tables S2, S3, and S4.

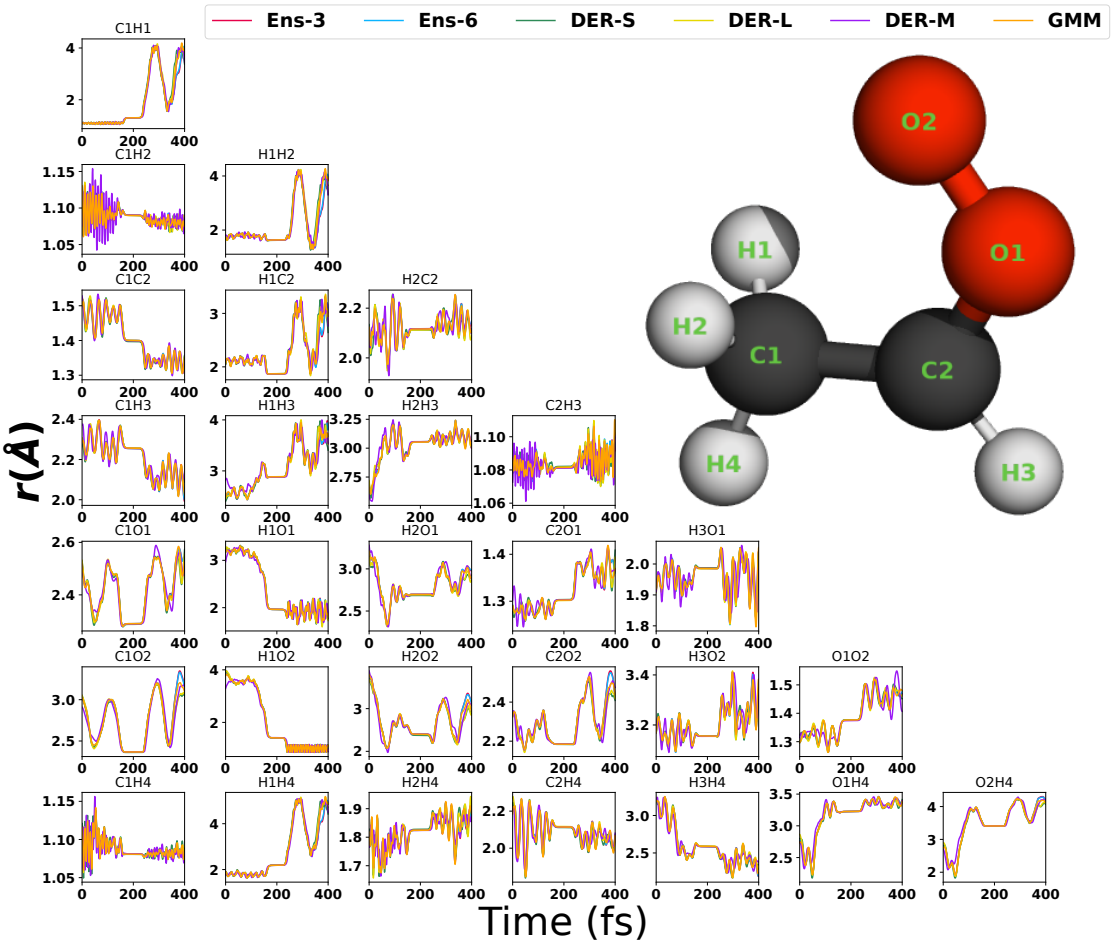

Figure S6: Atom-atom separation time series along the MDP for all models tested in this work. Each panel reports the distance between two atoms. The inset molecule displays the labelling of the atoms.

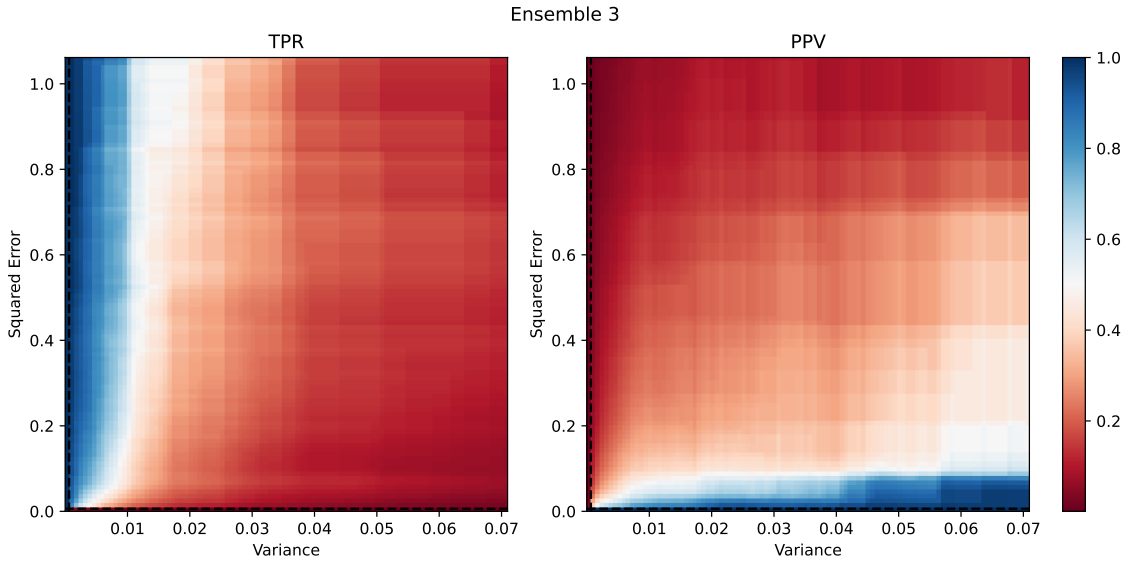

Figure S7: True Positive Rate (Left) and Positive Predictive Value (Right) for the Ens-3 model.

