Supplementary Material for *Substituting density functional theory in reaction barrier calculations for hydrogen atom transfer in proteins*

Kai Riedmiller[a], Patrick Reiser[b,c], Elizaveta Bobkova[a], Kiril Maltsev[a], Ganna Gryn'ova[a,d], Pascal Friederich[b,c,*], Frauke Gräter[a,d,*]

[a] *Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. *E-mail: frauke.graeter@h-its.org*

[b] *Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Engler-Bunte-Ring 8, 76131 Karlsruhe, Germany. *E-mail: pascal.friederich@kit.edu*

[c] *Institute of Nanotechnology, Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1: 76344 Eggenstein-Leopoldshafen, Germany.*

[d] *Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany*

1 Methods

1.1 Synthetic Structure generation

The synthetic systems are build from a pool of structures. This pool contains single amino acids, the modified amino acids dihydroxyphenylalanine (dopa) and hydroxyproline, obtained from pubchem.[1] Additionally, modified versions of these structures are used, in which the free carboxyl group is converted into an amid group to closer mimic the chemical situation in a protein. Water, as well as nine amino acid fragments derived from alanine, glutamine, glycine, hydroxyproline and proline, resulting from breaks in the backbone of a protein are also added to the pool.

All structures in the pool are considered in every possible charge state. From this pool of structures two, not necessarily the same, structures are selected. Within each of them, a hydrogen atom is randomly picked. The hydrogens and their bonds are placed on top of each other, one of the structures is rotated 180° around an axis perpendicular to the bond. Then the structures are pulled apart, one structure is rotated around the bond of the hydrogen and tilted around the hydrogen of the other structure. Also see Figure 2 A in the main text for the definition of the translation, rotation and tilt. The translations are sampled from a standard exponential distribution, the tilt angle from a normal distribution with standard deviation of 45, the rotations are sampled in 30° increments. If any atom from one structure ends up being closer than 2 Å to any atom of the other structure, this structure pair is discarded. Additionally, systems get excluded if in a cylinder around the transition path with a radius of 0.8 Åany atom is found, as this would lead to non-physical reactions paths.

1.2 Trajectory Structure generation

The trajectory systems are generated from a collagen model spanning one overlap and one gap region, which was obtained from Colbuilder.[2] Sequences from *Loxodonta africana*, *Pongo abelii* and *Rattus norvegicus* were used, paired with the divalent HLKNL and the trivalent PYD crosslink at various positions. These collagen models are simulated under tension, i.e., the peptide chain ends are pulled apart. On each collagen chain, a force of 1 nN, or 3 nN per triple helix is exerted. Four different pulling methods are used: I) the triple helices are pulled from both ends, II) and III) they are pulled from one side, while the other is fixed in place, and IV) on different triple helices different forces are exerted, drawn from a Gaussian distribution with $F_{av} = 1nN$ mean force and width $\sigma = F_{av}/3$. In this last scheme, the outer ring of triple helices is still pulled at the average force to prevent triple helices from sliding. More details on the pulling simulations are given by Rennekamp et al.,[3] where the same simulations are used.

The simulations can further by divided in two groups, one containing two radicals as a result of a backbone break, the other consisting of intact collagen system without radicals. For the latter, HAT reactive systems are sampled by H-H distance. As the energy barrier is highly correlated with this distance, small distance were emphasized when sampling. >98% of the samples have translation <3 Å, >50% <2 Å. (Figure S1) Systems, where atoms are closer than 0.8 Åto the transition path of the reacting hydrogen are excluded in the same way as for synthetic systems. The radicals in the other type are result of a homolytic bond breakage performed using KIMMDY.[4] In these trajectories, only potential HAT reactions involving the existing radicals are considered. Again, systems with atoms intersecting the transition path of the hydrogen where excluded. In contrast to the fully saturated systems, the exact end position of the reacting hydrogen is not known and is therefore guessed from the geometry. In case of an ambiguous endpoint, like the hydrogen in an alcohol group, the smallest distance on a 109.5° cone around the oxygen is sampled.

To ensure chemical sensible systems, capping groups are added to the cut-out sections from the trajectory. For the N terminus these are acetyl groups and for the C terminus $NH-CH_3$ groups. To know where to add capping groups, first we defined the groups we want to keep: all complete amino acids with atoms in beta position to the reacting hydrogen, or alpha position to the radical heavy atom. The next atoms from adjacent amino acids are then used to construct the capping groups. In Figure S3 A, this is show. Here, the reacting amino acids and the atoms used for capping are drawn solid, while the environment which will be removed is drawn translucent. The result of the capping is shown in Figure S3 B. If the selected amino acids are part of the same

protein chain, but with one other amino acid in between, this amino acid is replaced by glycine to decrease the atom amount (Figure S3 C to E).

1.3 DFT Optimization

As mentioned in the main text, due to the closed-shell parameters of our input MD structures, it is necessary to optimize the structures to obtain realistic barriers. We froze most of the system during the optimization to restrict it to correcting wrong MD parameters, while not perturbing the system any more than necessary. In Figure S5, we visualize the atom layers around the reacting hydrogen. Each layer includes all hydrogen atoms connected to any member of the respective layer.
Layer 1 is highlighted in green and what we ultimately used in the main text. Increasing the optimization region by one bond, layer 2 (yellow) already allows in the shown example complete movement of the side chain. This does lead to the situation, that in some optimization steps (start, TS, end) a hydrogen bond between the NH_3 and an oxygen in the backbone is formed, and in others not. This disturbs the calculated HAT barrier. Furthermore, this rearrangement can be sampled in MD, therefore we want the DFT optimization to stick to other degrees of freedom not covered by MD.

