Electronic Supplementary Information (ESI)

Specialising and Analysing Instruction-Tuned and Byte Level Language Models for Organic Reaction Prediction

Jiayun Pang,*[a] and Ivan Vulić,*[b]

a. *School of Science, Faculty of Engineering and Science, University of Greenwich Medway Campus, Central Avenue, Chatham Maritime, ME4 3RL, United Kingdom*
Email: j.pang@gre.ac.uk
b. Language Technology Lab, University of Cambridge, 9 West Road, Cambridge CB3 9DA, UK. E-mail: iv250@cam.ac.uk

Contents: Further examples of SHAP analysis

1. Further examples of SHAP analysis.

1) Further to the generation of ketoxime from ketone example we provide in the main text, we have carried out SHAP analysis on two additional reactions. We randomly selected the reactions with more complex reaction conditions (e.g. reagents) and product structure.

The first reaction is a reductive amination. The reactants are a secondary amine and a ketone. The presence of acetic acid catalyses the imine formation and the sodium triacetoxyborohydride, a mild reducing agent is used to reduce the iminium ions (Figure S1 and S2).

In the FlanT5 SHAP figure (Figure S3), there is strong positive impact from CH_3COOH to the formation of the fluorine-substituted aromatic ring (one of the reactants) in the product. This is interesting as CH_3COOH acts as a catalyst to the formation of the iminium ion intermediate. However, the ByT5 model does not reflect this feature (Figure S4). Neither the FlanT5 nor the ByT5 model recognise the importance of triacetoxyborohydride in the reaction.

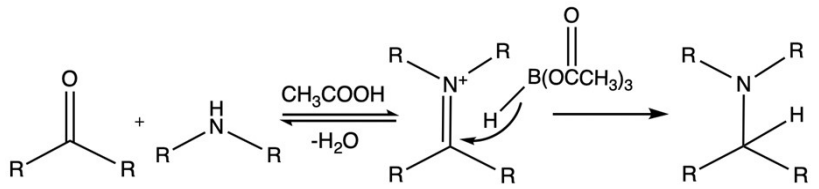

Figure S1: The reaction mechanism of reductive amination.

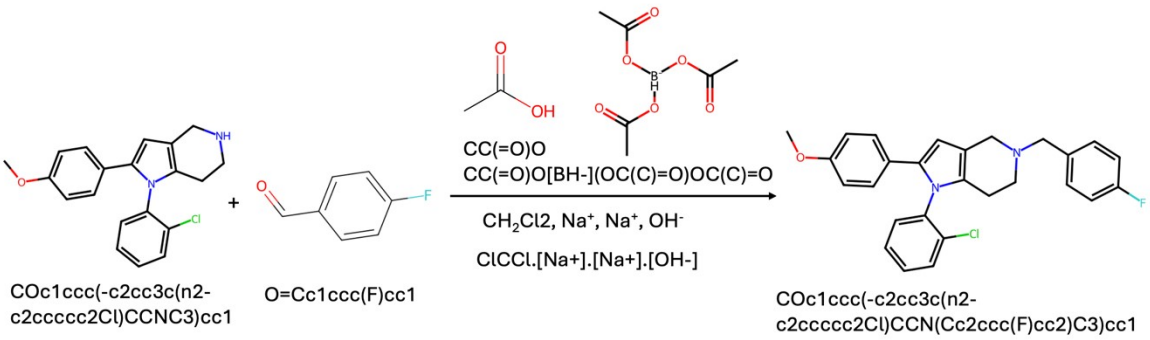

Figure S2: The reaction with SMILES of the reactants, reagents and product.