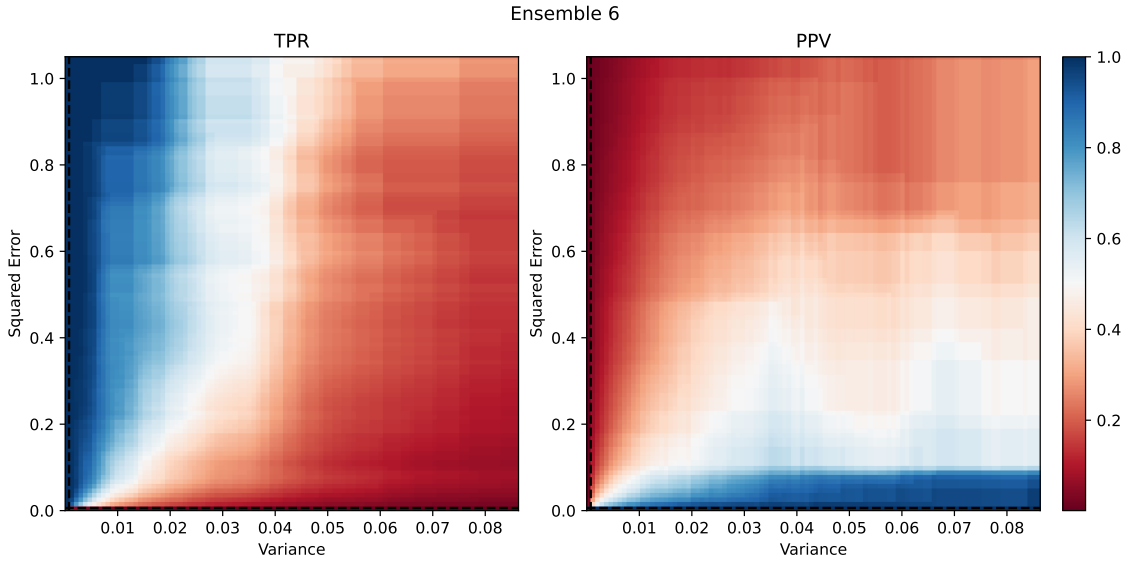

Figure S8: True Positive Rate (left) and Positive Predictive Value (right) for the Ens-6 model.

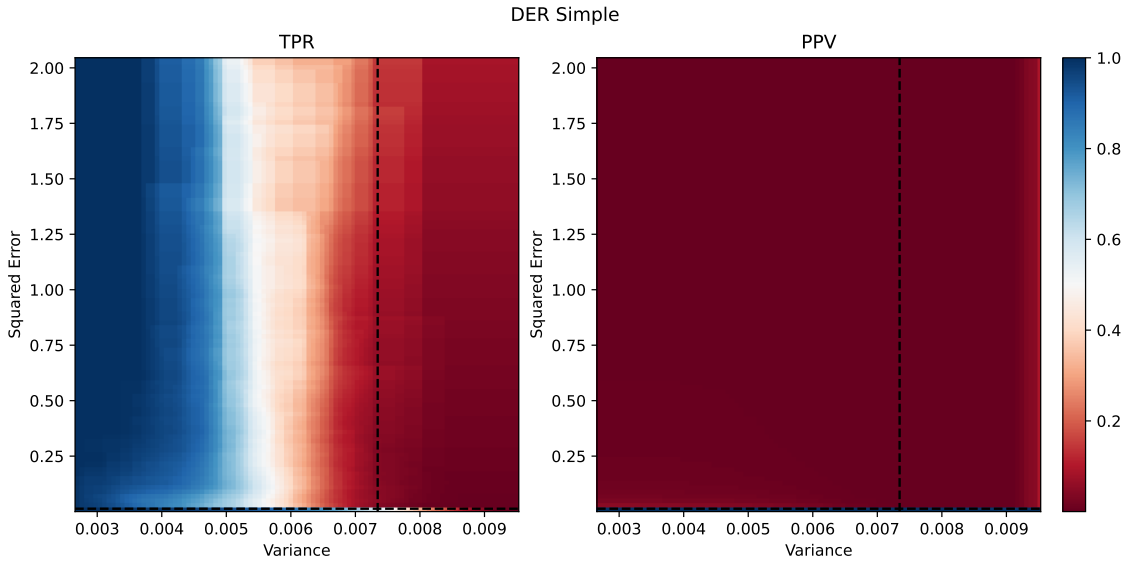

Figure S9: True Positive Rate (left) and Positive Predictive Value (right) for DER-S.

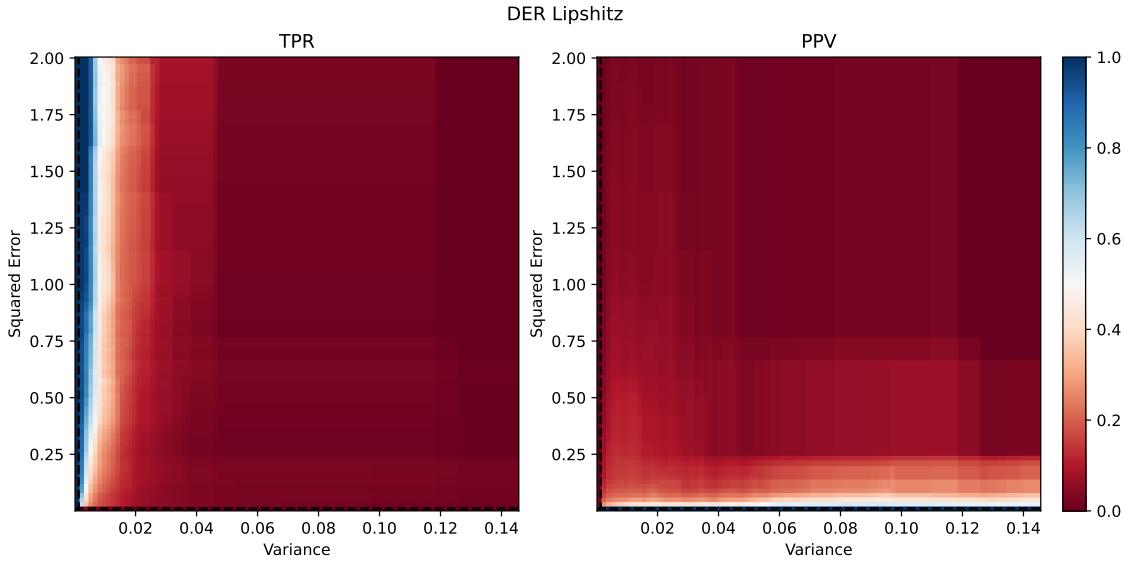

Figure S10: True Positive Rate (left) and Positive Predictive Value (right) for DER-L.