Unsurprisingly, bigger optimized regions do lead to on average lower barriers, as can be seen in panel B and C of Figure S5.

1.4 Hyperparameter

Following hyperparameters were used for training the ensemble model. The complete training routines can be found on github: `https://github.com/HITS-MBM/HAT_prediction_GNN`

```
"loss": "MAE",
"lr_start": 0.0008,
"lr_scheduler": "cos",
"lr_fraction": 0.01,
# Only central nodes of GNN pooled
"out_emb": "poi",
# all mlp layers have same size
"mlp_style": "static",
"mlp_layers": 2,
"mlp_size": 128,
# mlp repetitions with skip connections
"mlp_rep": 1,
# message passing iterations
"depth": 2,
"equiv_norm": false,
"node_norm": false,
# pooling going from gnn to mlp
"pooling": "sum",
"epochs": 200,
"batchsize": 128,
"val_split": 0.1,
# optional output normalization
"scale": false,
# maximum translation
"max_dist": null,
# minimum translation
"min_dist": null,
```

1.5 Random Forest Regressor

A random forest regressor is compared against the developed ensemble GNN. It is used as implemented in scikit-learn[5]. 100 trees are used, the maximum tree depth is unrestricted. No hyperparameter optimization was performed as the mode serves only as a point of reference.

1.6 Software

L-MBTR descriptors are calculated using DScribe 1.2.2.[6] Neural Networks are build using Tensorflow 2.10,[7] graph neural networks use additionally KGCNN.[8]

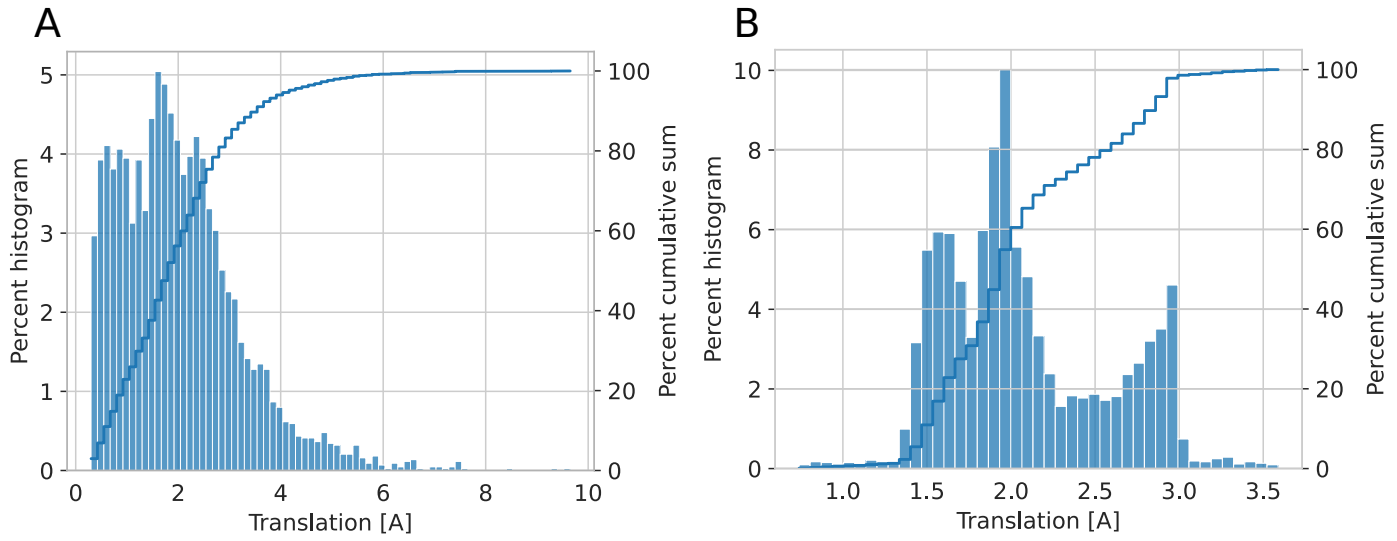

Fig. S1 Histograms of (A) the synthetic data and (B) the trajectory data over translation distance of the reacting hydrogen.

