Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting information for "Machine learning the quantum flux-flux correlation function for catalytic surface reactions"

Brenden G. Pelkie[*,1], Stéphanie Valleau[*,2]

* Department of Chemical Engineering, University of Washington, Seattle, Washington 98115, United States

Contents

[1] email: bgpelkie@uw.edu
[2] email: valleau@uw.edu

1. Minimum energy paths

Minimum energy paths were obtained from the Catalysis-hub energies of the reactants, products and transition states by fitting those values to single asymmetric Eckart barriers, $V_{\text{eckart}}(x)$ (Equation S1) or a function composed of the product of a logistic function and skewed normal function (Equation S2). The choice of the fitting function was determined based on the error of the fit.

$$V_{\text{eckart}}(x) = \frac{V_1(1-\alpha)}{1+e^{-\frac{2\pi x}{w_1}}} + \frac{V_1\left(1+\sqrt{\alpha}\right)^2}{4\cosh^2\left(\frac{\pi x}{w_1}\right)} \tag{S1}$$

The Eckart barrier function depends on 3 parameters: V_1 is the height of the barrier, w_1 the width of the barrier, and α the asymmetry of the barrier.

$$V_{\text{skewnormal}}(x) = v * \underbrace{\left[\frac{L}{1+e^{-k(x-x_0)}}\right]}_{Logistic\ function} * \underbrace{\left[2 * \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right] * \frac{1}{2}\left[1+\text{erf}\left(\frac{x*a-\mu}{\sigma\sqrt{2}}\right)\right]}_{Skewed\ normal\ distribution} \tag{S2}$$

The skewed normal barrier function depends on 5 parameters: L the horizontal asymmetry of the barrier, k the width of the logistic function, a the horizontal asymmetry of the barrier, v the overall vertical scaling of the barrier (used to fit activation energy), and σ the overall width of the skewed normal distribution.

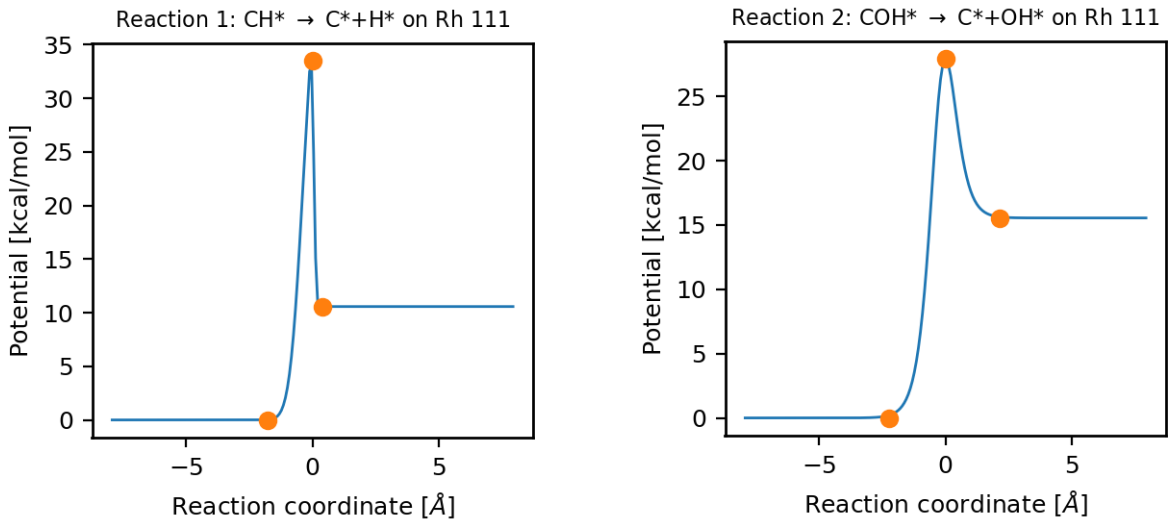

Figure S1: Plots of minimum energy path fits (skew logistic or Eckart) to reactions 1 and 2 reactant, product and transition state energies. See Table 1 of the manuscript for the list of reactions. Orange dots (●) represent (from left to right) the reactant, transition state and product. Blue line (—) represents barrier fit to reaction coordinate.