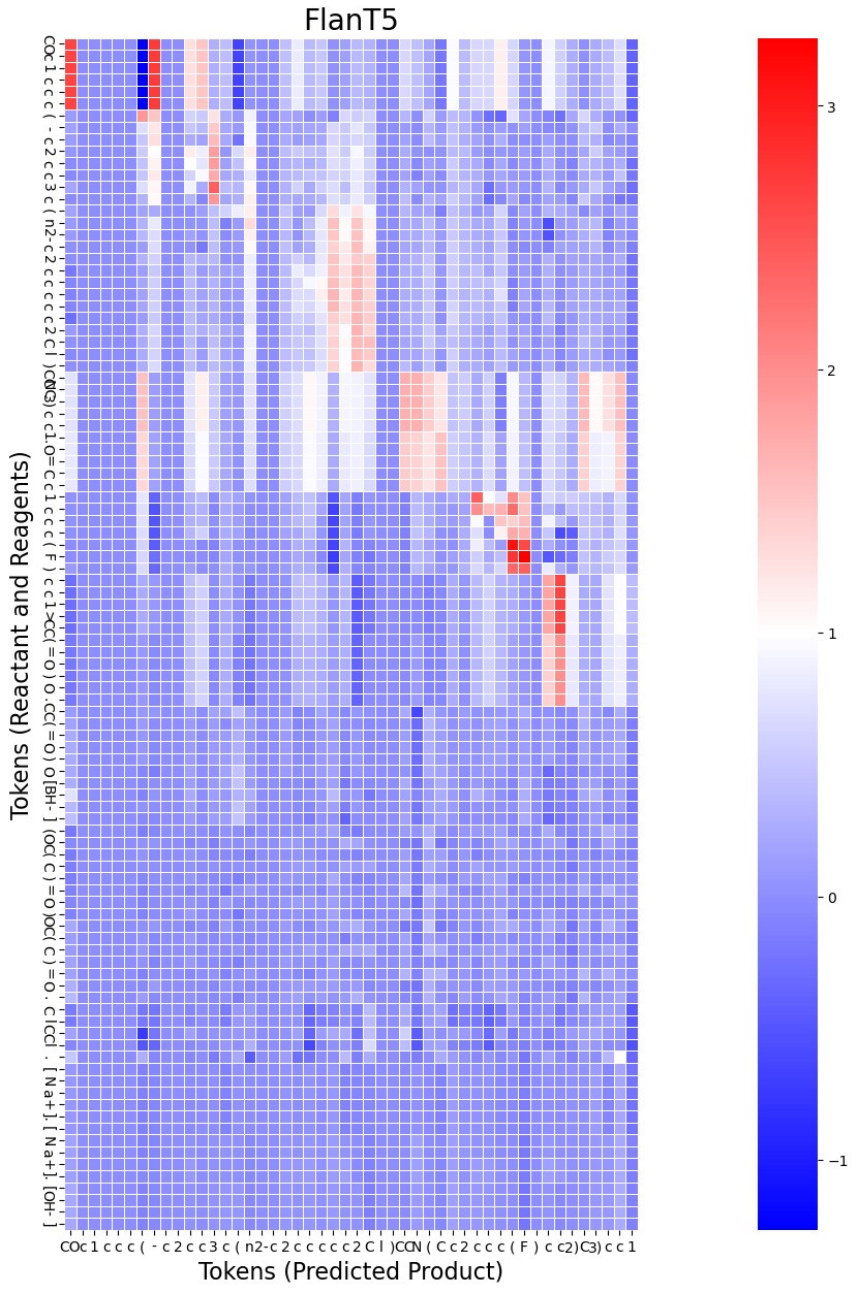

Figure S3: Computed Shapley values for the reductive amination reaction using the FlanT5 model