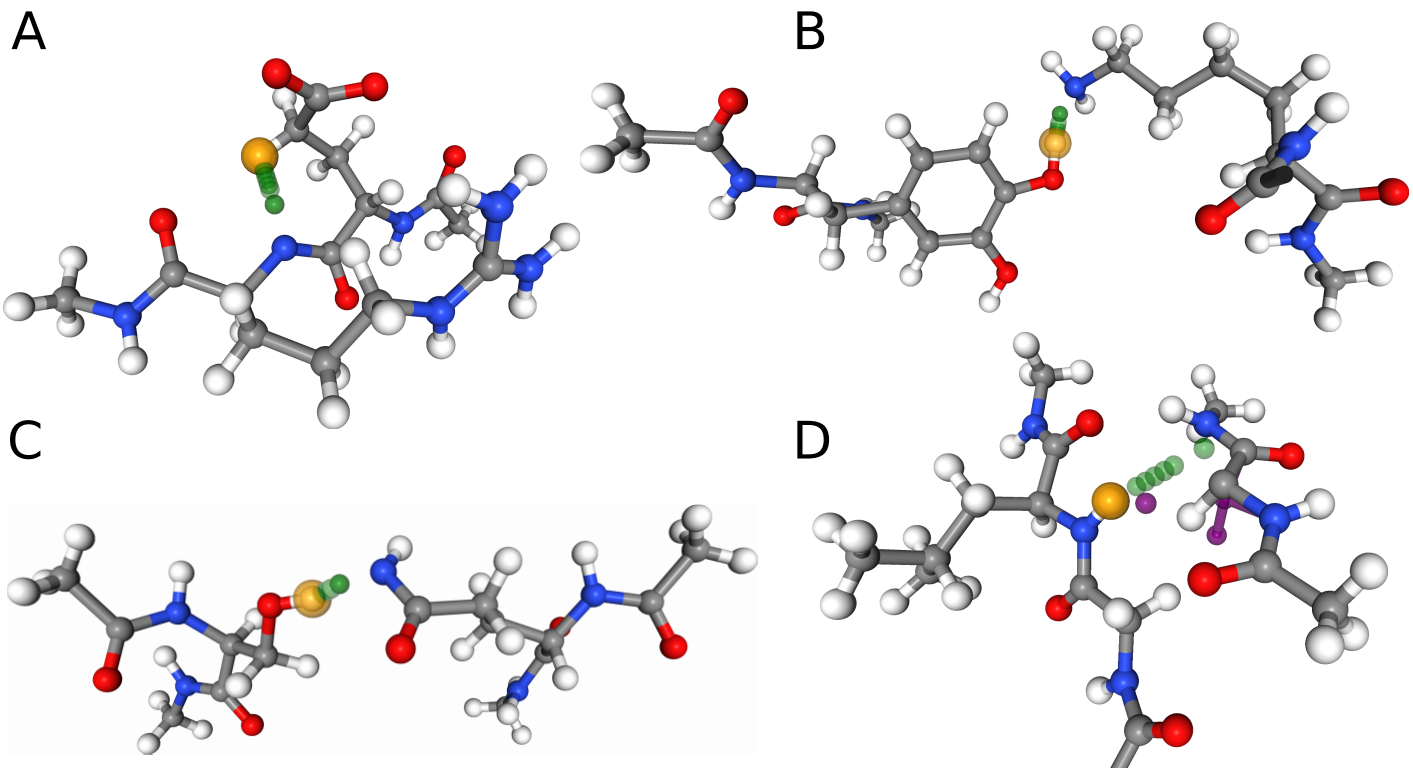

Fig. S2 Example systems with low barriers in A-C, high barrier in D. All barriers in kcal/mol. The start position of the hydrogen is highlighted orange, the interpolated transition path green. (A) $Ea = 19.8; Ea_{opt} = 3.6$ The strong decrease of the barrier during optimization is due to the donating CH_2 group adapting sp^2 conformation. (B) $Ea = 0.3; Ea_{opt} = 0.4$ The barrier in the reverse direction has also a low barrier of $Ea = 8.7$ (C) $Ea = 14.8; Ea_{opt} = 13.0$ Very little rearrangement necessary during the reaction. (D) $Ea = 112.0; Ea_{opt} = 108.8$ The high barrier is caused by another hydrogen interfering with the reaction path. In purple, the optimized TS is shown, where the hydrogen is pushed out of the way. This pushing opens the transition path, but comes with an energy penalty, leading to the high barrier. Additionally, the reacting hydrogen has to travel 2.9 Å.