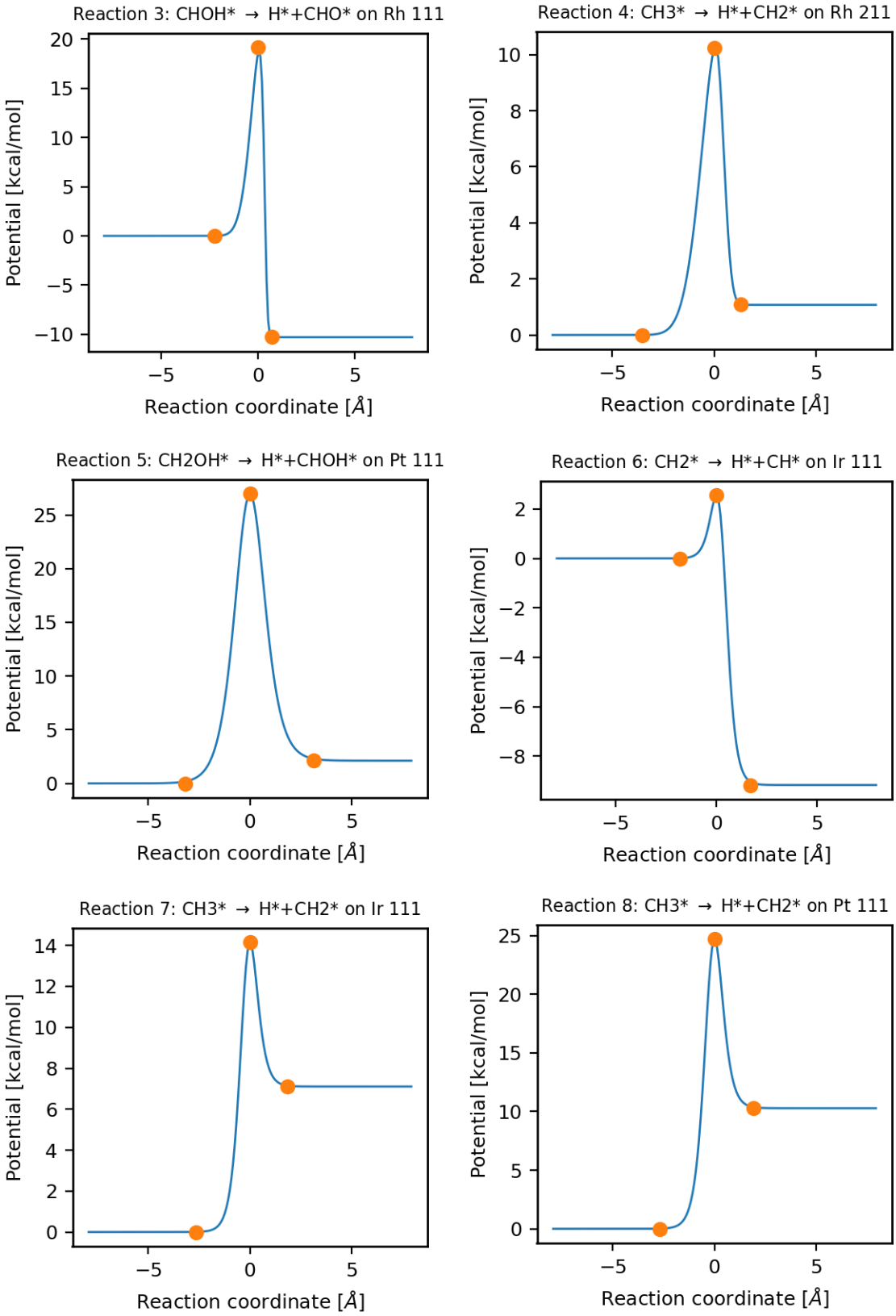

Figure S2: Plots of minimum energy path fits (skew logistic or Eckart) to reactions 3 through 8 reactant, product and transition state energies. See Table 1 of the manuscript for the list of reactions. Orange dots (●) represent (from left to right) the reactant, transition state and product . Blue line (—) represents barrier fit to reaction coordinate.

