Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information: Machine Learning of Stability Scores from Kinetic Data

Veerupaksh Singla, Qiyuan Zhao, and Brett M. Savoie*

Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, 47906

E-mail: bsavoie@purdue.edu

1 Detailed Methods

1.1 Reaction rate parameters

The rate constants for the barrierless radical recombination reactions were obtained through the collision theory based group additivity rules implemented in Reaction Mechanism Generator (RMG).[1,2] For the other four reaction types, the rate constants (k) were calculated based on the transition state theory (TST). Specifically, the Eyring equation with the shallow Wigner tunneling correction was utilized (Eq. 1).[3]

$$k = \kappa \frac{k_B T}{h} \left(\frac{RT}{P}\right)^{m-1} e^{\frac{-\Delta G^{\ddagger}}{RT}} \tag{1}$$

The tunneling correction term, κ, was calculated using Eq. 2[4] for proton-transfer reactions such as hydrogen abstraction and isomerization, whereas for initiation and radical decomposition reactions,

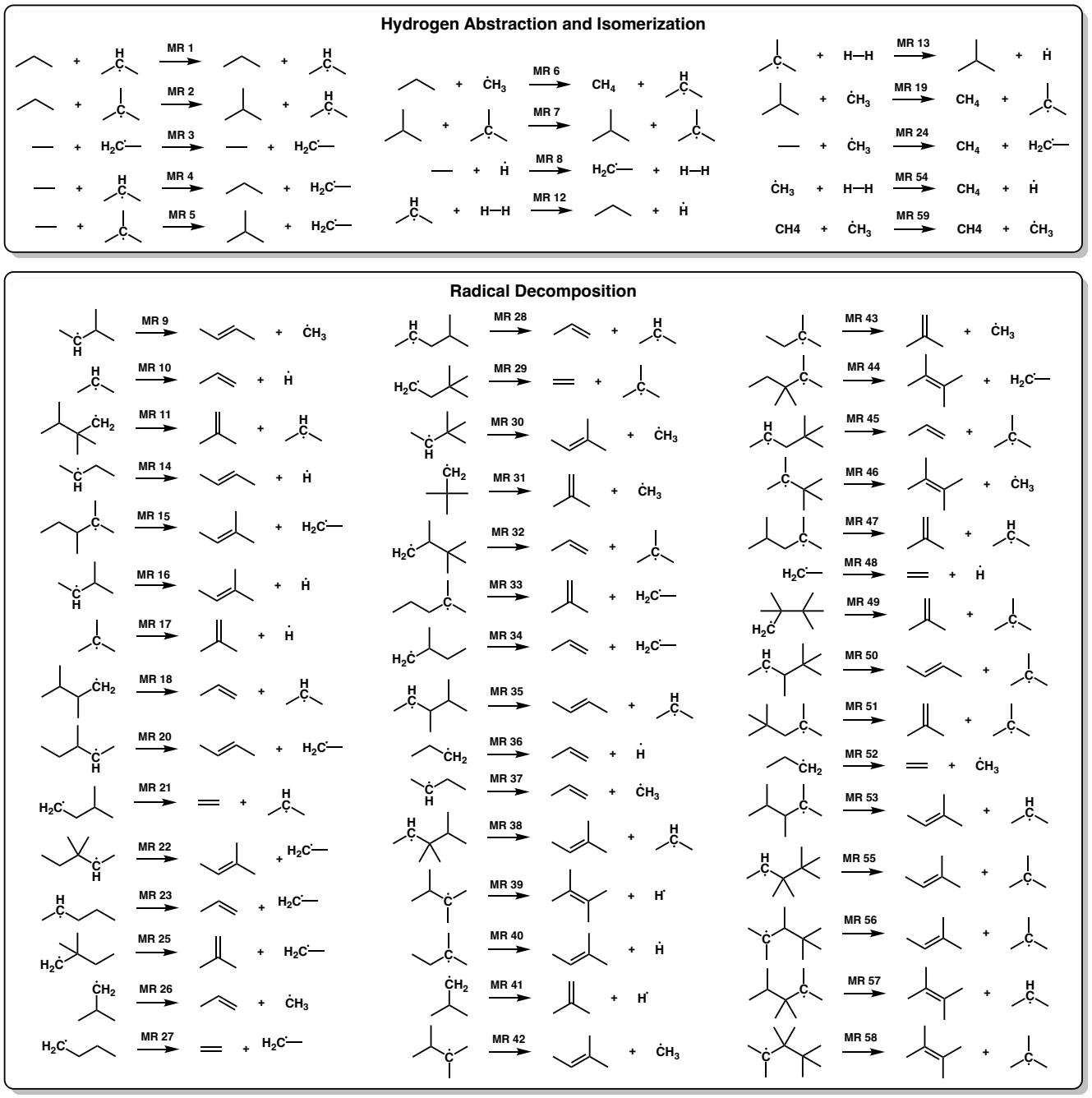

Figure S1: List of all 59 model reactions in the alkane pyrolysis system.