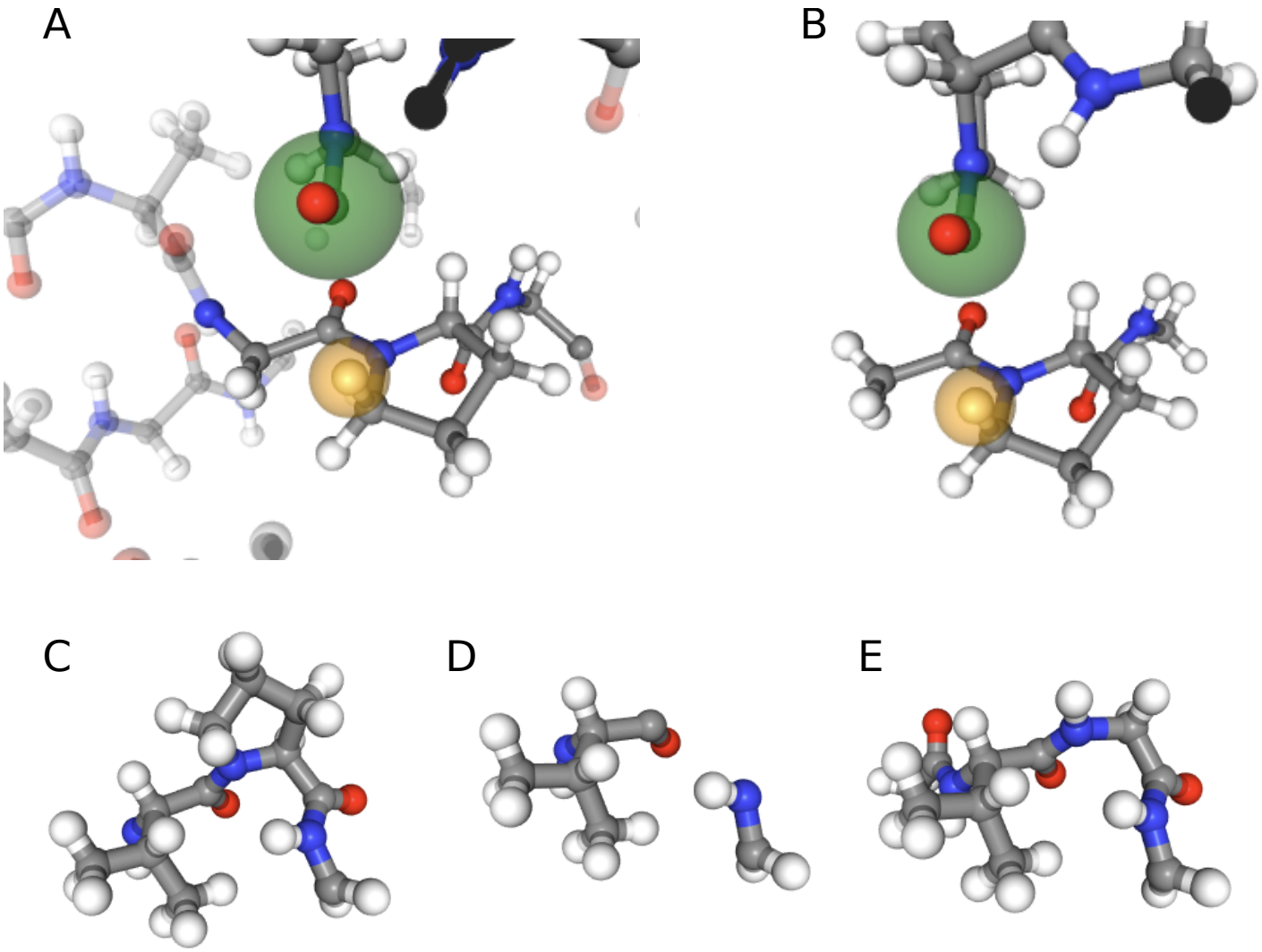

Fig. S3 Capping of trajectory systems. (A) A trajectory system with its context around shown translucently. (Same as Figure 2 C) (B) Same system as in A now isolated with the capping groups. (C) A intramolecular HAT reaction between the radical (bottom right) and the adjacent CH_3 group. (D) Same system as in C. The bridging amino acid between the reacting groups is removed. (E) Same system as in C and D, with capping groups added. A glycine replaces the removed amino acid, in the background an acetyl group caps the N terminus.

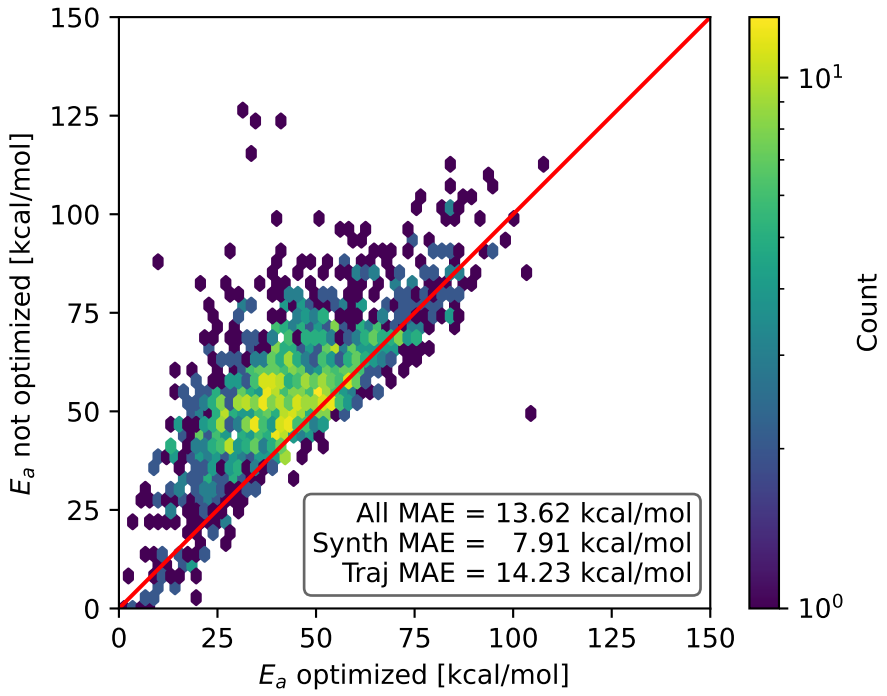

Fig. S4 Histogram of optimized energy barriers versus not optimized energy barriers.