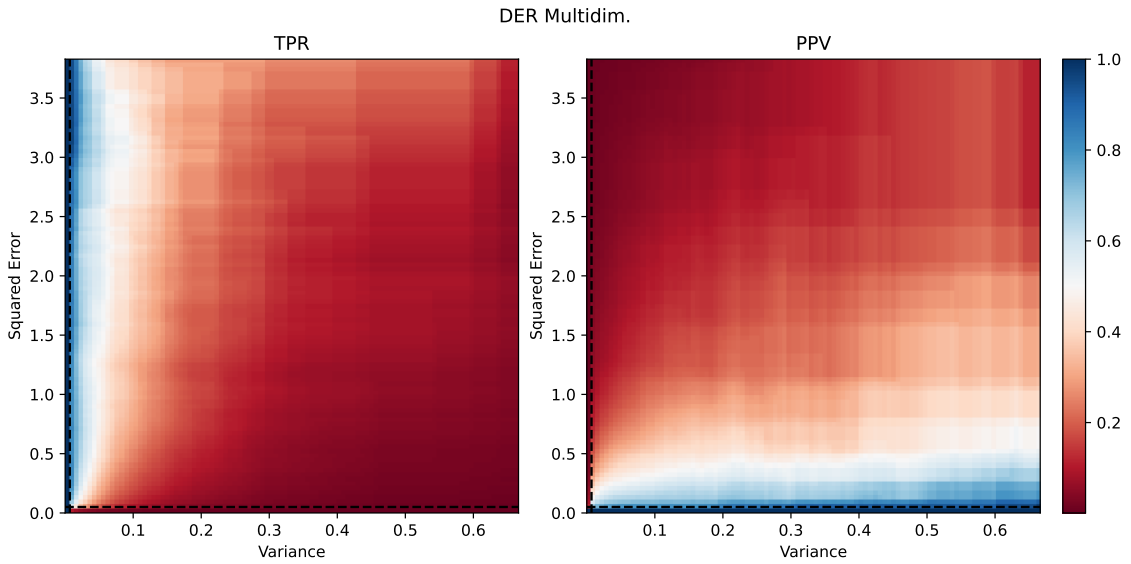

Figure S11: True Positive Rate (left) and Positive Predictive Value (right) for DER-M.

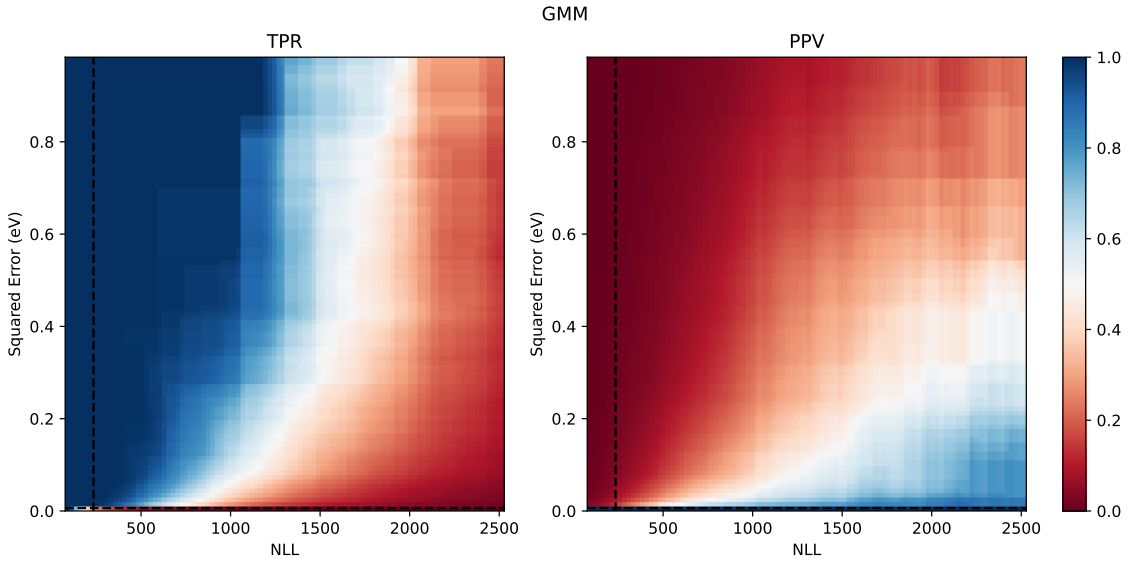

Figure S12: True Positive Rate (left) and Positive Predictive Value (right) for Gaussian Mixture Model.

