catalysts

Supplementary Materials: Conformational Landscapes of Halohydrin Dehalogenases and Their Accessible Active Site Tunnels

Miquel Estévez-Gay [1], Javier Iglesias-Fernández [1,3,*] and Sílvia Osuna [1,2,*]

[1] CompBioLab group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, c/Maria Aurèlia Capmany 69, 17003 Girona, Catalonia, Spain; miquel.estevez@udg.edu

[2] ICREA, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain

[3] Present address: Nostrum Biodiscovery, Carrer de Baldiri Reixac, 10–12, 08028, Barcelona, Catalonia, Spain

* Correspondence: jiglesiasfrn@gmail.com (J.I.-F.), silvia.osuna@udg.edu (S.O.)

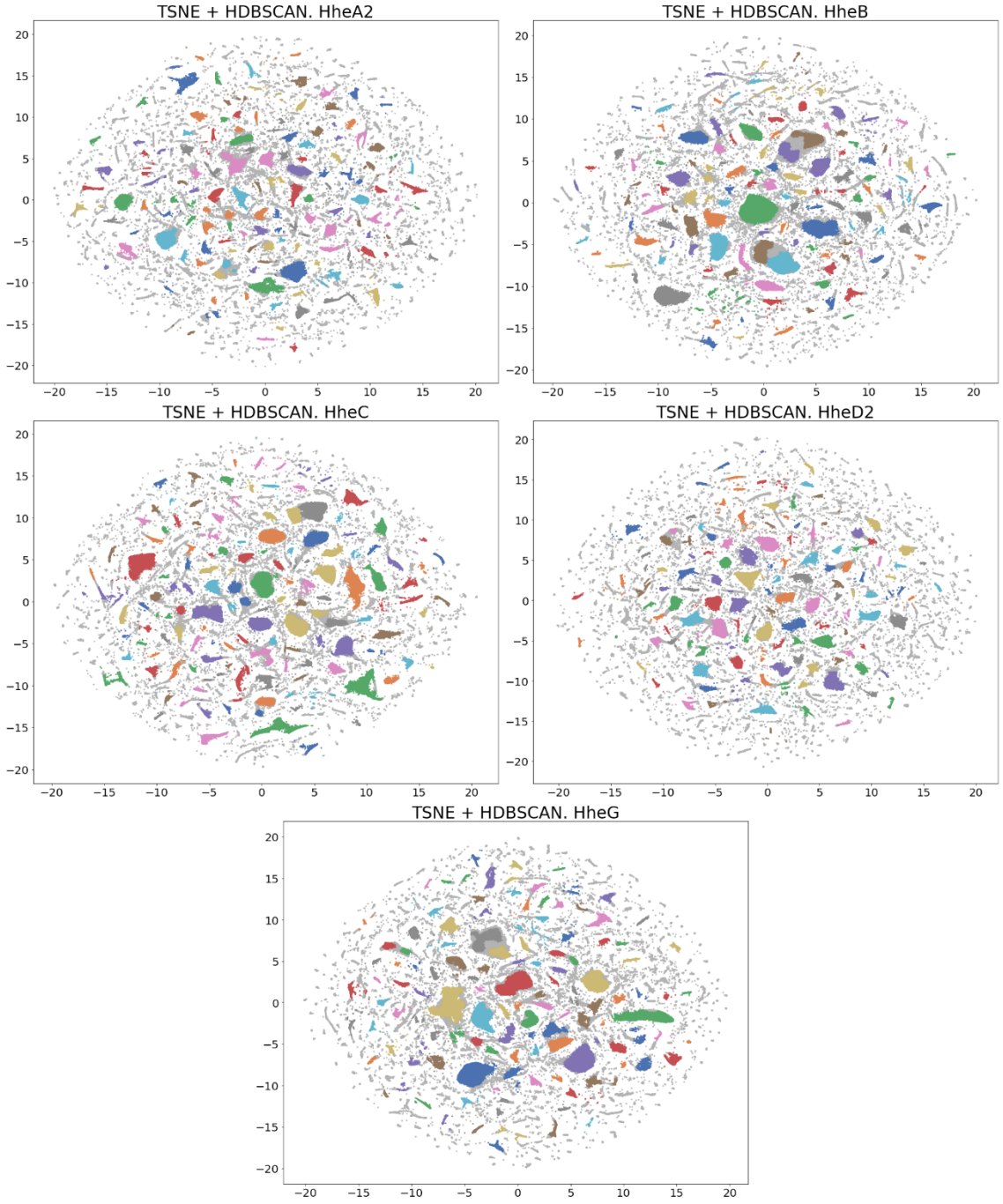