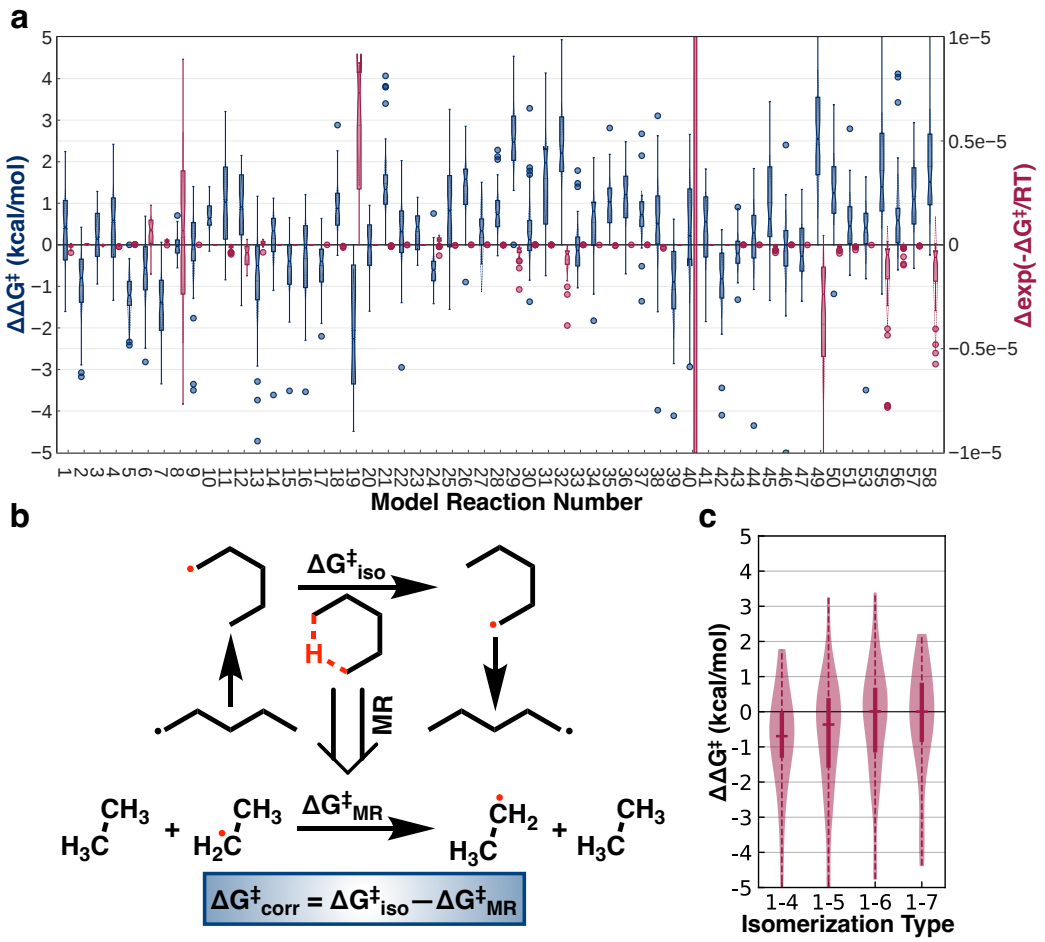

Figure S2: Model reaction benchmarking and correction for isomerization reactions. (a) Signed deviations in free energy of activation (left y-axis), and exponential term variation (right y-axis) of model reactions with respect to the benchmarking reactions. The solid boxes span from the first (q1) to the third (q3) quartile with the second quartile (q2/median) represented by the solid line and notch in the box. The whiskers range from $q1 - 1.5iqr$ to $q3 + 1.5iqr$ where iqr refers to the interquartile range (q3-q1). The dashed lines represent the mean and the standard deviation. (b) Generating correction factors for isomerization reactions from the corresponding model reaction. (c) Signed deviations in corrected free energy of activation of isomerization reactions with respect to the benchmarking reactions. The violin plots span from the minimum to the maximum value, bold vertical lines span from the first to the third quartile, and the bold horizontal tick is the median.

it was set to 1.

$$\kappa = 1 + \frac{1}{24}\left(\frac{h\omega^{\ddagger}}{k_B T}\right)^2 \tag{2}$$

In Eqs. 1- 2, k_B is the Boltzmann constant, h is Planck's constant, T is the operating temperature (i.e. 700 K), R is the ideal gas constant, P is the operating pressure (i.e., 1.01325 bar), m is the molecularity (i.e. the number of reactants) of the reaction, $\Delta G^{\ddagger}$ is the Gibbs energy of activation, and $\omega^{\ddagger}$ is the imaginary frequency of the transition state.