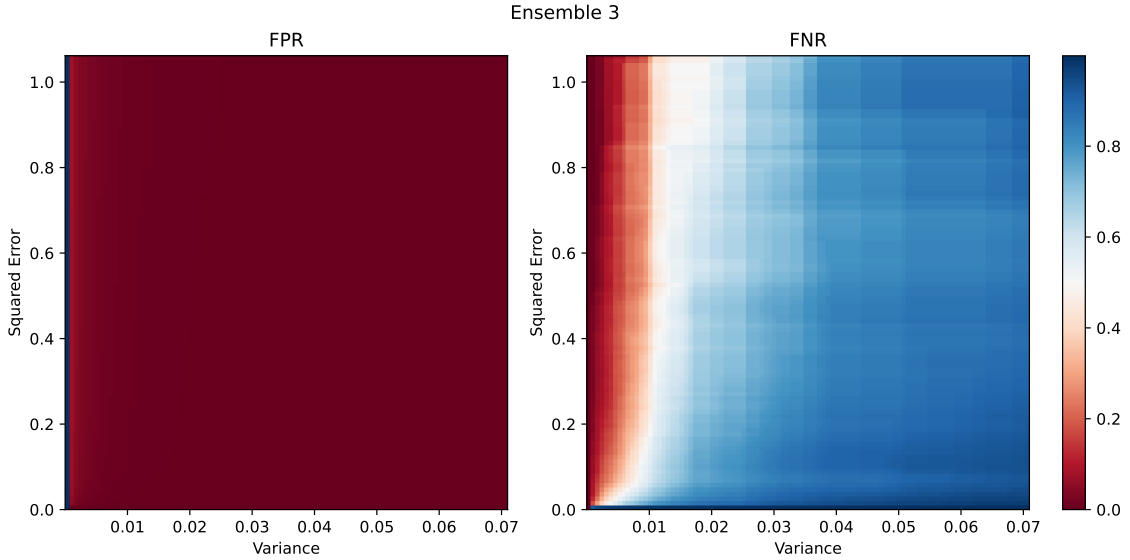

Figure S13: False Positive Rate (left) and False Negative Rate (right) for the Ens-3 model

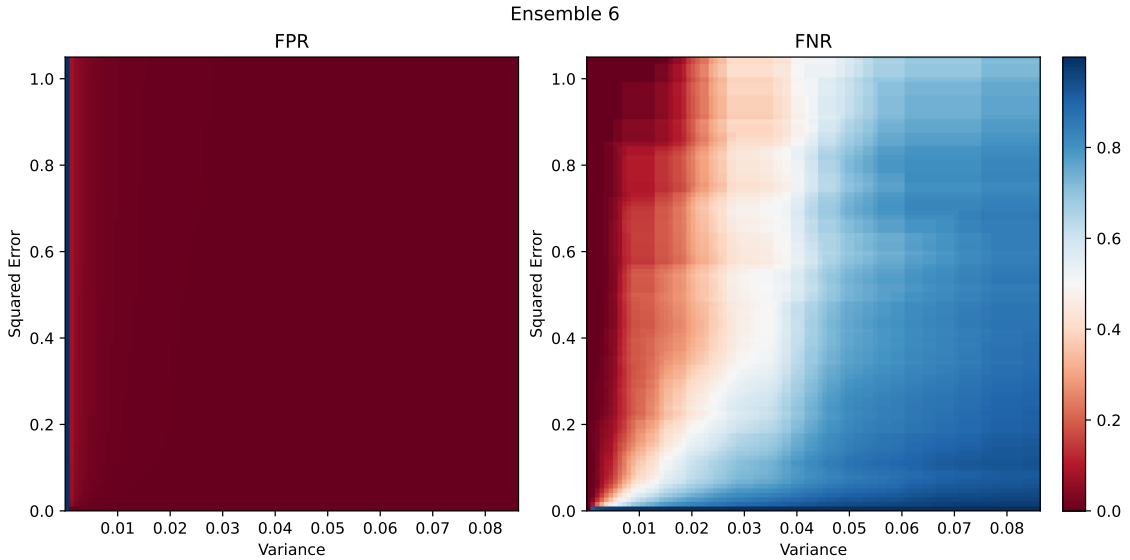

Figure S14: False Positive Rate (left) and False Negative Rate (right) for the Ens-6 model