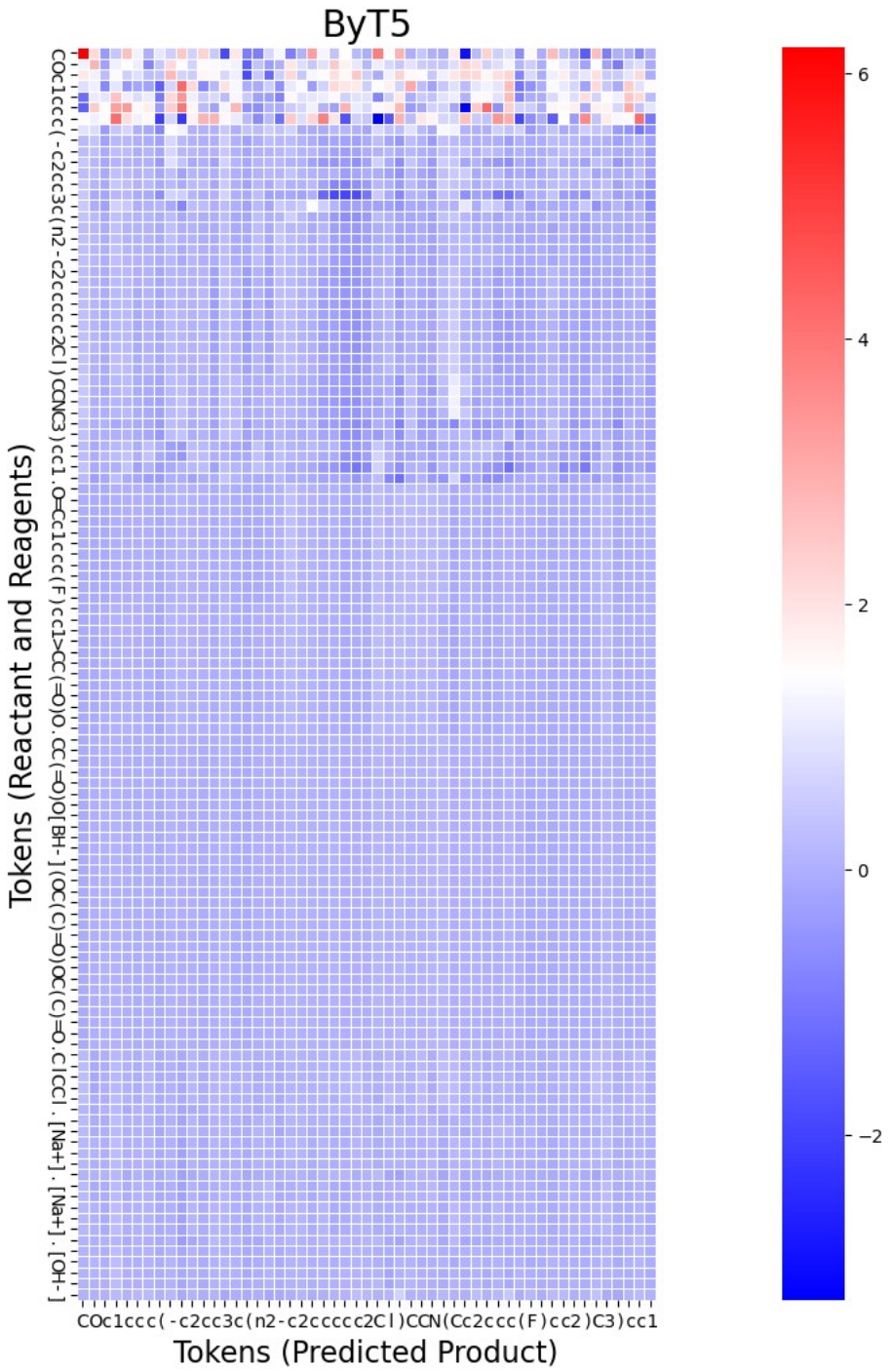

Figure S4: Computed Shapley values for the reductive amination reaction using the ByT5 model.

2) The second reaction is the reduction of ketone to secondary alcohol (Figure S5 and S6). For this reaction, Sodium Borohydride ($NaBH_4$) is the source of hydride ion (H^-) for the reduction. The SHAP analysis of FlanT5 model displays strong positive contribution from reagents BH_4^- and CH_3OH to the reaction, although they are mostly associated with the part of the product that is not the reaction centre (i.e. the newly formed OH) (Figure S7). The ByT5 model does not pick up significant contribution of relevant reagents based on their SHAP values (Figure S8).

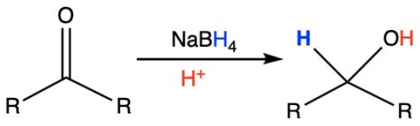

Figure S5: The reaction mechanism of the reduction of ketone to primary alcohol.

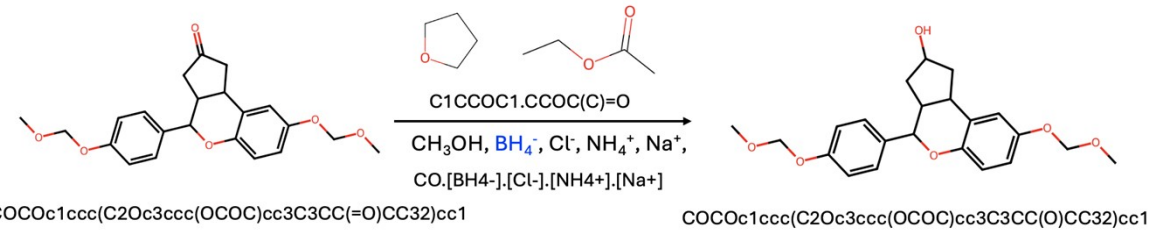

Figure S6: The reaction with SMILES of the reactants, reagents and product.