Since initiation reactions are inverse radical recombination reactions, their Gibbs energies of activation were estimated by the corresponding bond dissociation free energies (BDFE),predicting by ALFABET.[5] For the remaining three reaction types, i.e., hydrogen abstraction, isomerization, and radical decomposition, $\Delta G^{\ddagger}$ was estimated by a model reaction based approach (Fig. 1a.ii), to manage the exponential growth in the number of reactions as the alkane length increases.

Model reactions were defined based on the uniqueness of the reaction centers (i.e., the atoms involved in a bond break or formation during the reaction) and their bonded neighbors. The smallest representative of each reaction was used to calculate $\Delta G^{\ddagger}$ from a transition state search (described next) and this value was used as an estimate for all reactions in that class as defined by the reaction centers and bonded neighbors. The smallest representative reactions for each class were generated by hydrogenating the under-coordinated bonded neighbors of the reaction centers to a degree that maintained the reaction center hybridization. An illustration of the approach is shown in Fig. 1a.ii in the main text.

Generating model reactions for all possible hydrogen abstraction, isomerization, and radical decomposition reactions associated with arbitrary alkane pyrolysis resulted in a fixed set of 59 model reactions (Fig. S1) that were characterized with the YARP v2.0 package.[6] Using YARP v2.0 on the model reaction geometries, double-ended transition state searches were performed to generate initial transition state geometry guesses that were then refined at the DFT level.[7] Up

to four independent double-ended searches were performed using distinct reactant and product conformers generated by YARP to promote the discovery of the lowest barrier transition states.[8] The reactant, product, and transition state geometries were optimized at the ωB97X-D/def2-TZVP[9,10] level of theory and the lowest $\Delta G^{\ddagger}$ amongst the discovered transition states for each reaction was used for the kinetic simulations. The Rigid Rotor-Harmonic Oscillator (RRHO) approximation was used to estimate the free energies associated with all DFT calculations.

A detailed benchmarking of the model reaction approach for the alkane pyrolysis network is in Fig. S2a. For each model reaction, activation barriers of at least 40 corresponding actual reactions were calculated at the same level of theory. Note that MR48, MR52, MR54 and MR59 have no other mapped actual reactions but themselves, and hence were not benchmarked. Additionally, a comparison of the errors in exponential terms was made in Fig. S2a. The huge exponential factor variation for MR 40 is explained by the fact that the activation barrier of MR 40 is very small (3 kcal/mol), which causes the rate constant variation to be relatively high even with a 1 kcal/mol change in activation barrier.

Because isomerization reactions are internal hydrogen abstraction reactions and the model reaction approach only conserves the first adjacent neighbors to reacting centers, the model reactions $\Delta G^{\ddagger}$ corresponding to isomerization do not contain the contribution of the ring-based transition state. To account for this, the differences between the simplest observed n-alkyl isomerizations and the corresponding H abstraction model reaction were used as correction factors for each isomerization type (1-4, 1-5, 1-6, or 1-7). These correction factors were then subtracted from the corresponding H abstraction model reactions to obtain the $\Delta G^{\ddagger}$ values of isomerization reactions in the network (Fig. Fig. S2b). Benchmarking of the isomerization correction factors is presented in Fig. S2c.

1.2 Kinetic modeling for half lives

With the ability to generate the pyrolysis reaction network and obtain the corresponding kinetic parameters of these reactions, Cantera[11] was used to solve the rate equations to obtain the half-life of each alkane starting with the pure alkane in an isothermal-isobaric (700 K, 1.01325 bar) ideal gas batch reactor.

The huge reaction networks pose a challenge due to the memory and processing power required to solve these networks in Cantera. From the estimates of possible unique reactants or reactant pairs (for unimolecular and bimolecular reactions, respectively) for n-alkanes, the reactions networks grow at an exponential rate with increasing alkane length. These estimates are very conservative since the actual number of reactions will be much higher than the unique reactants/reactant pairs since each reactant/reactant pair can lead to multiple reactions depending on the number of unique reaction sites in the reactants, and furthermore, the branched isomeric alkanes will have an even higher number of reaction channels compared to n-alkanes. The pruning (or reduction or condensation) of reaction mechanisms is a problem of great interest, and over the years a large number of numerical and analytical methods have been developed to tackle this challenge based on the complexity of the problem.[12]