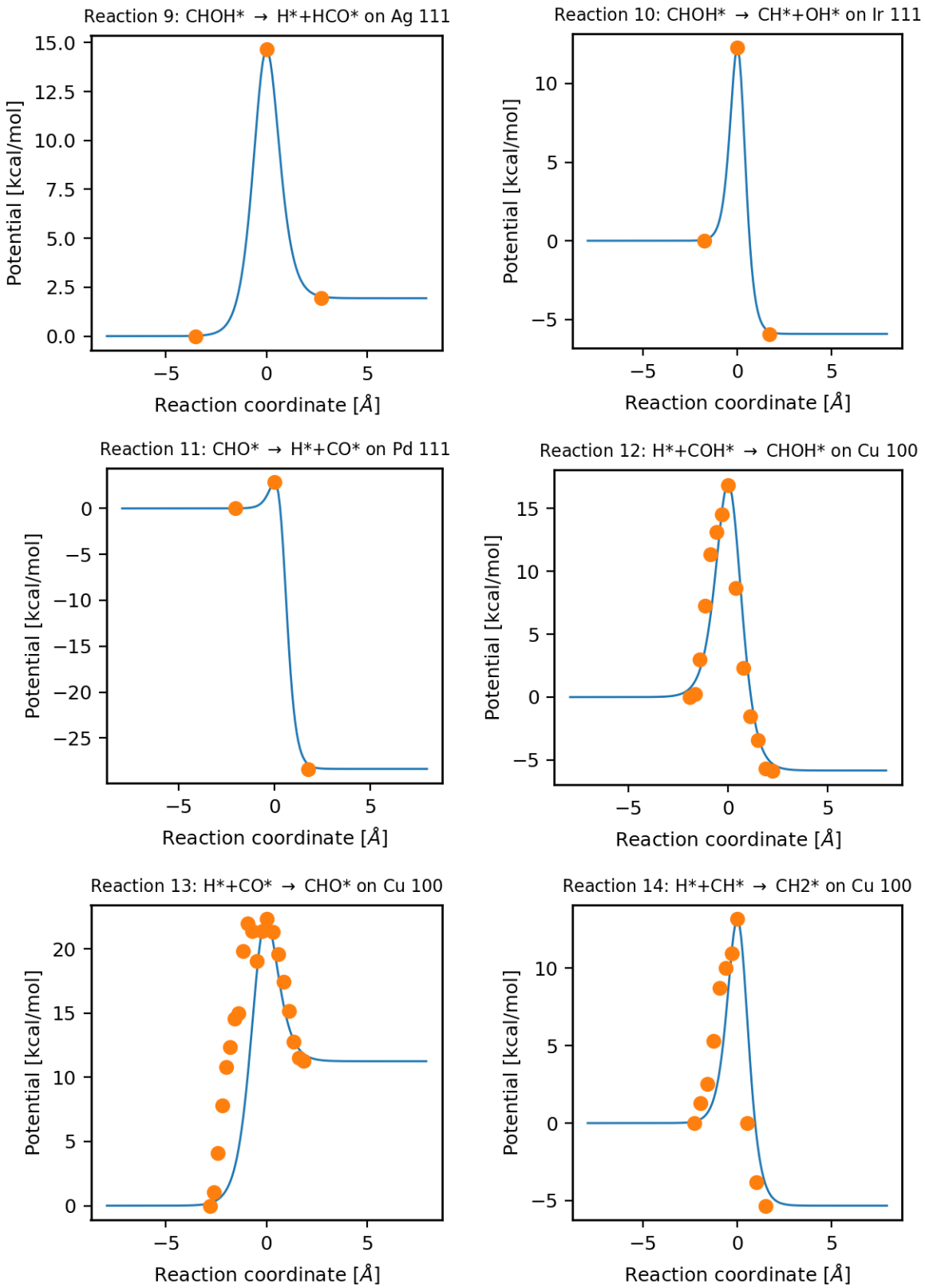

Figure S3: Plots of minimum energy path fits (skew logistic or Eckart) to reactions 9 through 14 reactant, product and transition state energies (reactions 9-11) or minimum energy paths (reactions 12, 13, and 14) (See Table 1 of the manuscript for the list of reactions. Orange dots (•) represent reaction path images. When three dots are shown (from left to right) they indicate the reactant, transition state and product. When more dots are present the first and last indicate the reactant and product and the one at highest potential represents the transition state. The rest indicate images which connect these. Blue line (—) represents barrier fit to reaction coordinate.

2. Discrete Variable Representation and $C_{ff}(t)$ calculation

Table S1: Description of the DVR grid parameters used to compute the flux-flux correlation function for each reaction. Randomly selected temperatures for each reaction are listed in the third column. The fourth column lists the spacing Δx in atomic units between points on the DVR grid. The fifth column lists the grid half-width, *L*. The complete grid covered the range [-*L*, *L*] with N_{DVR} points.