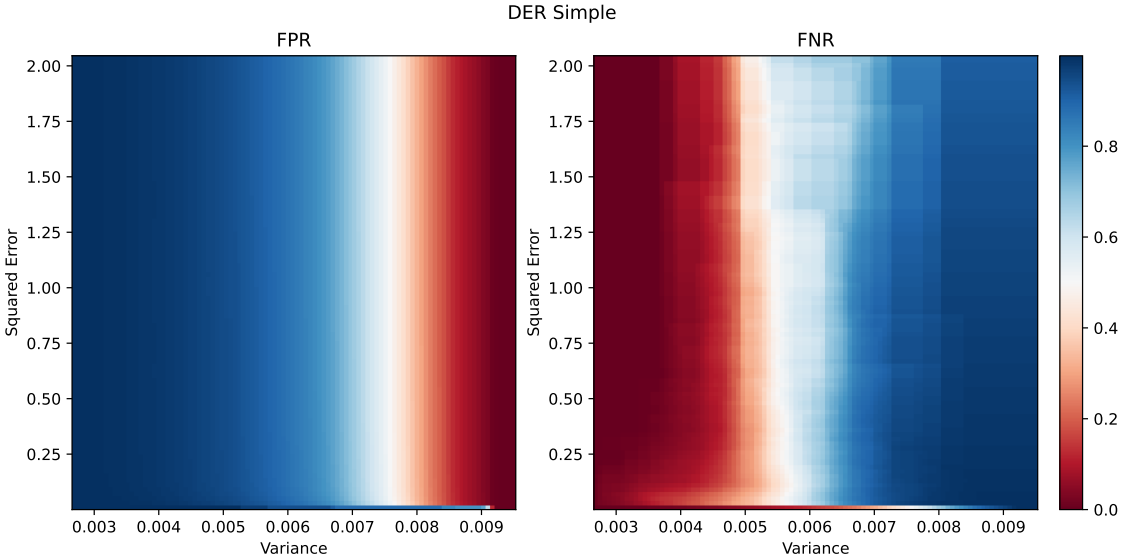

Figure S15: False Positive Rate (left) and False Negative Rate (right) for DER-S

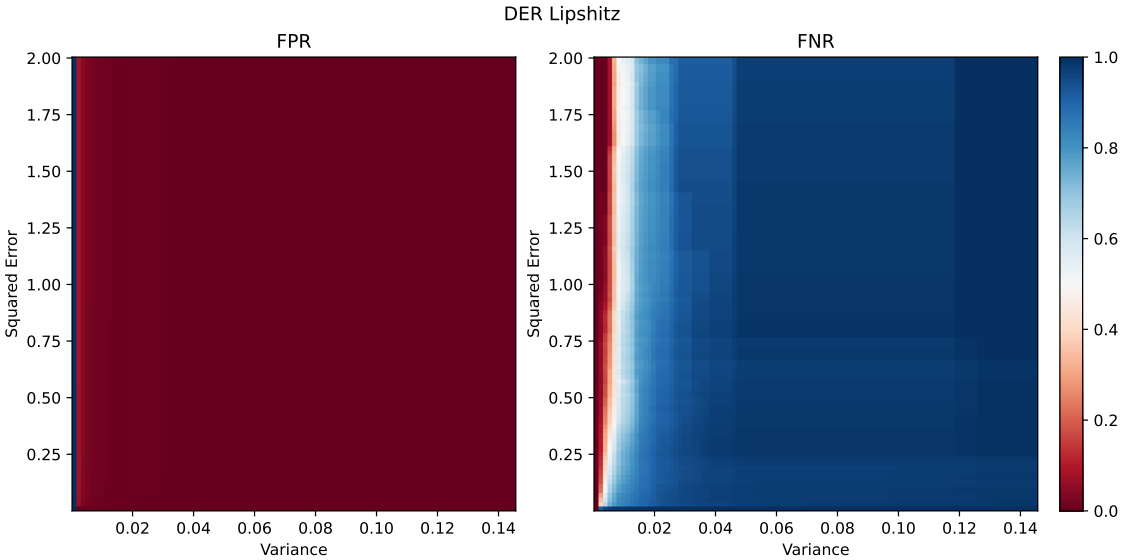

Figure S16: False Positive Rate (left) and False Negative Rate (right) for DER-L

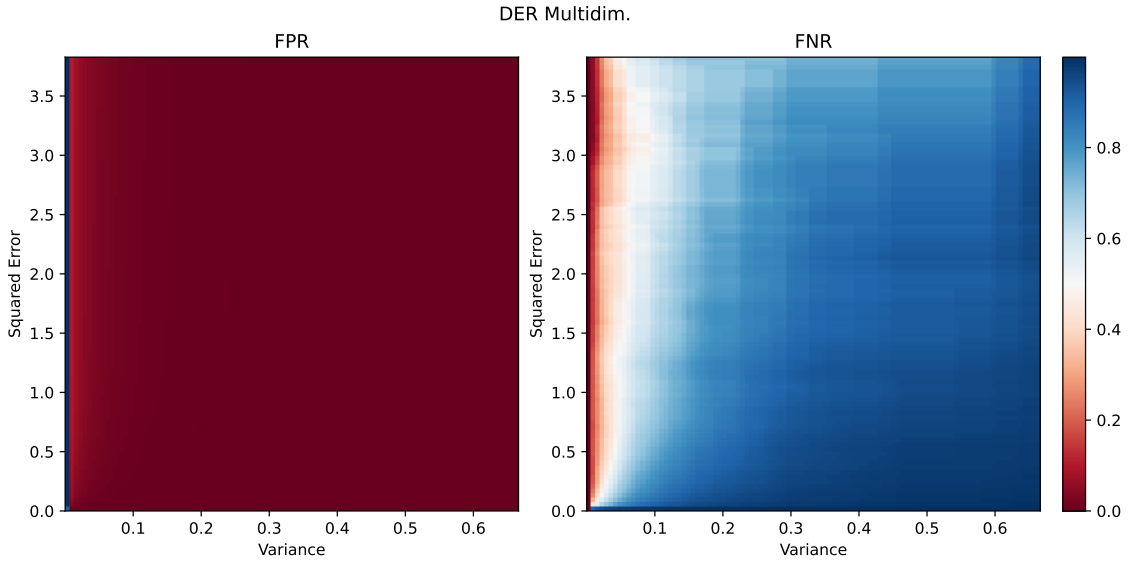

Figure S17: False Positive Rate (left) and False Negative Rate (right) for DER-M

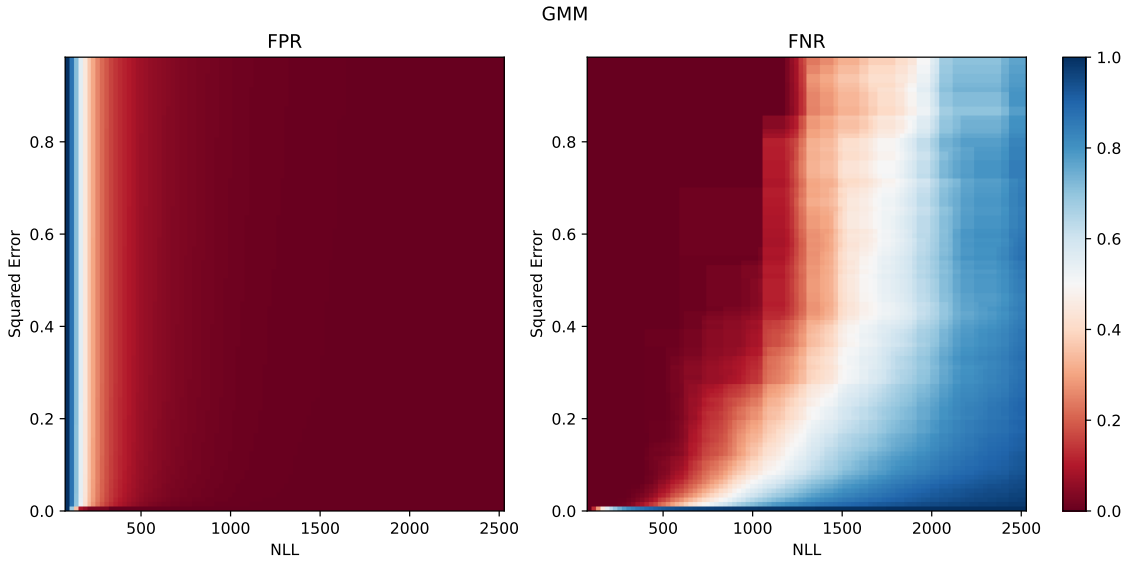

Figure S18: False Positive Rate (left) and False Negative Rate (right) for Gaussian Mixture Model