A simple and robust method to prune reaction networks, especially when reaction rates are known, is to model the reaction kinetics and remove redundant reactions with extremely low relative fluxes (or net rate of production) which are essentially not contributing to the overall kinetics but still consuming computational power. We used this method since it naturally integrated with our efforts to characterize reaction rate constants for all alkane pyrolysis reactions, and to obtain alkane pyrolysis half-lives from these rate constants and the reaction networks. However, as already mentioned, given the super-exponential growth of reaction networks with increasing alkane length, pruning post network generation was not viable. So, we implemented a depth-wise kinetic

pruning strategy wherein at each reaction network generation depth calculated with respect to the starting alkane, the half-life of starting alkane was kinetically modeled (using the network generated up to that depth) and reactions with low relative fluxes across all time steps were discarded before moving on to generating reactions for the next depth (Fig. 1a.iii). The threshold maximum relative flux to be used to prune reactions was obtained to be 10^{-9} via trial-and-error comparison of alkane half-lives from the pruned networks to the half-lives obtained from full reaction networks for all isomers between C_2H_6 and $C_{10}H_{22}$. That is, for the pruning step at each depth, all reactions with maximum relative fluxes over the simulation window of less than 10^{-9} that of the highest flux reactions were discarded. The benchmarking of this kinetic pruning approach has been illustrated in the main text (Fig. 2b).

1.3 Predicting Stability

The kinetic simulations described in the previous sections resulted in a dataset of 32,421 alkanes and their corresponding half lives under pyrolysis conditions. A 90:10 train:test split was randomly generated for models training. To provide a more rigorous test of transferability, four other case-studies were performed using distinct training and testing splits, as already mentioned in section 2.2 of the main text.

The stability score was developed as a relative metric to compare the thermal stability, based on the pyrolysis half-life data, of two or more alkanes. Rather than training the models to predict the half-lives directly, we hypothesized that a relative measure of stability might be more transferable and easier to learn. The Stability Score can be understood to be a log-scaled relative half-life of an alkane with the half-lives scaling exponentially with the Stability Score. Here, the stability score was learned by training the models to minimize a hinge loss function that penalizes incorrect predictions of the pair-wise relative stability of samples. Given a set of n alkanes: $\{m_1, m_2, ..., m_n\}$, and the corresponding log10 and z-normalized actual and model predicted half lives: $\{t_1, t_2, ..., t_n\}$

and $\{\hat{t}_1, \hat{t}_2, ..., \hat{t}_n\}$, the hinge loss function is defined as

$$L = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} max\left(0, \left[(t_i - t_j) - (\hat{t}_i - \hat{t}_j)\right] \cdot \frac{t_i - t_j}{|t_i - t_j|}\right) \tag{3}$$

where the summations are over all unique pairs of samples in the training data. This loss function penalizes relative stability differences between species that are smaller than the log10 half-life difference in training data. This prevents overfitting by ensuring that if the predicted difference in stabilities is at least as much as the training data, the value inside the summation will be zero and will not affect the model parameter optimization.

A feed-forward MLP (Multi-Layer Perceptron) with 2048 length 8-bit integer folded Morgan fingerprints from RDKit as inputs, five hidden layers with 300 nodes each, with ReLU activation with bias, and a single linear output was implemented in python 3.10 with Tensorflow[13] 2.10.0 and Keras[14] (Fig. 1b). Due to the large variation in the alkane half-life values, the model was trained on the natural log and z-score normalized half-life values. Once trained, the model predictions were then wrapped in a sigmoid function and scaled from 0 to 100 for a more human readable format, and these scaled values were called the Stability Scores. The Adam optimizer with a constant learning rate of 0.001 was used for training. Training was stopped based on minimizing validation set performance. A batch size of 128 was used while training.

A Chemprop[15] model with alkane SMILES[16] and hinge loss was also trained for comparison. Both the MLP and Chemprop models were trained on all the split types.

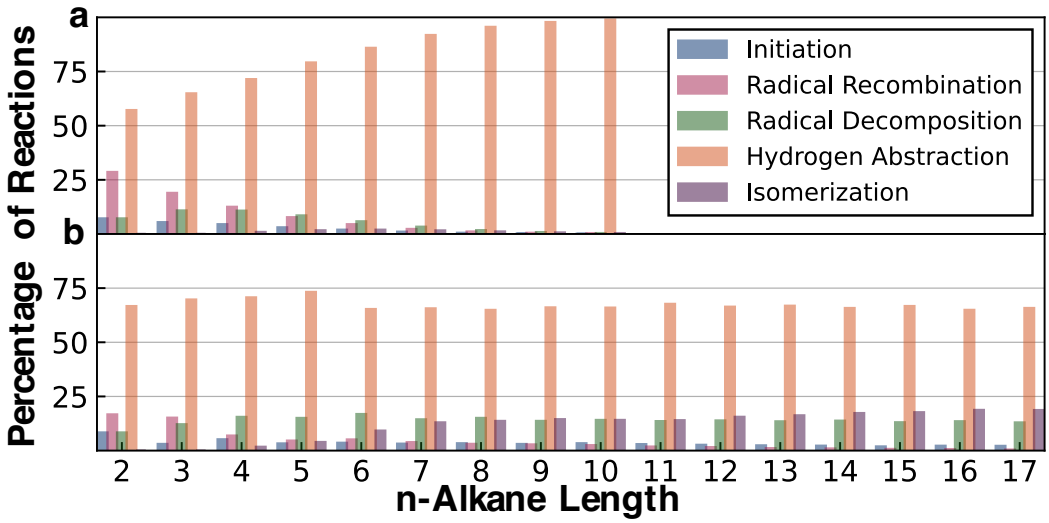

Figure S3: Comparing n-alkane reaction network ERS percentage composition (a) before and (b) after pruning.