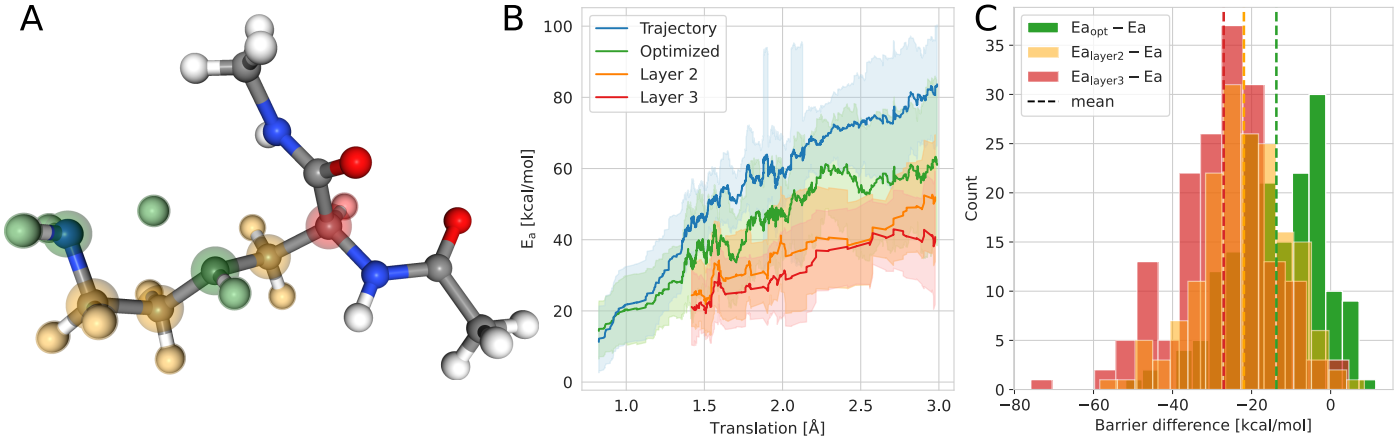

Fig. S5 Analysis of different sized optimization regions. (A) Illustration of different possible optimization regions. The ultimately used region is highlighted green, the next bigger region layer 2 yellow, and layer 3 red. (B) Comparison of the non-optimized trajectory systems, the same systems optimized as in the main text, and these systems optimized at layer 2 and layer 3. (C) Histogram of the barriers optimized at all different layers in relation to the non-optimized ones.

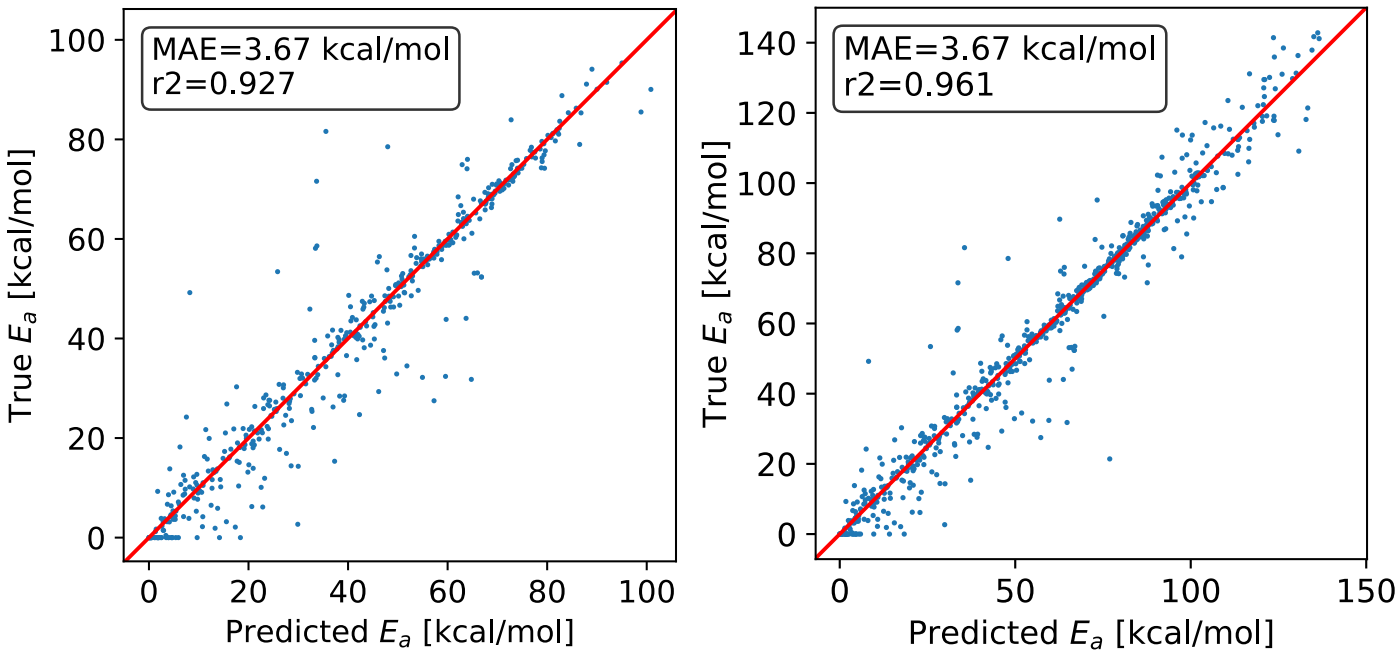

Fig. S6 Performance of the ensemble model on synthetic test data. Left: Translation <2 Å, Right: Translation <3 Å