Figure S1. Generated t-SNE spaces [1] (one for each HHDH) using the 20th most kinetically-relevant tICA dimensions [2]. The t-SNE space is then clusterized with HDBscan algorithm [3]. Each cluster is colored differently on the t-SNE space. A representative structure from each cluster is then used for further tunnel analysis.

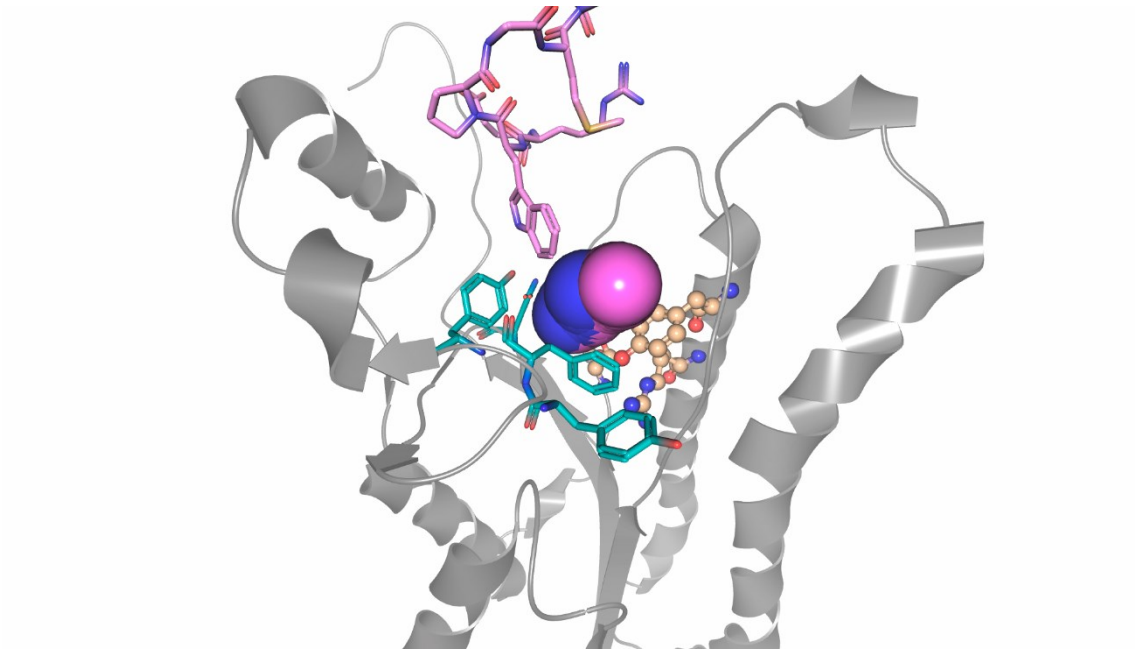

Figure S2. Tunnel T1 representation from the CaverAnalyst [4] calculations for the HheC X-ray structure considering either the C-terminal part of the neighbour monomer (results in purple) or without (i.e. monomeric structure shown in blue). T1 obtained in both cases are rather similar and the computed bottleneck radius (BR) are 2.28 and 2.30 Å for the structure with and without the C-terminal part, respectively.