2 Reaction networks ERS (Elementary Reaction Step) composition

Fig. S3 describes the percentage composition of the ERS (Elementary Reaction Steps) in some complete and pruned n-alkane reaction networks illustrating the domination of H-abstraction reactions.

3 Experimental Kinetic Data

Table S1 lists all the experimental reactions and rate parameters used.[17] Table S2 shows the reactions which were used from Table S1 for the corresponding reaction schemes.

Table S1: All Reactions and Kinetic Parameters

No.	Reaction	A (s^{-1} or L mol^{-1} s^{-1})	E_a (kcal/mol)
1.	$C_2H_6 \longrightarrow 2\,CH_3\cdot$	4.0×10^{16}	87.5
2.	$C_2H_6 + H\cdot \longrightarrow C_2H_5\cdot + H_2$	1.0×10^{11}	9.7
3.	$C_2H_6 + CH_3\cdot \longrightarrow C_2H_5\cdot + CH_4$	3.8×10^{11}	16.5
4.	$C_2H_5\cdot \longrightarrow C_2H_4 + H\cdot$	3.2×10^{13}	40.0
5.	$C_2H_5 + H\cdot \longrightarrow C_2H_6$	4.0×10^{10}	0
6.	$2\,CH_3\cdot \longrightarrow C_2H_6$	1.3×10^{10}	0
7.	$C_3H_8 \longrightarrow C_2H_5\cdot + CH_3\cdot$	2.0×10^{16}	84.5
8.	$C_3H_8 + H\cdot \longrightarrow 1-C_3H_7\cdot + H_2$	1.0×10^{11}	9.7
9.	$C_3H_8 + H\cdot \longrightarrow 2-C_3H_7\cdot + H_2$	9.0×10^{10}	8.3
10.	$C_3H_8 + CH_3\cdot \longrightarrow 1-C_3H_7\cdot + CH_4$	3.4×10^{10}	11.5
11.	$C_3H_8 + CH_3\cdot \longrightarrow 2-C_3H_7\cdot + CH_4$	4.0×10^{9}	10.1
12.	$C_3H_8 + C_2H_5\cdot \longrightarrow 1-C_3H_7\cdot + C_2H_6$	1.2×10^{9}	12.6
13.	$C_3H_8 + C_2H_5\cdot \longrightarrow 2-C_3H_7\cdot + C_2H_6$	8.0×10^{8}	10.4
14.	$1-C_3H_7\cdot \longrightarrow C_2H_4 + CH_3\cdot$	4.0×10^{13}	32.6
15.	$1-C_3H_7\cdot \longrightarrow C_3H_6 + H\cdot$	2.0×10^{13}	38.4
16.	$2-C_3H_7\cdot \longrightarrow C_3H_6 + H\cdot$	2.0×10^{13}	38.7
17.	$1-C_3H_7\cdot + H\cdot \longrightarrow C_3H_8$	1.0×10^{10}	0
18.	$2-C_3H_7\cdot + H\cdot \longrightarrow C_3H_8$	1.0×10^{10}	0
19.	$C_2H_5\cdot + CH_3\cdot \longrightarrow C_3H_8$	3.2×10^{9}	0