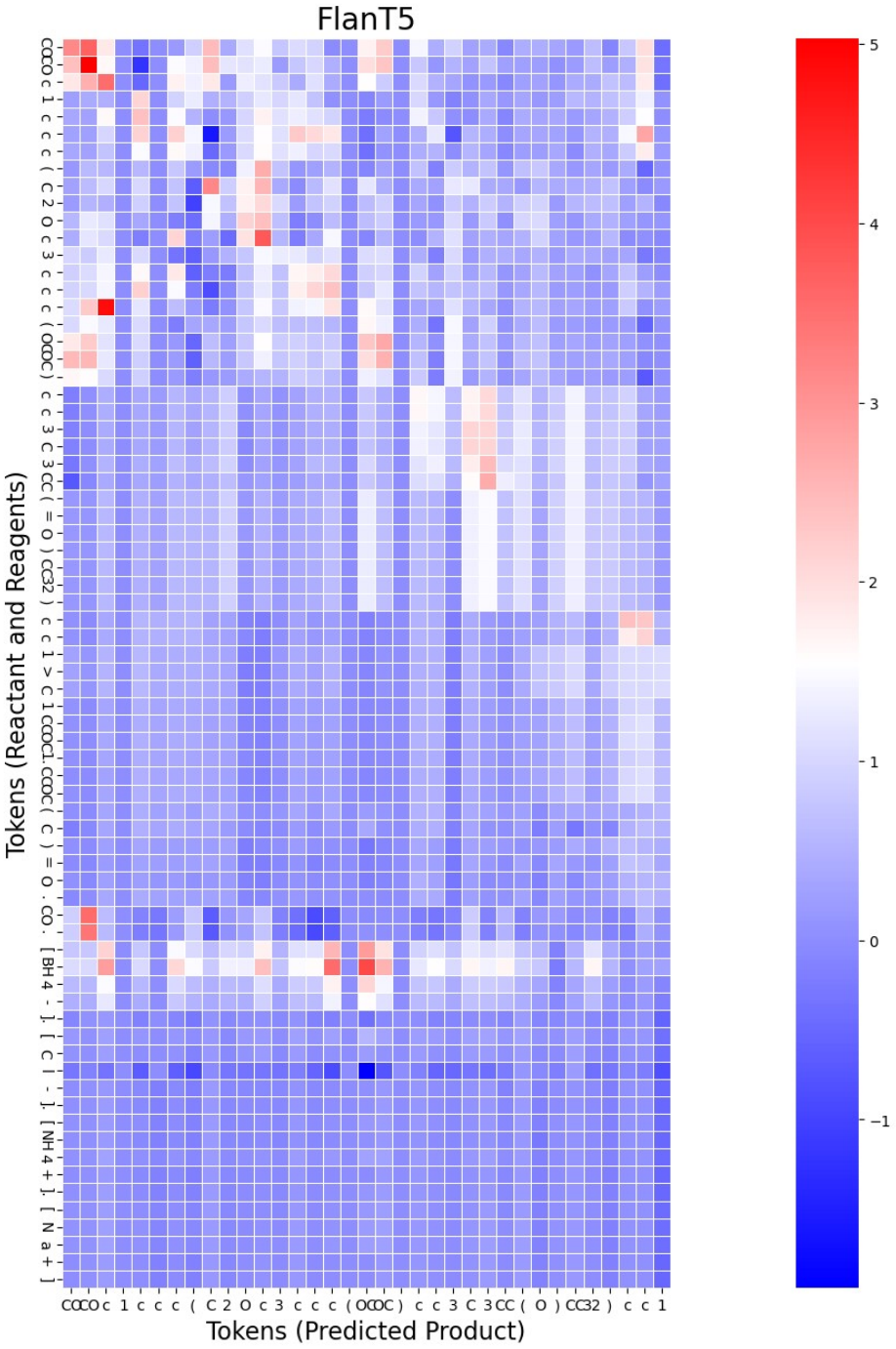

Figure S7: Computed Shapley values for the reduction of ketone to secondary alcohol using the FlanT5 model.