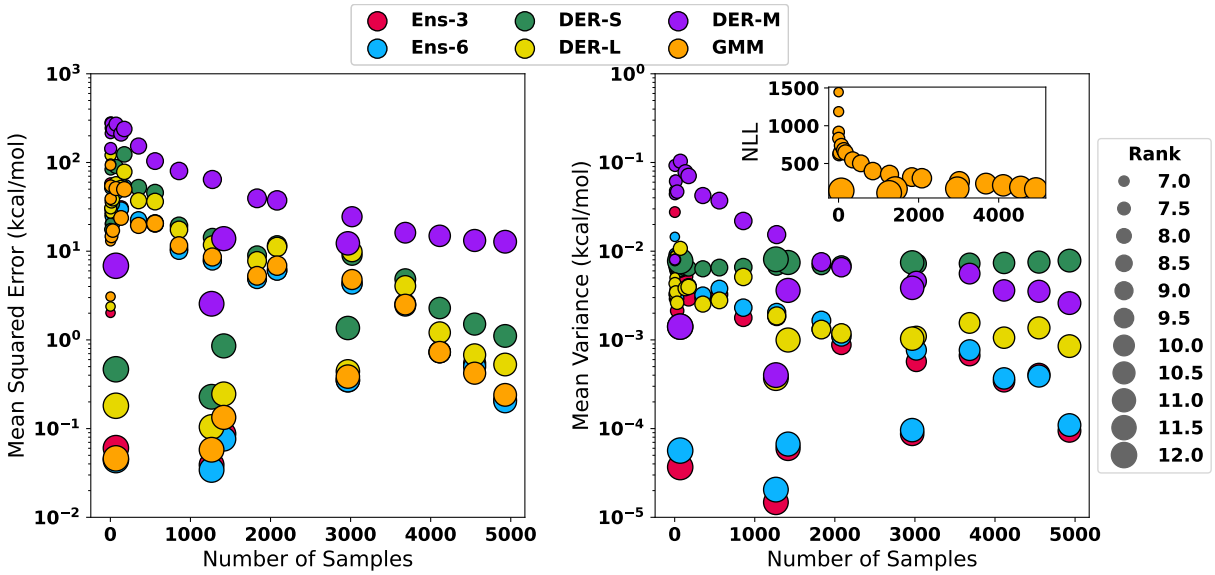

Figure S19: Changes in the mean square error (left) and mean variance (right) with respect to the number of samples in each class. The size of the scatter point is scaled with the ranking number. For the GMM model, the NLL is used to estimate the uncertainty. Notice that the y-axis scale is logarithmic.

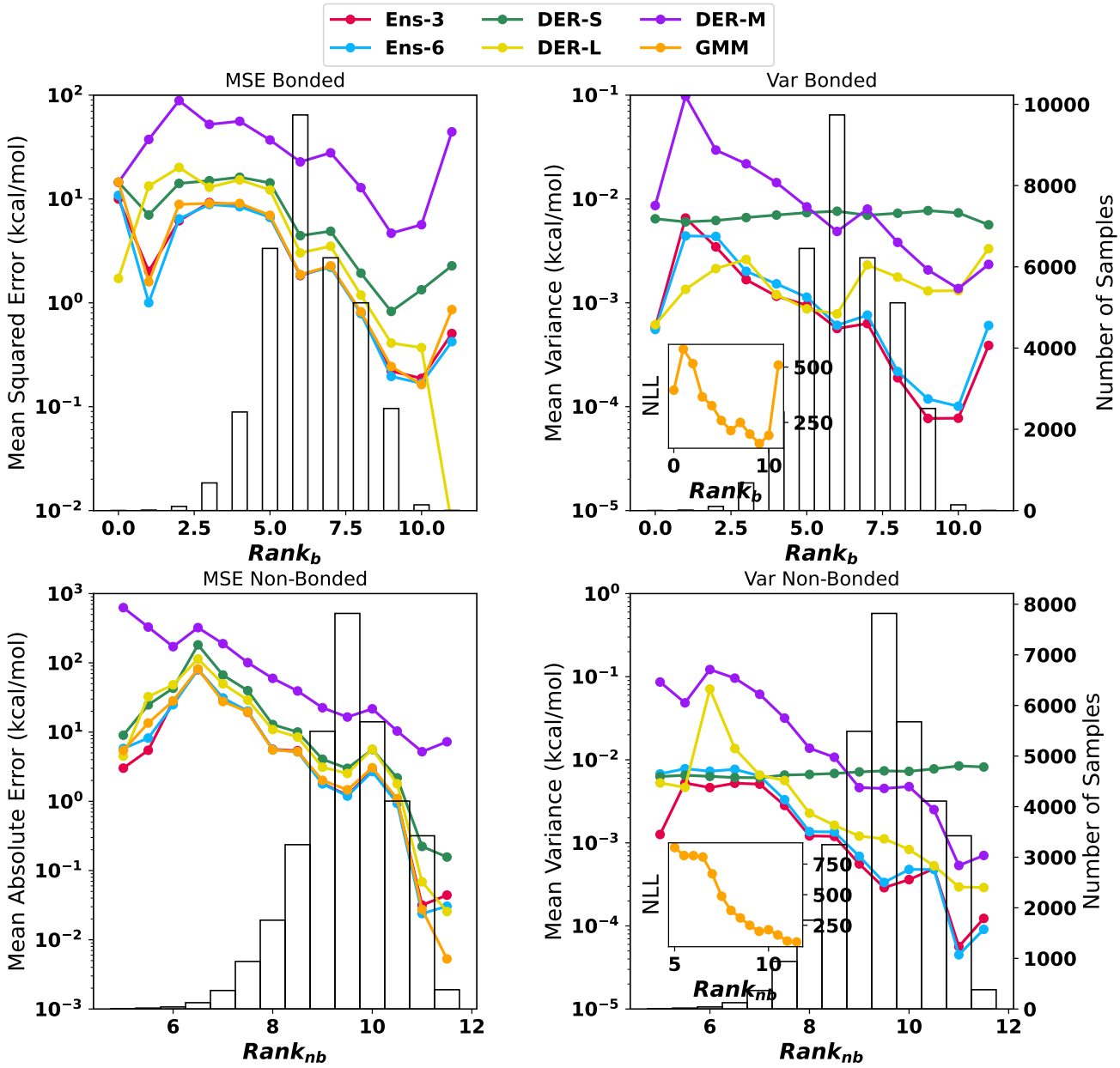

Figure S20: Changes in the mean square error (left) and mean variance (right) with respect to the rank of the molecules in the test set divided by contributions to bond (top) and non-bonded (bottom). In the background, a histogram of the number of samples with the same rank. For the GMM model, the NLL is used to estimate the uncertainty; therefore, the inset shows the changes in the NLL with respect to the rank. Notice that the y-axis scale is on logarithmic units.