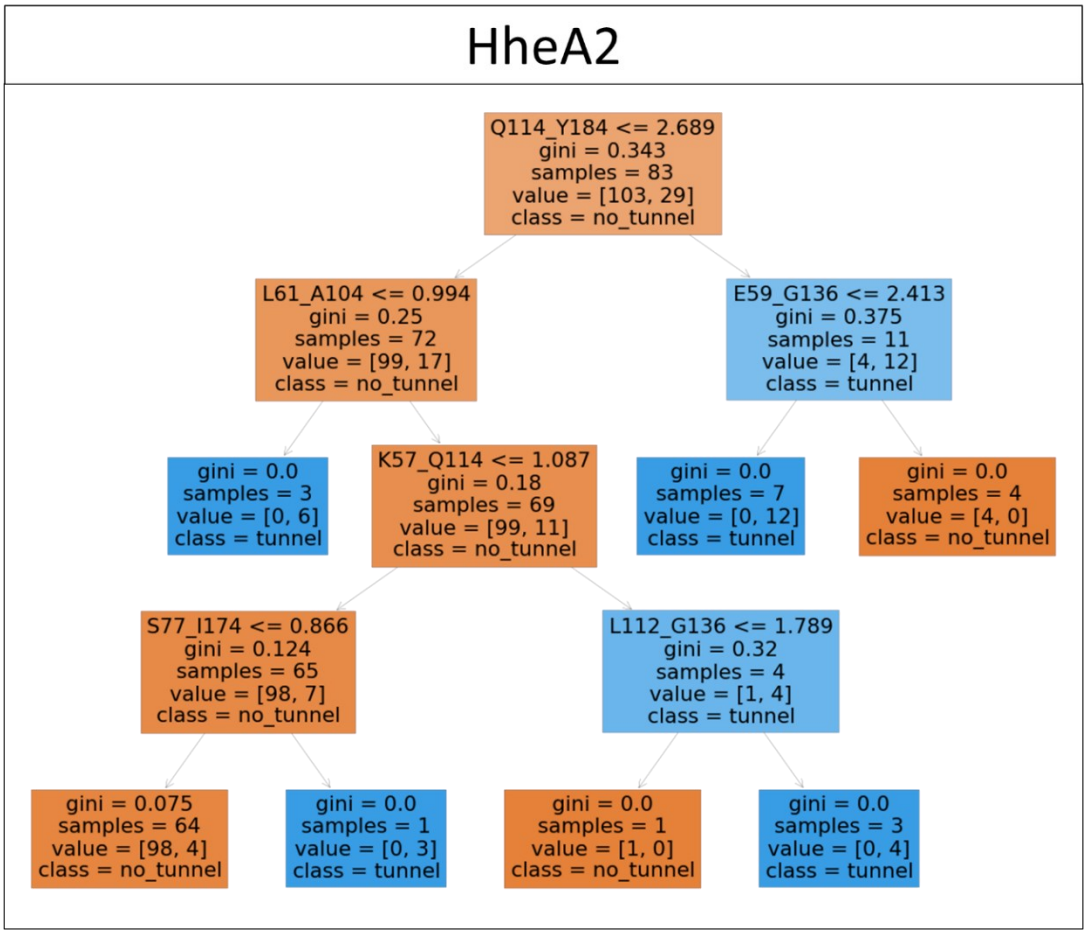

Figure S3. Schematic representation of a selected decision tree from the random forest classifier for HheA2. Input features are the most important "closest-heavy" distances involved in T2 formation (at

any point, computed using CaverAnalyst), whereas the target variable is the presence or absence (T/F) of the tunnel T2. After shuffling the dataset and splitting it in 80% training/ 20% prediction sets, the score (i. e. accuracy) is 0.88.