Continued on next page

Table S1 – continued from previous page

No.	Reaction	A (s^{-1} or L mol^{-1} s^{-1})	E_a (kcal/mol)
20.	$n{-}C_4H_{10} \longrightarrow 2\,C_2H_5\cdot$	1.5×10^{16}	82.1
21.	$n{-}C_4H_{10} \longrightarrow 1{-}C_3H_7\cdot + CH_3\cdot$	9.0×10^{16}	85.4
22.	$i{-}C_4H_{10} \longrightarrow 2{-}C_3H_7\cdot + CH_3\cdot$	2.0×10^{16}	82.0
23.	$n{-}C_4H_{10} + H\cdot \longrightarrow 1{-}C_4H_9\cdot + H_2$	1.5×10^{11}	9.7
24.	$n{-}C_4H_{10} + H\cdot \longrightarrow 2{-}C_4H_9\cdot + H_2$	9.0×10^{10}	8.4
25.	$i{-}C_4H_{10} + H\cdot \longrightarrow i{-}C_4H_9\cdot + H_2$	1.0×10^{11}	8.4
26.	$n{-}C_4H_{10} + CH_3\cdot \longrightarrow 1{-}C_4H_9\cdot + CH_4$	3.5×10^{10}	11.6
27.	$n{-}C_4H_{10} + CH_3\cdot \longrightarrow 2{-}C_4H_9\cdot + CH_4$	3.5×10^{9}	9.5
28.	$i{-}C_4H_{10} + CH_3\cdot \longrightarrow i{-}C_4H_9\cdot + CH_4$	9.5×10^{9}	9.0
29.	$n{-}C_4H_{10} + C_2H_5\cdot \longrightarrow 1{-}C_4H_9\cdot + C_2H_6$	2.0×10^{9}	12.6
30.	$n{-}C_4H_{10} + C_2H_5\cdot \longrightarrow 2{-}C_4H_9\cdot + C_2H_6$	4.5×10^{8}	10.4
31.	$i{-}C_4H_{10} + C_2H_5\cdot \longrightarrow i{-}C_4H_9\cdot + C_2H_6$	1.5×10^{9}	10.4
32.	$n{-}C_4H_{10} + 1{-}C_3H_7\cdot \longrightarrow 2{-}C_4H_9\cdot + C_3H_8$	2.0×10^{8}	10.4
33.	$n{-}C_4H_{10} + 2{-}C_3H_7\cdot \longrightarrow 2{-}C_4H_9\cdot + C_3H_8$	2.0×10^{8}	12.6
34.	$i{-}C_4H_{10} + 2{-}C_3H_7\cdot \longrightarrow i{-}C_4H_9\cdot + C_3H_8$	1.0×10^{8}	13.4
35.	$1{-}C_4H_9\cdot \longrightarrow C_2H_4 + C_2H_5\cdot$	1.6×10^{12}	28.0
36.	$1{-}C_4H_9\cdot \longrightarrow 1{-}C_4H_8 + H\cdot$	1.0×10^{13}	36.6
37.	$2{-}C_4H_9\cdot \longrightarrow 1{-}C_4H_8 + H\cdot$	2.0×10^{13}	39.8
38.	$i{-}C_4H_9\cdot \longrightarrow i{-}C_4H_8 + H\cdot$	3.3×10^{14}	36.0
39.	$2{-}C_4H_9\cdot \longrightarrow C_3H_6 + CH_3\cdot$	2.5×10^{13}	31.9
40.	$i{-}C_4H_9\cdot \longrightarrow C_3H_6 + CH_3\cdot$	8.0×10^{13}	33.0

Continued on next page

Table S1 – continued from previous page

No.	Reaction	A (s^{-1} or L mol^{-1} s^{-1})	E_a (kcal/mol)
41.	$i{-}C_4H_9\cdot \longrightarrow 2{-}C_4H_8 + H\cdot$	4.0×10^{13}	36.6
42.	$1{-}C_4H_9\cdot + H\cdot \longrightarrow n{-}C_4H_{10}$	1.0×10^{10}	0
43.	$2{-}C_4H_9\cdot + H\cdot \longrightarrow n{-}C_4H_{10}$	1.0×10^{10}	0
44.	$i{-}C_4H_9\cdot + H\cdot \longrightarrow i{-}C_4H_{10}$	1.0×10^{10}	0
45.	$1{-}C_3H_7\cdot + CH_3\cdot \longrightarrow n{-}C_4H_{10}$	3.2×10^{9}	0
46.	$C_2H_5\cdot + C_2H_5\cdot \longrightarrow n{-}C_4H_{10}$	3.2×10^{9}	0

Table S2: Alkane Pyrolysis Schemes

Alkane	Reactions Considered
Ethane	1-6
Propane	1-19
n-Butane	1-46
iso-Butane	1-46

4 Models Training Details

The learning curves for all split types for the MLP model are shown in Fig. S4 and for the Chemprop model are shown in Fig. S5. 10% of each training split was withheld as validation data for determining stopping for training. The Best Validation model was used for testing. It can be observed that across all splits and models, the training and validation losses drop quickly within the first 20 epochs and the majority of the training happens within the first 100 epochs with another 200 epochs being required for fine tuning. Also, it is important to note that the loss values for MLP and Chemprop cannot be directly compared because of the different default