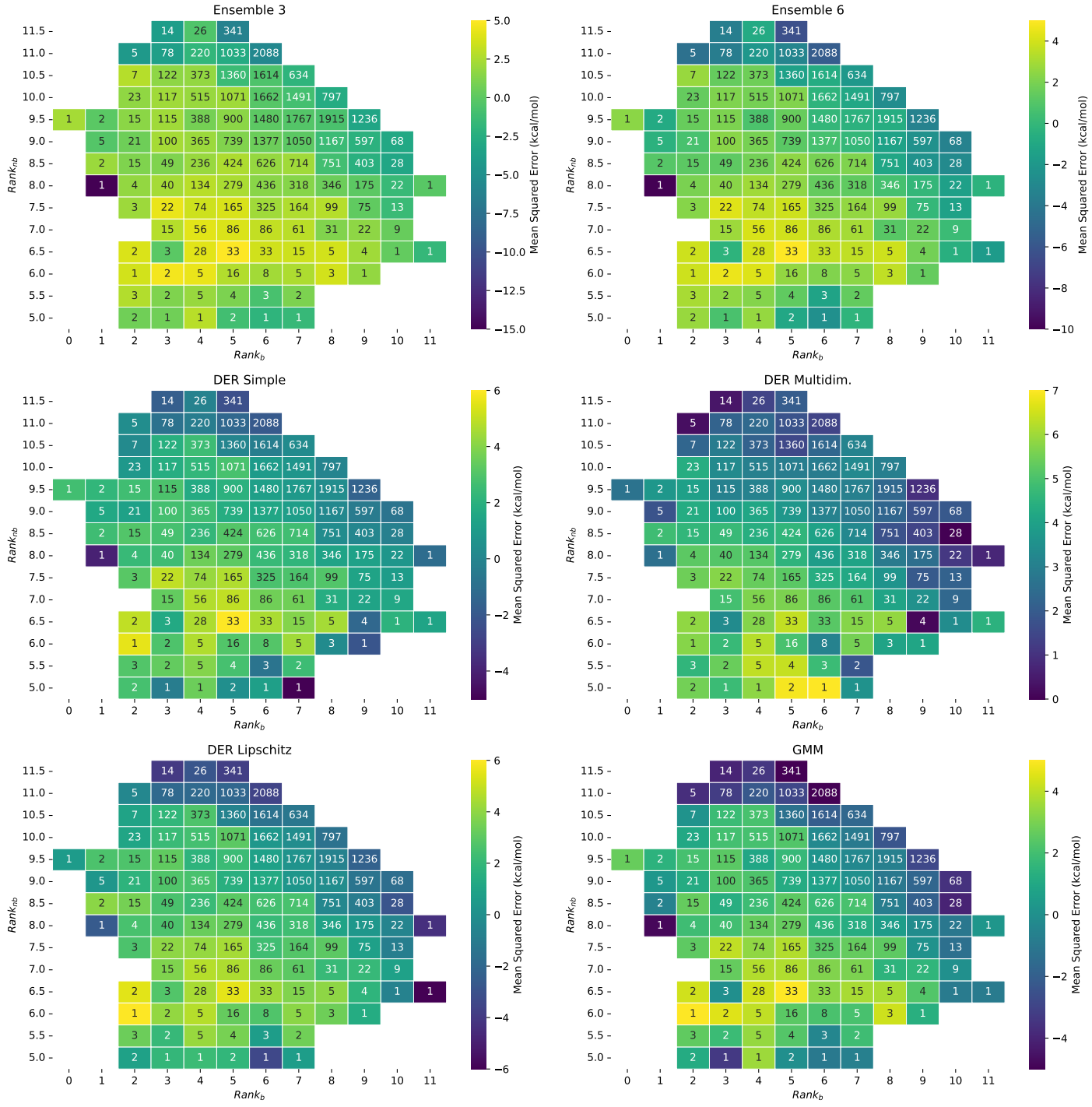

Figure S21: Map of influence of rank values for in and outside distribution of bond and non-bonded distances with respect to the error. The colour bar indicates the logarithm of the Mean Square Error and is normalised to its minimum and maximum values. The numbers inside each box are the number of samples for that score. The box is empty if no samples were found with that combination.