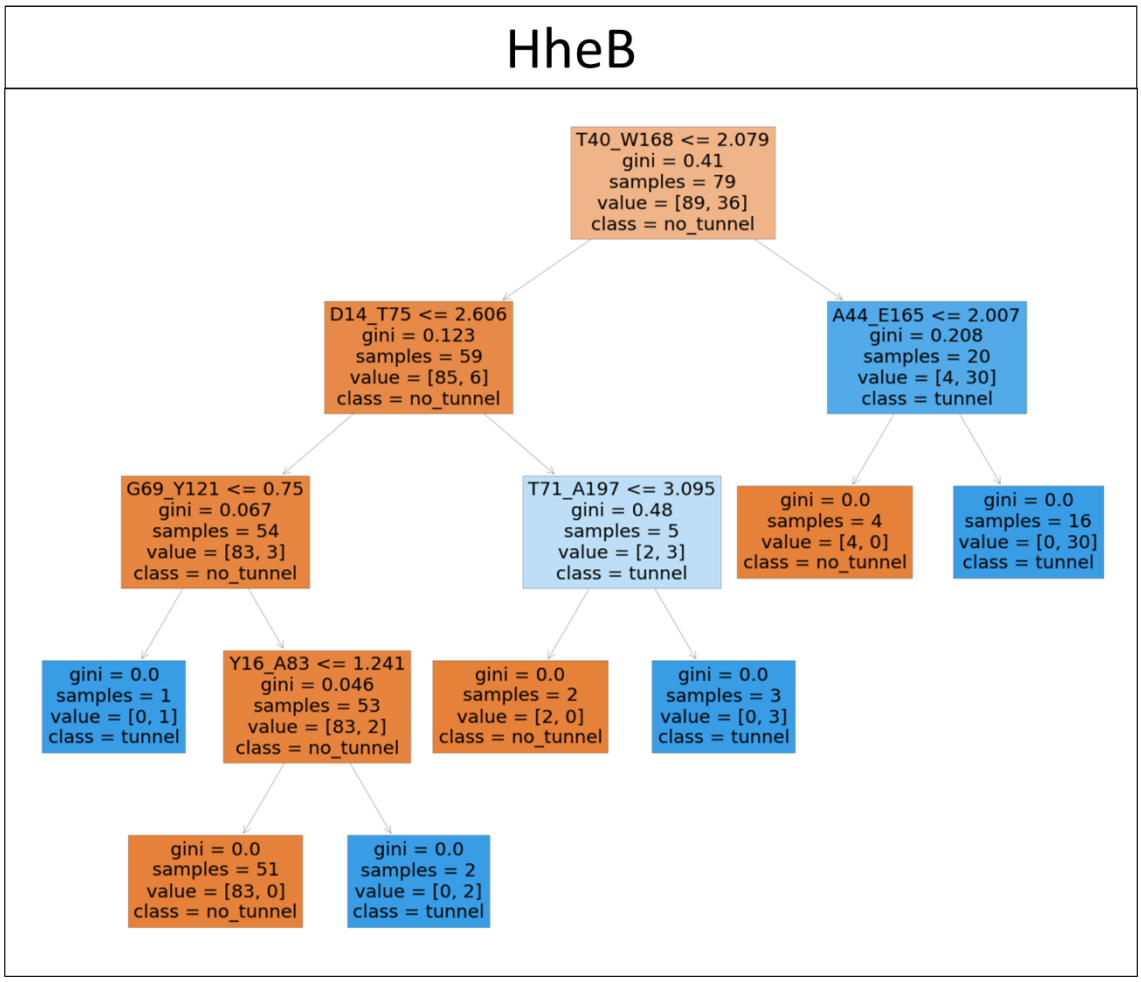

Figure S4. Schematic representation of a selected decision tree from the random forest classifier for HheB. Input features are the most important "closest-heavy" distances involved in T2 formation (at any point, computed using CaverAnalyst), whereas the target variable is the presence or absence (T/F) of the tunnel T2. After shuffling the dataset and splitting it in 80% training/ 20% prediction sets, the score (i. e. accuracy) is 0.94.