N_{react}	Reaction	T [K]	Δx [au]	*L* [au]	N_{DVR}
1	$CH^*+* \rightarrow C^* + H^*$	[300.0, 324, 356, 388]	0.001	10	20001
2	$COH^*+* \rightarrow C^*+OH^*$	[300.0, 308, 356, 381]	0.0075	8	2133
3	$CHOH^*+* \rightarrow CHO^*+H^*$	[300.0, 304, 355, 392]	0.00075	10	26667
4	$CH_3^*+* \rightarrow CH_2^*+H^*$	[150.0, 180, 277, 394]	0.0075	14	3733
5	$CH_2OH^*+* \rightarrow CHOH^*+H^*$	[300.0, 301, 359, 394]	0.001	14	28001
6	$CH_2^*+* \rightarrow CH^*+H^*$	[150.0, 181, 310, 316]	0.005	8	3201
7	$CH_3^*+* \rightarrow CH_2^*+H^*$	[150.0, 212, 284, 326]	0.0075	8	2133
8	$CH_3^*+* \rightarrow CH_2^*+H^*$	[300.0, 326, 360, 382]	0.01	8	1601
9	$CHOH^*+* \rightarrow HCO^*+H^*$	[200.0, 241, 276, 340]	0.0075	14	3733
10	$CHOH^*+* \rightarrow CH^*+OH^*$	[200.0, 263, 278, 388]	0.005	8	3201
11	$CHO^*+* \rightarrow CO^*+H^*$	[250.0, 254, 310, 373]	0.005	8	3201
12	$COH^*+H^* \rightarrow CHOH^*$	[300.0, 321, 337, 383]	0.0025	10	8001
13	$CO^*+H^* \rightarrow CHO^*$	[250.0, 296, 336, 381]	0.0075	10	2667
14	$CH^*+H^* \rightarrow CH_2^*$	[250.0, 266, 321, 353]	0.01	12	2401

Flux-flux correlation functions, $C_{ff}(t)$, were computed using software developed in-house on a time grid with a time spacing $\Delta t_1 = 35$ [au time] for the first 2000 au time, and $\Delta t_2 = 70$ [au time] for the rest of the time period. Calculations were run to a minimum stopping time of 10,000 [au time]. To calculate rate constant products $k(T) \cdot Q_R(T)$, the $C_{ff}(t)$ time series data was integrated using a trapezoidal integration scheme until a change of $C_{ff}(t)$ value of less than 1% was observed between 2 time points, or until time 8000 [au] was reached.

3. Grid search on input features for $k(T) \cdot Q_R(T)$

For every set of input features tested, the following GPR kernels, and every possible combination of 2 kernels, were iterated over:

- Matern
- Radial Basis Function
- Rational Quadratic
- Pairwise Kernel
- White Noise Kernel

An initial kernel length scale of 1.0 and length scale bounds of (1x10-5, $1x10^5$) were used for the Matern and Radial basis function kernels. Gaussian Process Regressor training was done using 50 optimizer restarts.

Table S2: Representation and kernel hyperparameter grid search results for temperature split data. Green shading indicates selected input features and kernel. Optimal representations and kernels were selected based on train set MAE. This search was done using the default value of the noise parameter, $\alpha = 10^{-10}$. Results in this table are from models trained with this α value.

Representation	Optimal kernel on train set	Train set MAE
Coulomb Matrix Difference, inverse temperature, and reaction energy, all minmax scaled	60.8^2 * Matern(length_scale=4.49, nu=1.5) + 15.5^2 * PairwiseKernel(gamma=1997.944185658677, metric=linear)	9.92×10^{-11}
Encoded Bonds Difference, inverse temperature, and reaction energy, all minmax scaled.	7.69^2 * Matern(length_scale=29.6, nu=1.5) + 306^2 * Matern(length_scale=29.6, nu=1.5)	1.11×10^{-10}

Table S3: Representation and kernel hyperparameter grid search results for reaction split. Green shading indicates selected input features and kernel. Optimal representations and kernels were selected based on train set MAE. This search was done using the default value of the noise parameter, $\alpha = 10^{-10}$. Results in this table are from models trained with this α value.