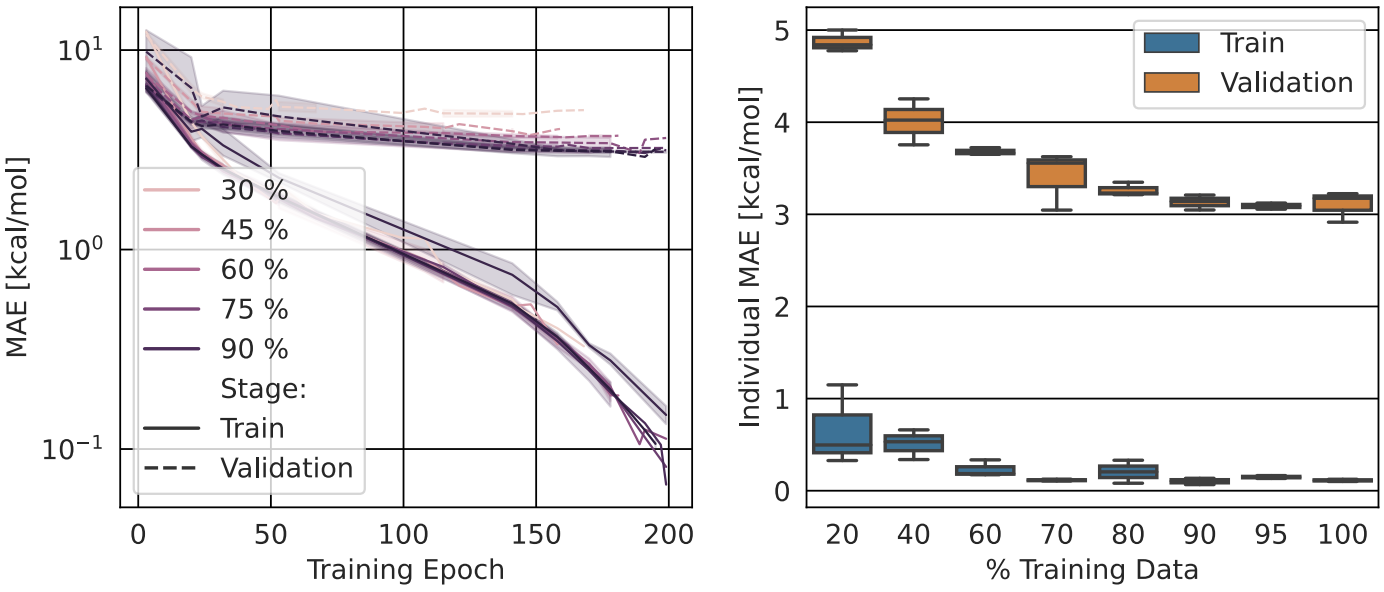

Fig. S7 Training and validation performance of GNNs for non-optimized barriers. Left: Training curves at different points of learning curve, shaded area is the 95% confidence interval. Right: Box plot of learning curve of training and validation data.

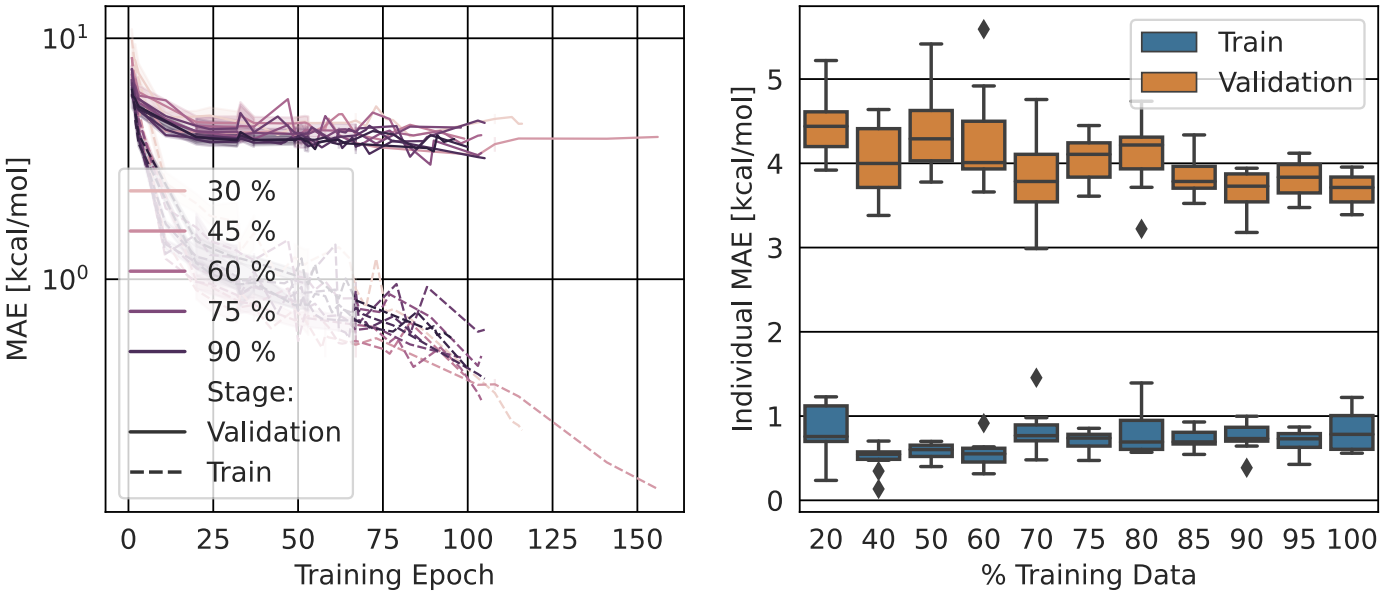

Fig. S8 Training and validation performance of GNNs for transfer learning optimized barriers. Left: Training curves at different points of learning curve, shaded area is the 95% confidence interval. Right: Box plot of learning curve of training and validation data.