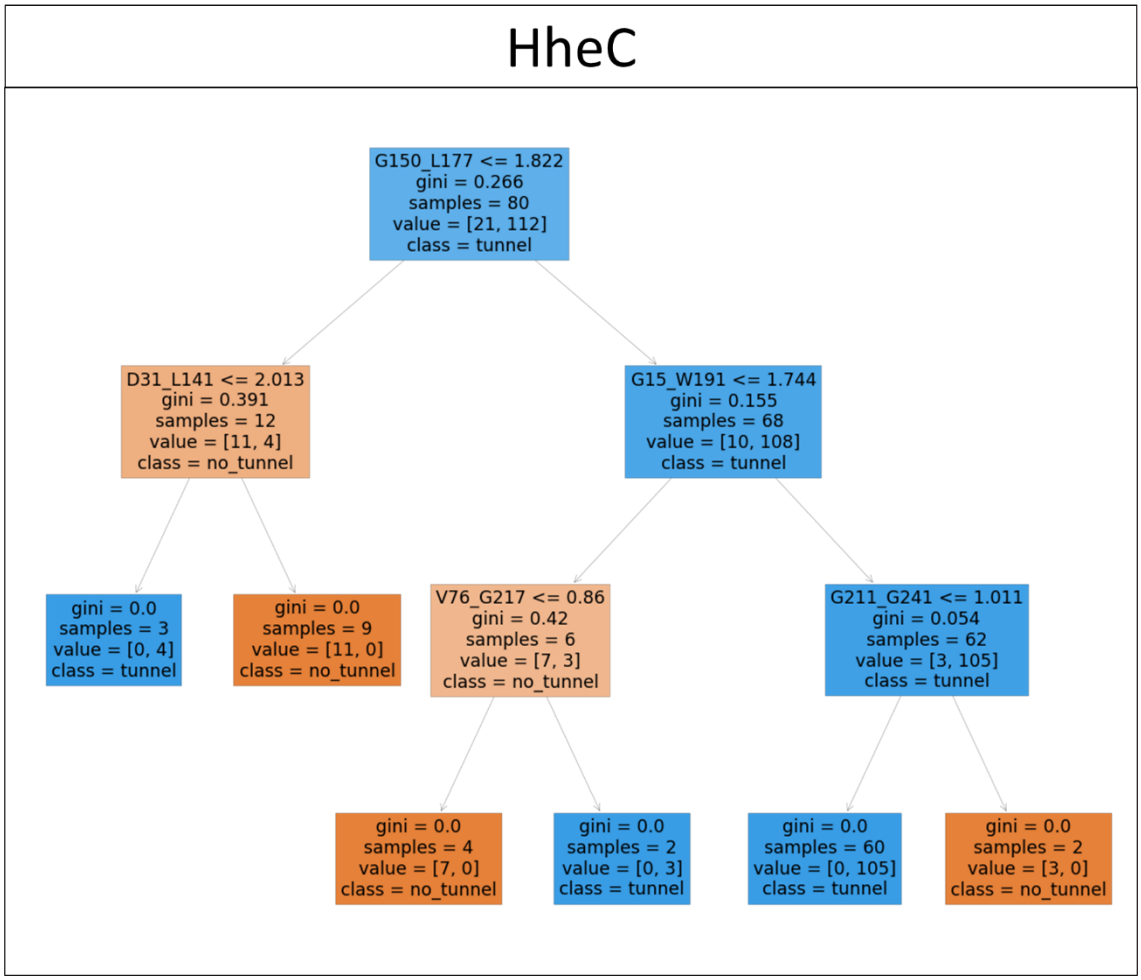

Figure S5. Schematic representation of a selected decision tree from the random forest classifier for HheC. Input features are the most important "closest-heavy" distances involved in T2 formation (at any point, computed using CaverAnalyst), whereas the target variable is the presence or absence (T/F) of the tunnel T2. After shuffling the dataset and splitting it in 80% training/ 20% prediction sets, the score (i. e. accuracy) is 0.83.