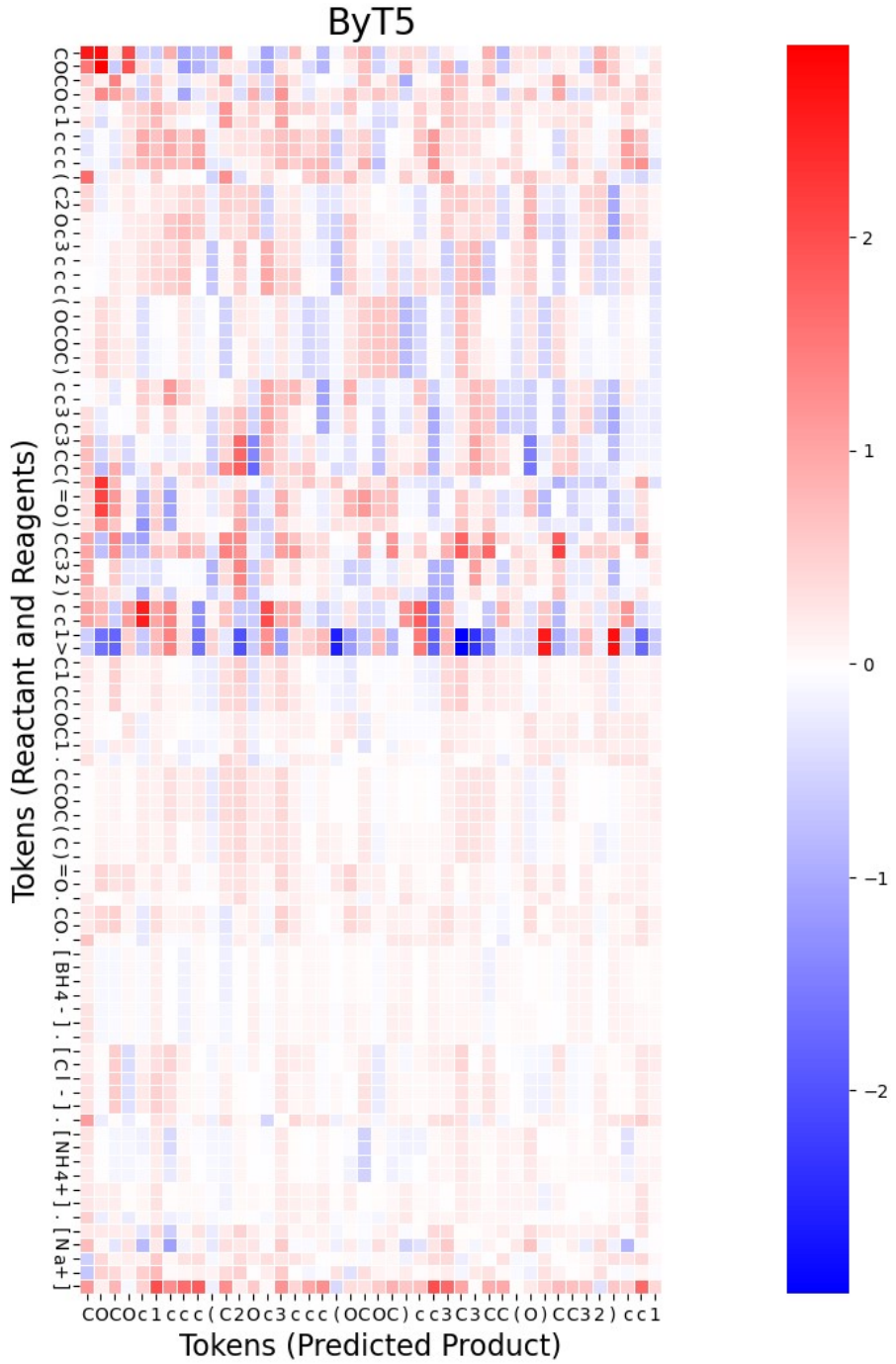

Figure S8: Computed Shapley values for the reduction of ketone to secondary alcohol using the ByT5 model.