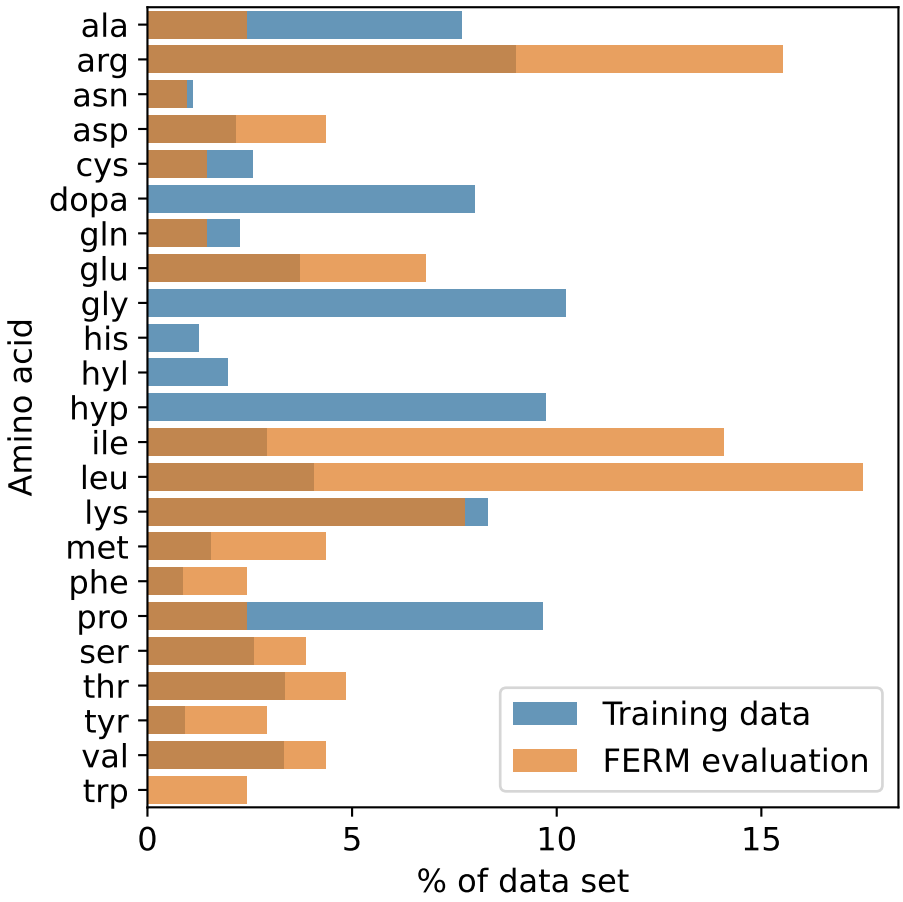

Fig. S9 Amino acid distribution in the training data set, including synthetic and trajectory data, compared to the amino acid distribution in the test data of the F0F1 domains of FERM.