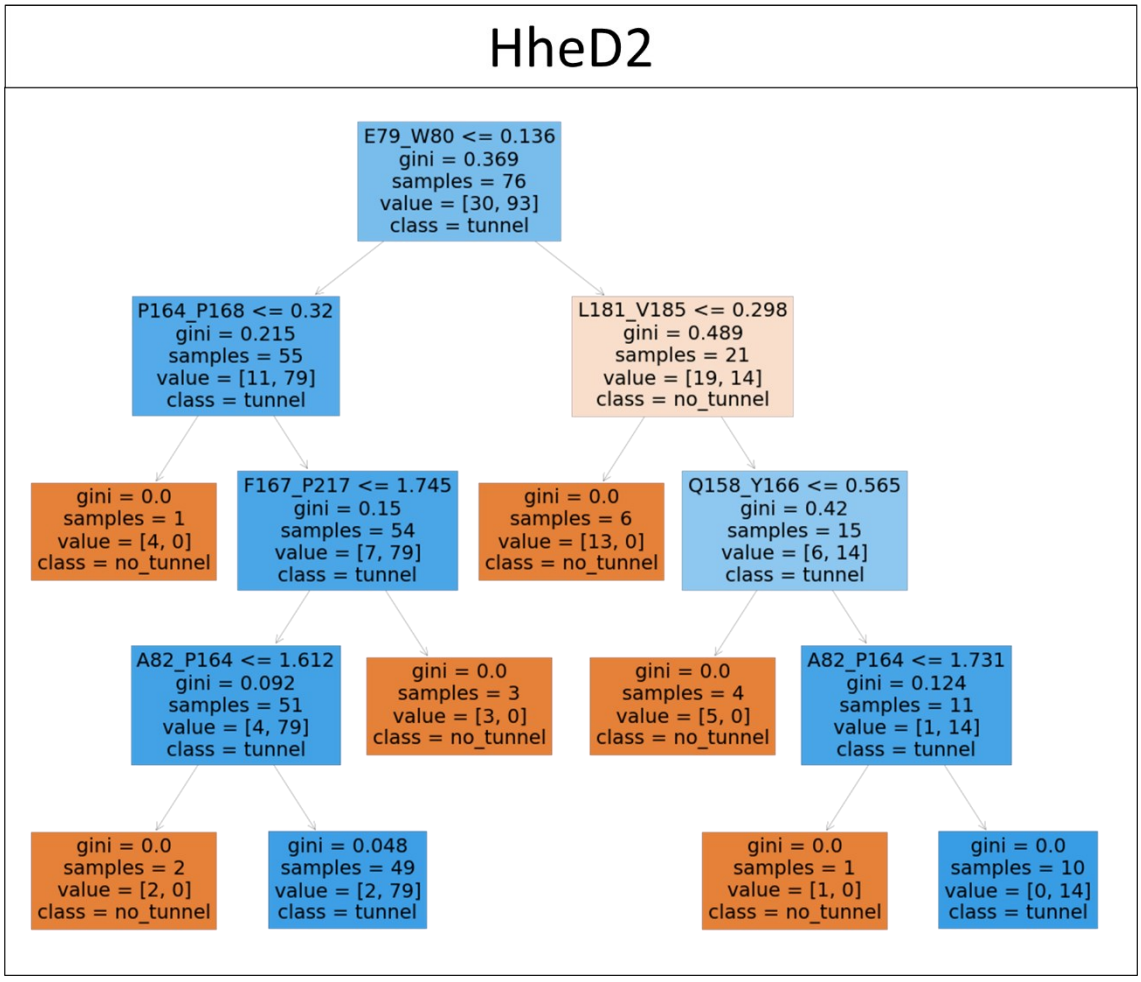

Figure S6. Schematic representation of a selected decision tree from the random forest classifier for HheD2. Input features are the most important "closest-heavy" distances involved in T2 formation (at any point, computed using CaverAnalyst), whereas the target variable is the presence or absence (T/F) of the tunnel T2. After shuffling the dataset and splitting it in 80% training/ 20% prediction sets, the score (i. e. accuracy) is 0.80.