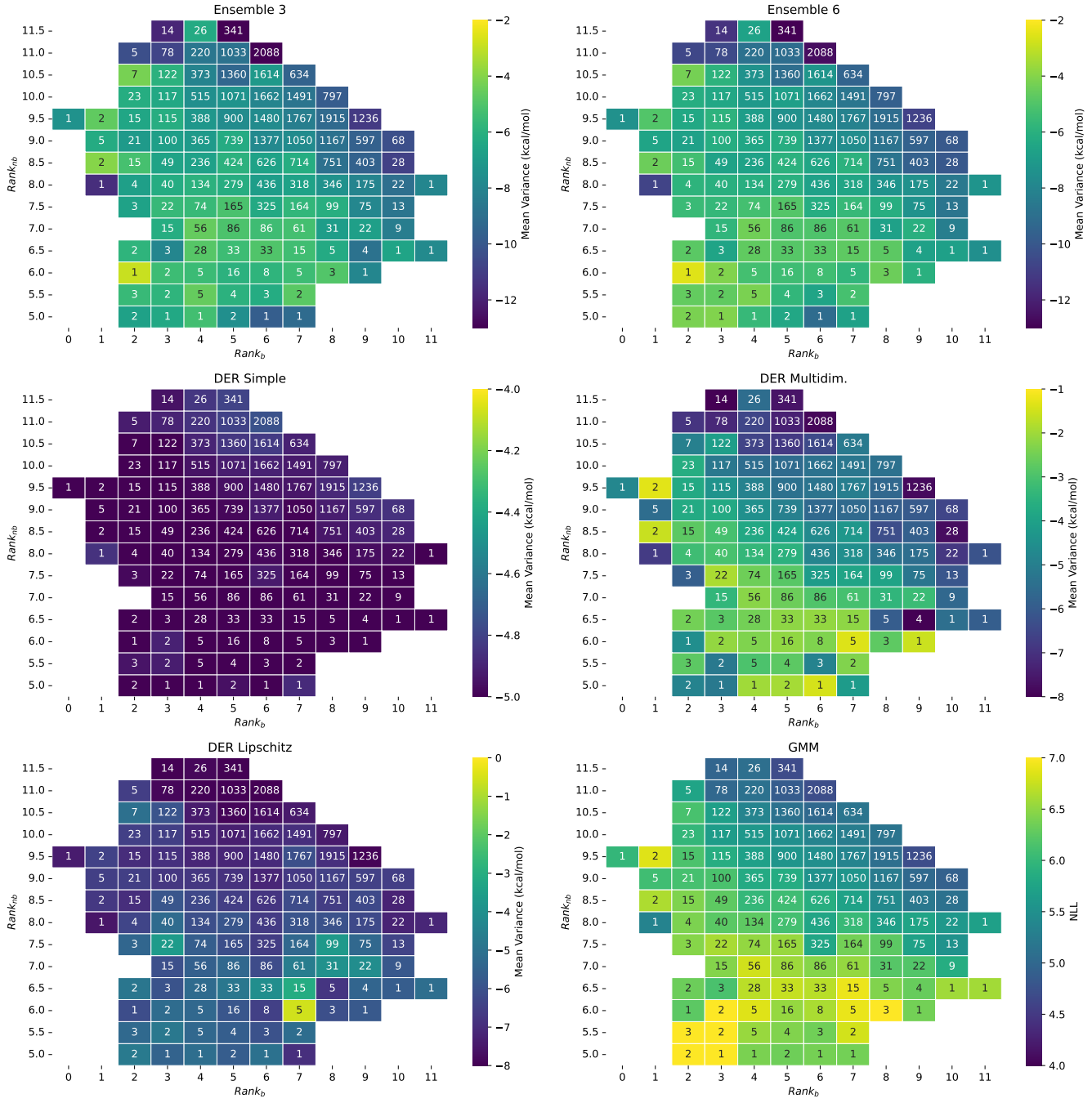

Figure S22: Map of influence of rank values for in and outside distribution of bond and non-bonded distances with respect to the variance. The colour bar indicates the logarithm of the mean variance except for GMM, which shows the NLL and is normalised to its minimum and maximum values. The numbers inside each box are the number of samples for that score. The box is empty if no samples were found with that combination.

References

(S1) Meinert, N.; Lavin, A. Multivariate deep evidential regression. *arXiv preprint arXiv:2104.06135* **2021**,

(S2) Murphy, K. P. *Probabilistic machine learning: Advanced topics*; MIT press, 2023.

(S3) Brereton, R. G. The t-distribution and its relationship to the normal distribution. *J. Chemom.* **2015**, *29*, 481–483.

(S4) Unke, O. T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **2019**, *15*, 3678–3693.

(S5) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(S6) Kahle, L.; Zipoli, F. Quality of uncertainty estimates from neural network potential ensembles. *Phys. Rev. E* **2022**, *105*, 015311.

(S7) Vazquez-Salazar, L. I.; Boittier, E. D.; Meuwly, M. Uncertainty quantification for predictions of atomistic neural networks. *Chem. Sci.* **2022**, *13*, 13068–13084.

(S8) Watt, J.; Borhani, R.; Katsaggelos, A. K. *Machine learning refined: Foundations, algorithms, and applications*; Cambridge University Press, 2020.

(S9) Lee, T. J.; Rice, J. E.; Scuseria, G. E.; Schaefer, H. F. Theoretical investigations of molecules composed only of fluorine, oxygen and nitrogen: determination of the equilibrium structures of FOOF,$(NO)_2$ and FNNF and the transition state structure for FNNF cis-trans isomerization. *Theor. Chim. Acta* **1989**, *75*, 81–98.

(S10) Lee, T. J.; Taylor, P. R. A diagnostic for determining the quality of single-reference electron correlation methods. *Int. J. Quantum Chem.* **1989**, *36*, 199–207.

(S11) Wang, J.; Manivasagam, S.; Wilson, A. K. Multireference character for 4d transition metal-containing molecules. *J. Chem. Theory Comput.* **2015**, *11*, 5865–5872.

(S12) Janssen, C. L.; Nielsen, I. M. New diagnostics for coupled-cluster and Møller–Plesset perturbation theory. *Chem. Phys. Lett.* **1998**, *290*, 423–430.

(S13) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M.; Celani, P.; Györffy, W.; Kats, D.; Korona, T.; Lindh, R.; Mitrushenkov, A.; Rauhut, G.; Shamasundar, K. R.; Adler, T. B.; Amos, R. D.; Bennie, S. J.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hesselmann, A.; Hetzer, G.; Hrenar, T.; Jansen, G.; Köppl, C.; Lee, S. J. R.; Liu, Y.; Lloyd, A. W.; Ma, Q.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Miller III, T. F.; Mura, M. E.; Nicklass, A.; O'Neill, D. P.; Palmieri, P.; Peng, D.; Pflüger, K.; Pitzer, R.; Reiher, M.; Shiozaki, T.; Stoll, H.; Stone, A. J.; Tarroni, R.; Thorsteinsson, T.; Wang, M.; Welborn, M. MOLPRO, version 2019, a package of ab initio programs. 2019.

(S14) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C., et al. The atomic simulation environment – a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.