Representation	Optimal kernel on train set	Train set MAE
Coulomb Matrix Difference, inverse temperature, and reaction energy, all rescaled (minmax scaler)	128^2 * Matern(length_scale=11.5, nu=1.5) + 0.0114^2 * Matern(length_scale=189, nu=1.5)	5.69×10^{-10}
Encoded Bonds Difference, inverse temperature, and reaction energy, all rescaled (minmax scaler)	7.31^2 * PairwiseKernel(gamma=0.004392928734334214, metric=linear) + 316^2 * Matern(length_scale=35.3, nu=1.5)	6.60×10^{-10}

4. Input features for $C_{ff}(t)$ Cauchy Curve fit

Coulomb matrix difference geometry features were used for the $C_{ff}(t)$ Cauchy curve fit problem, as these were found to work best for the $kQ(T)$ problem. Coulomb matrix features, inverse temperature, and reaction energy were used as input features and were rescaled using the minmax scaler. The Cauchy fit scale parameter was also rescaled.

Table S4: Best kernels based on train set MAE for Cauchy curve fit parameter prediction task. This search was done using the default value of the noise parameter, $\alpha = 10^{-10}$. Results in this table are from models trained with this α value.

Split	Fit kernel parameters	Train set MAE
Temperature	316^2 * Matern(length_scale=1e+05, nu=1.5) + 316^2 * RationalQuadratic(alpha=1.02, length_scale=1.35)	5.86×10^{-11}
Reaction	223^2 * Matern(length_scale=2.15, nu=1.5) + 141^2 * PairwiseKernel(gamma=0.000679, metric=linear)	6.71×10^{-11}

5. Predicted $k(T) \cdot Q_R(T)$ using temperature split

Here we report the role of undersampled temperature values on the $k(T) \cdot Q_R(T)$ prediction accuracy as a function of temperature (Figure S4).

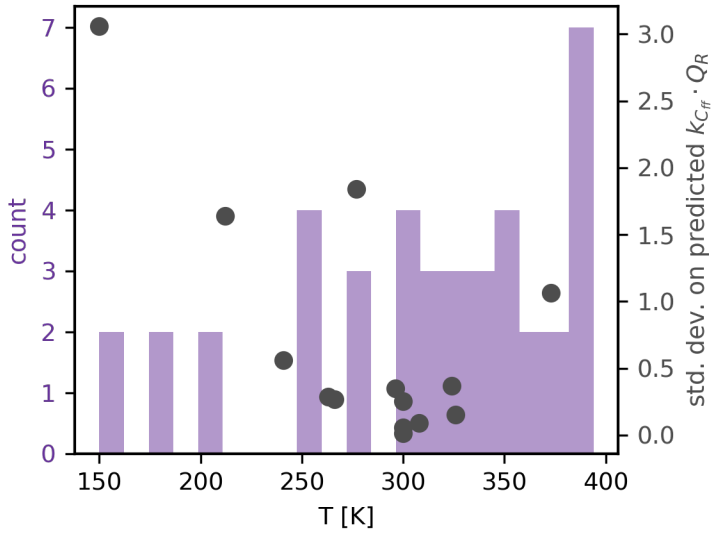

Figure S4: (y-axis left hand side) count of entries in training dataset binned by temperature (x-axis) compared to (y-axis right hand side) standard deviation on predicted value of $k(T) \cdot Q_R(T)$ for test set entries. We see that where the density of training points is higher the predicted error is lower, as expected.

6. Arrhenius Plots for reaction-wise test set predictions

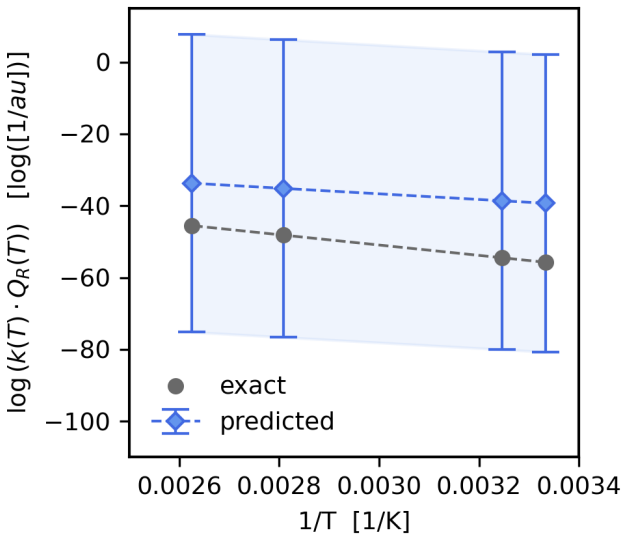

Figure S5: Predicted value of reaction rate constant product as a function of 1/T for the reaction COH*+* C*+OH* on Rh(111)), taken from the test set (Table 1 – reaction 2).