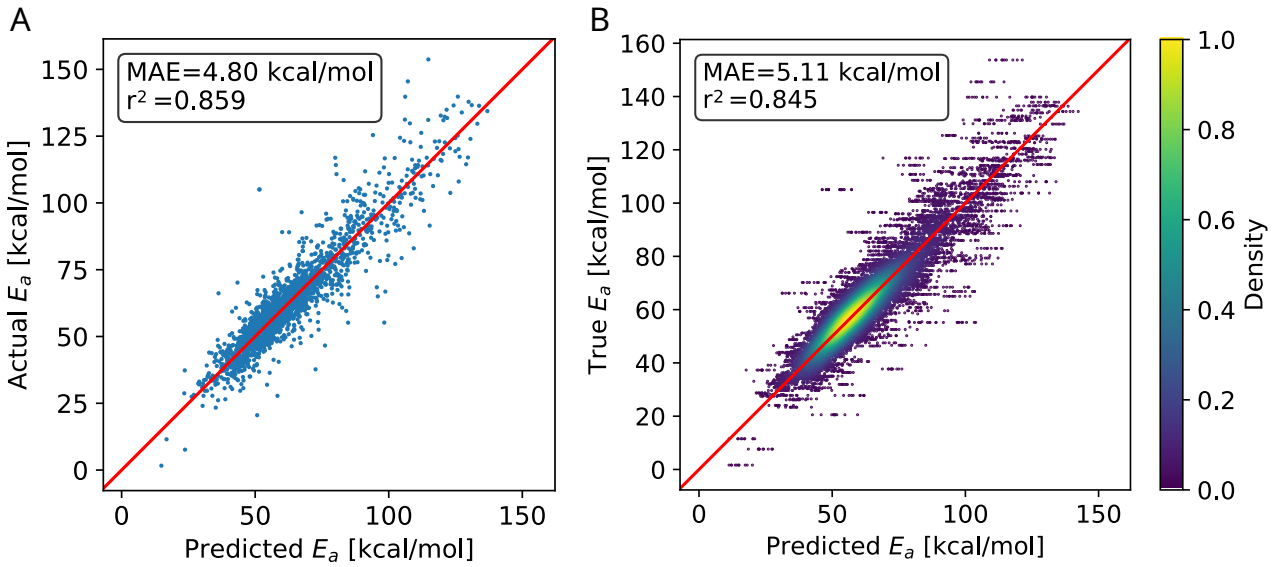

Fig. S10 Performance of the feed-forward neural network using L-MBTR as input. The model is evaluated on all trajectory data. In A, the ensemble model performance is plotted, B shows the performance of ten individual models in a density-colored scatter plot.

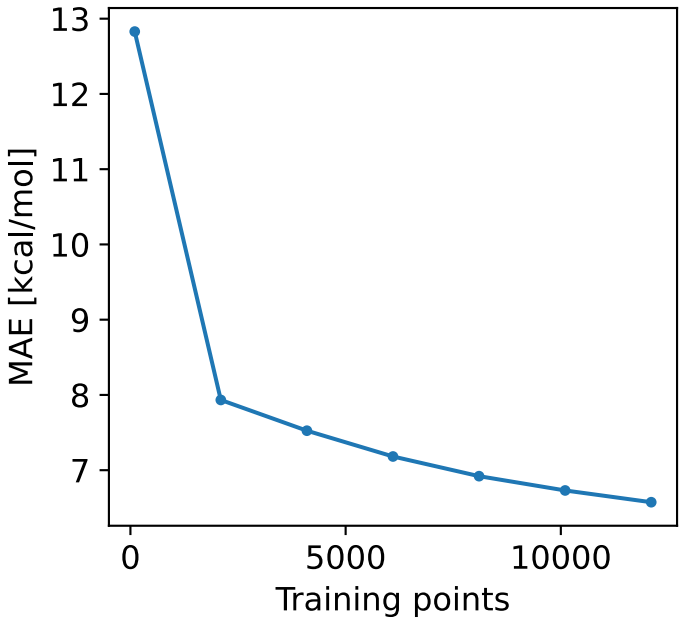

Fig. S11 Learning curve of a random forest model using L-MBTR as input with up to 12100 training points.

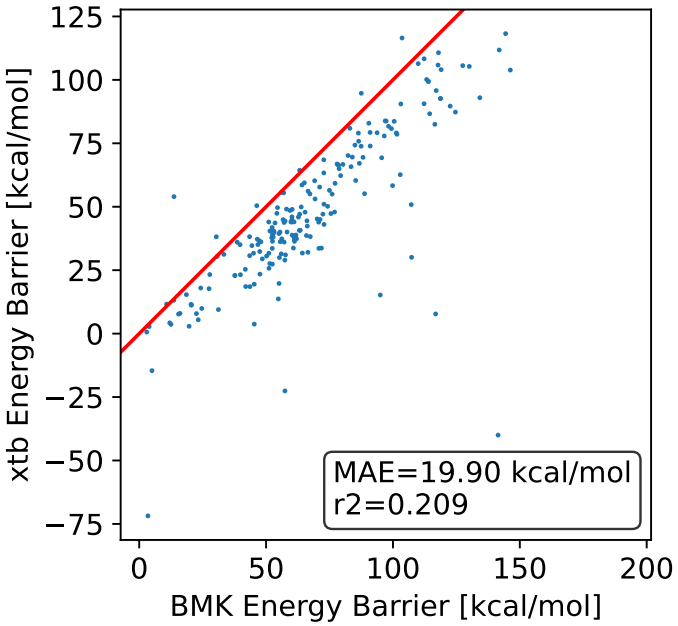

Fig. S12 Reaction barriers predicted with BMK vs. with GFN2-xTB[9]. 100 structures were drawn randomly from the combined synthetic and trajectory data set.

References

1 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2020, **49**, D1388–D1395.

2 A. Obarska-Kosinska, B. Rennekamp, A. Ünal and F. Gräter, *Biophys. J.*, 2021, **120**, 3544–3549.

3 B. Rennekamp, C. Karfusehr, M. Kurth, A. Ünal, D. Monego, K. Riedmiller, G. Gryn'ova, D. M. Hudson and F. Gräter, *Nature Communications*, 2023, **14**, 2075.

4 B. Rennekamp, F. Kutzki, A. Obarska-Kosinska, C. Zapp and F. Gräter, *Journal of Chemical Theory and Computation*, 2019, **16**, 553–563.

5 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

6 L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.

7 M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng, *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015, `https://www.tensorflow.org/`, Software available from tensorflow.org.

8 P. Reiser, A. Eberhard and P. Friederich, *Software Impacts*, 2021, **9**, 100095.

9 C. Bannwarth, S. Ehlert and S. Grimme, *Journal of Chemical Theory and Computation*, 2019, **15**, 1652–1671.