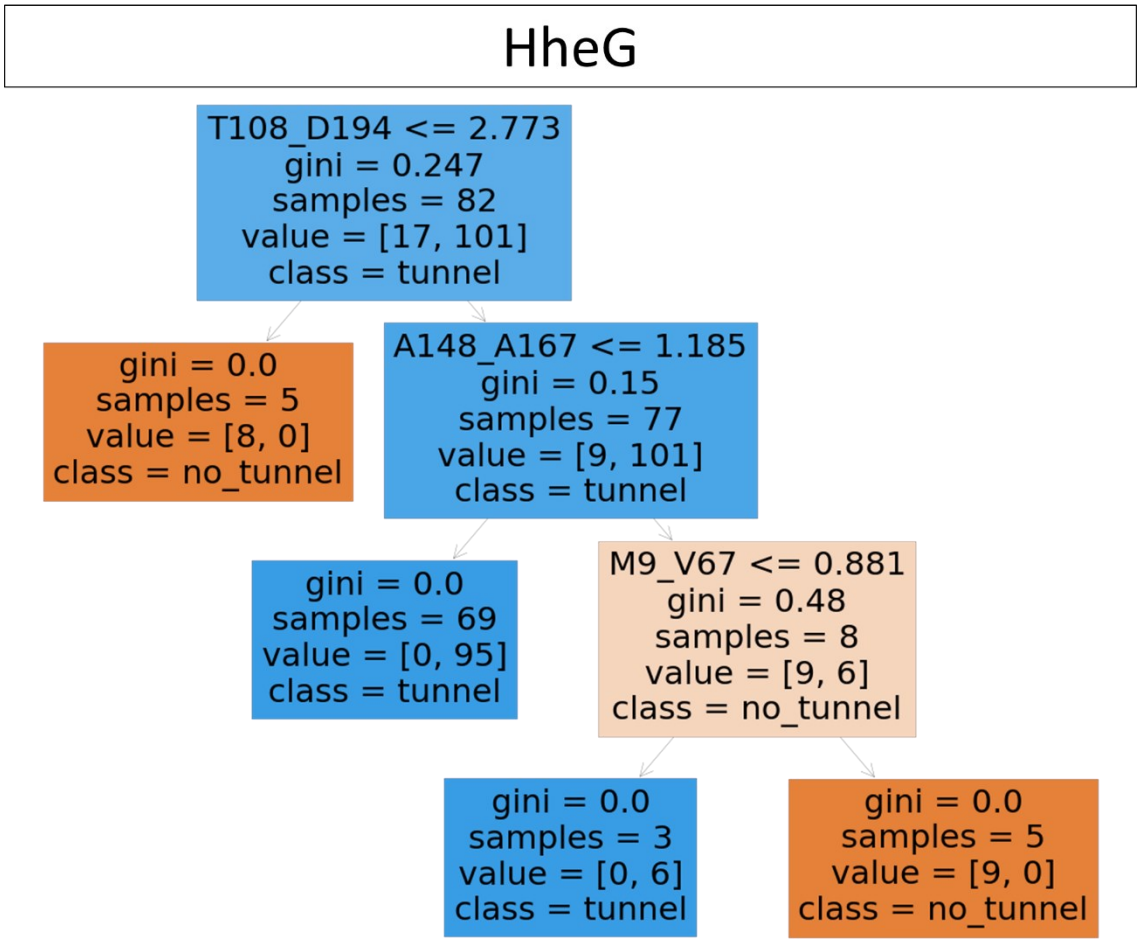

Figure S7. Schematic representation of a selected decision tree from the random forest classifier for HheG. Input features are the most important "closest-heavy" distances involved in T2 formation (at any point, computed using CaverAnalyst), whereas the target variable is the presence or absence (T/F) of the tunnel T2. After shuffling the dataset and splitting it in 80% training/ 20% prediction sets, the score (i. e. accuracy) is 0.90.