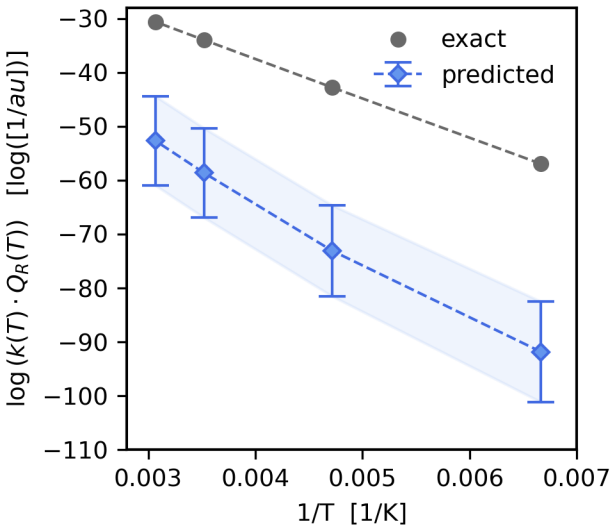

Figure S6: Predicted value of reaction rate constant product as a function of 1/T for the reaction of CH_3^*+*→ CH_2^*+H^* on Ir(111), taken from the test set (Table 1 – reaction 7).

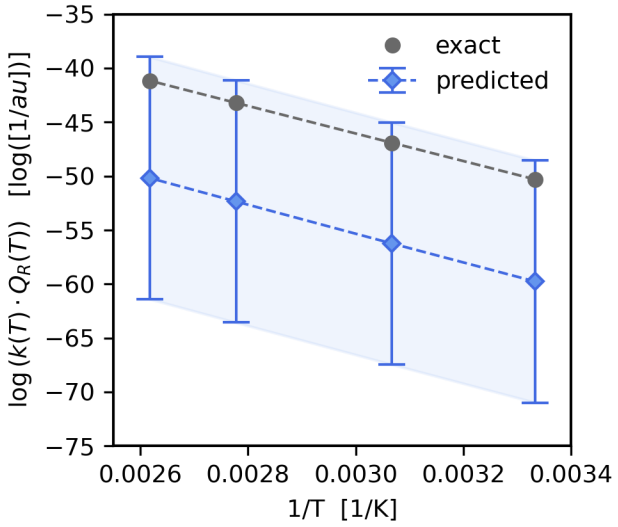

Figure S8: Predicted value of reaction rate constant product as a function of 1/T for the reaction of $CH_3^*+*\rightarrow CH_2^*+H^*$ on Pt(111), taken from the test set (Table 1 – reaction 8).

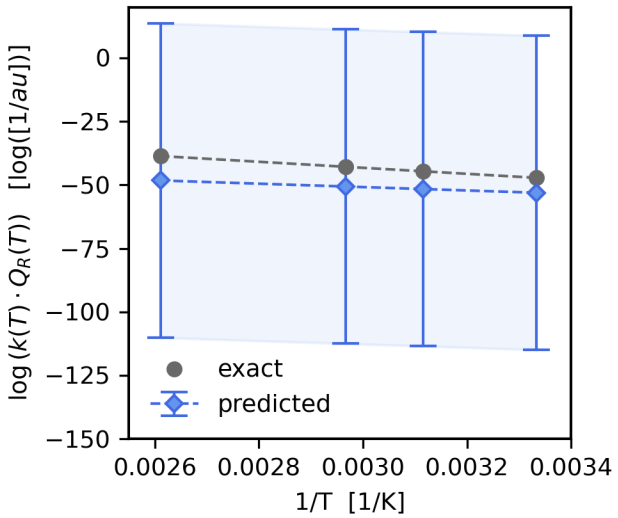

Figure S9: Predicted value of reaction rate constant product as a function of 1/T for the reaction of $COH^*+H^*\rightarrow CHOH^*$ on Cu(100), taken from the test set (Table 1 – reaction 12).

References

1 E. Komp and S. Valleau, *J. Phys. Chem. A.*, 2020, **124**, 8607–8613.