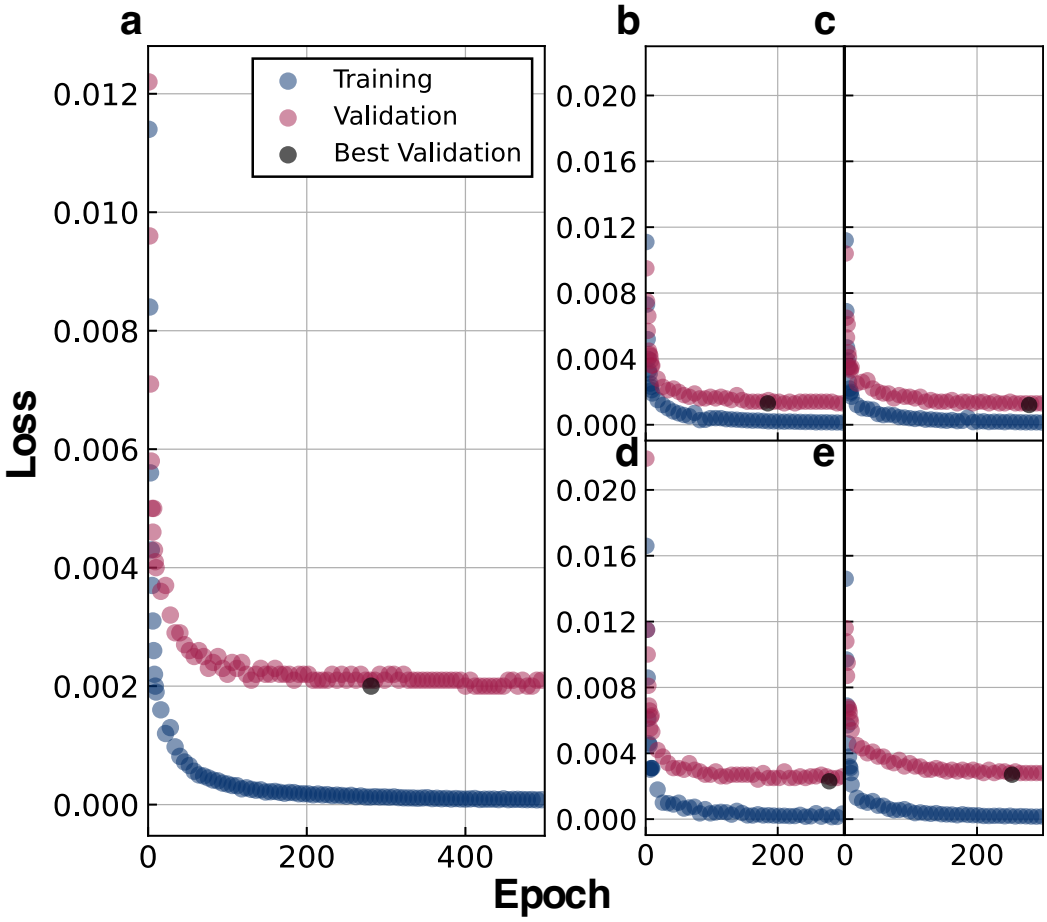

Figure S4: MLP learning curves (a) Random Split. (b) Total Branches ≤ 6. (c) Core Branches ≤ 4. (d) Backbone ≤ 10. (e) Length ≤ 16.