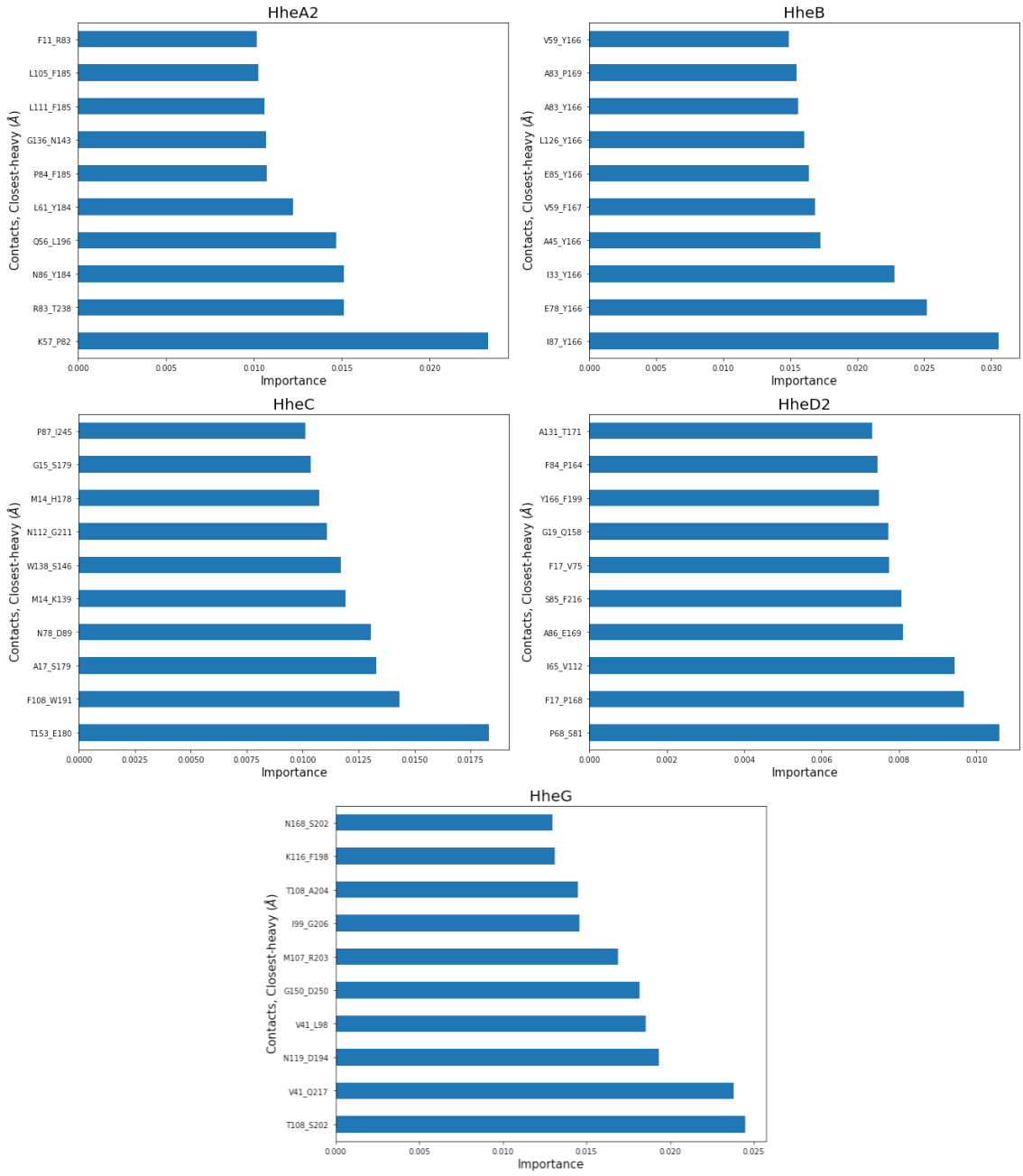

Figure S8. Bar plots showing for each HHDH studied the 10th most relevant features (Residue contacts based on closest-heavy distances) according to the feature importance extracted with a Random Forest Classifier.

References:

1. Zhou, H.; Wang, F.; Tao, P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *J. Chem. Theory Comput*. **2018**, *14*, 5499-5510, doi:10.1021/acs.jctc.8b00652.
2. Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102, doi:10.1063/1.4811489.
3. Campello, R.J.G.B.; Moulavi, D.; Sander, J. *Density-Based Clustering Based on Hierarchical Density Estimates;* Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.
4. Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P., et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* **2012**, *8*, e1002708, doi:10.1371/journal.pcbi.1002708.