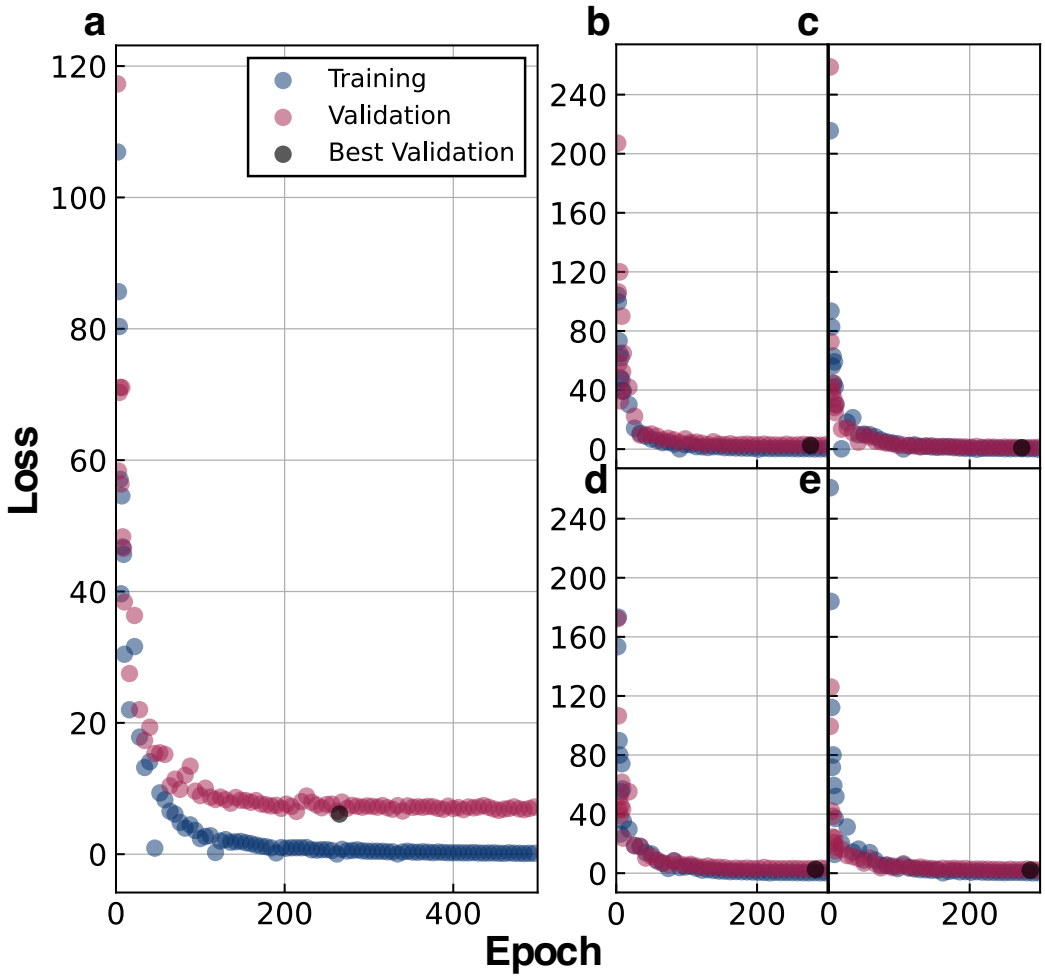

Figure S5: Chemprop learning curves (a) Random Split. (b) Total Branches $\leq$ 6. (c) Core Branches $\leq$ 4. (d) Backbone $\leq$ 10. (e) Length $\leq$ 16.

initialization.

References

(1) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. Reaction Mechanism Generator: Automatic Construction of Chemical Kinetic Mechanisms. *Computer Physics Communications* **2016**, *203*, 212–225, DOI: 10.1016/j.cpc.2016.02.013.

(2) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J.; Blondal, K.; West, R. H.; Goldsmith, C. F.; Green, W. H. Reaction Mechanism Generator v3.0: Advances in Automatic Mechanism Generation. *J. Chem. Inf. Model.* **2021**, *61*, 2686–2696, DOI: 10.1021/acs.jcim.0c01480.

(3) Eyring, H. The Activated Complex in Chemical Reactions. *J. Chem. Phys.* **1935**, *3*, 107–115, DOI: 10.1063/1.1749604.

(4) Wigner, E. On the Quantum Correction For Thermodynamic Equilibrium. *Phys. Rev.* **1932**, *40*, 749–759, DOI: 10.1103/PhysRev.40.749.

(5) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at near Chemical Accuracy with Sub-Second Computational Cost. *Nat Commun* **2020**, *11*, 2328, DOI: 10.1038/s41467-020-16201-z.

(6) Zhao, Q.; Savoie, B. M. Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces. *Angew. Chem., Int. Ed. n/a*, e202210693, DOI: https://doi.org/10.1002/anie.202210693.

(7) Zimmerman, P. M. Growing string method with interpolation and optimization in internal coordinates: Method and examples. *J. Chem. Phys.* **2013**, *138*, 184102.

(8) Zhao, Q.; Hsu, H.-H.; Savoie, B. M. Conformational Sampling for Transition State Searches on a Computational Budget. *J. Chem. Theory Comput.* **2022**, *18*, 3006–3016.

(9) Chai, J.-D.; Head-Gordon, M. Long-Range Corrected Hybrid Density Functionals with Damped Atom–Atom Dispersion Corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620, DOI: 10.1039/B810189B.

(10) 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.* **2005**, *7*, 3297–3305, DOI: 10.1039/B508541A.

(11) Goodwin, D. G.; Moffat, H. K.; Schoegl, I.; Speth, R. L.; Weber, B. W. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes. https://www.cantera.org, 2022; Version 2.6.0.

(12) Turányi, T.; Tomlin, A. S. In *Analysis of Kinetic Reaction Mechanisms*; Turányi, T., Tomlin, A. S., Eds.; Springer: Berlin, Heidelberg, 2014; pp 183–312, DOI: 10.1007/978-3-662-44562-4_7.

(13) Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; https://www.tensorflow.org/, Software available from tensorflow.org.

(14) Chollet, F., et al. Keras. https://keras.io, 2015.

(15) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388, DOI: 10.1021/acs.jcim.9b00237.

(16) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

(17) Sundaram, K. M.; Froment, G. F. Modeling of Thermal Cracking Kinetics. 3. Radical Mechanisms for the Pyrolysis of Simple Paraffins, Olefins, and Their Mixtures. *Ind. Eng. Chem. Fund.* **1978**, *17*, 174–182, DOI: 10.1021/i160067a006.