Big Data from Pharmaceutical Patents:

A Computational Analysis of Medicinal Chemists' Bread and Butter

Nadine Schneider[1a], Daniel M. Lowe[2], Roger A. Sayle[2], Michael A. Tarselli[1b], Gregory A. Landrum[1a]*

[1a]Novartis Institutes for BioMedical Research, Novartis Pharma AG, Novartis Campus, 4002 Basel, Switzerland

[1b] Novartis Institutes for BioMedical Research, 186 Massachusetts Avenue, Cambridge, MA 02139, USA

[2]NextMove Software Ltd, Innovation Centre, Unit 23, Science Park, Milton Road, Cambridge CB4 0EY, UK

Preparation and analysis of the SureChEMBL dataset

The SureChEMBL dataset we used in our study is also composed of molecules extracted from 1,568,011 granted US patents and US patent applications from the years between 1976 and October 2015 (applications here also only date back to 2001). In order to make it comparable to our reaction dataset, we only considered molecules extracted from the text of the claims and description sections of the patents. Note that our reaction-patent dataset is less than one tenth the size of the patent set from SureChEMBL. This is due to the extraction of whole reactions and not only molecules and due to the set of quality filters applied to the extracted data (details are given in the methods section). Additionally, in the SureChEMBL set molecules were extracted from patents from three different areas - Chemistry, Physics, and Human Necessities – while our dataset was restricted to patents only from pharmaceutical research.

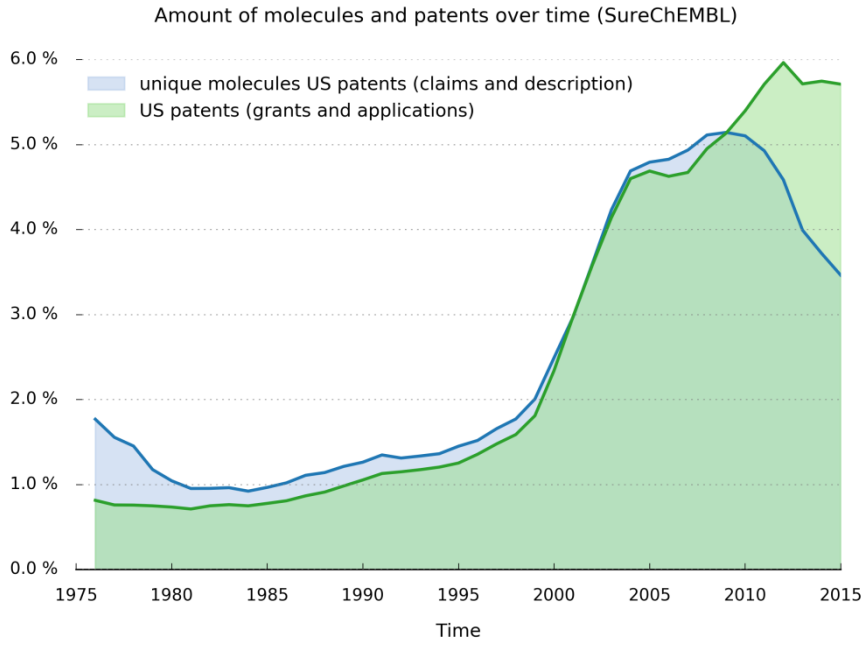


Figure S1: Distribution of unique molecules extracted from US patents and US patents per year in the time period comprising the years from 1976 until October 2015. This data was extracted from the SureChEMBL database (https://www.surechembl.org/search/). In light blue unique molecules extracted from the claims and description section of US patents (grants and applications) are shown. In light green US patents (grants and applications) are shown. Please note that the values in the plot were smoothed/averaged over 5 years for better visual clarity.

Evolution of the number of different reaction types over time

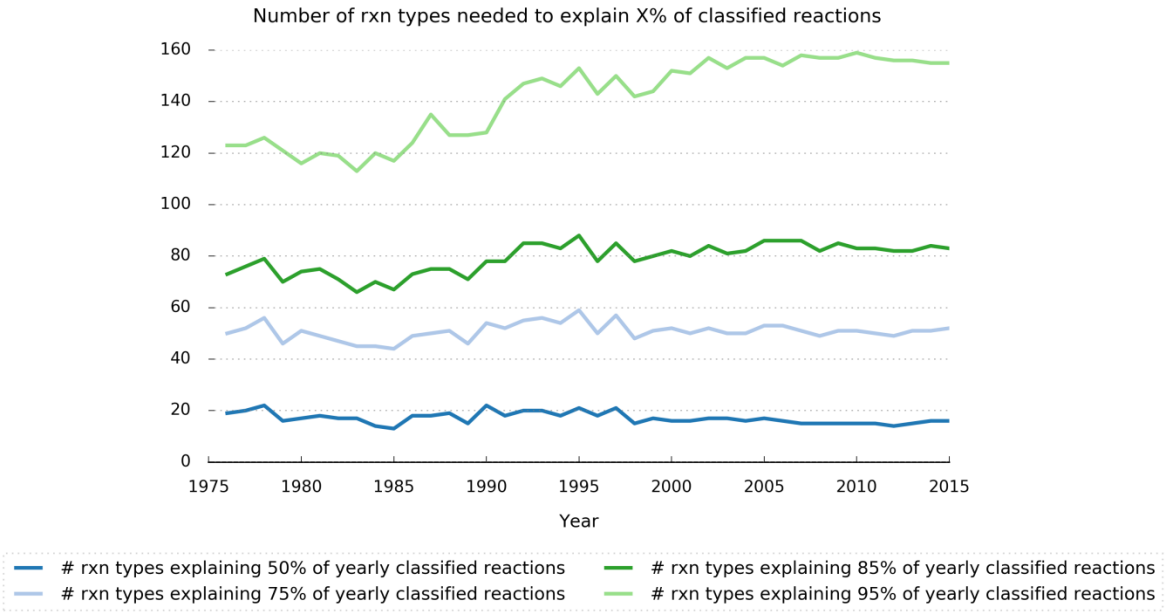


Figure S2: Evolution of the number of different reaction types being actively used over time. The number of reaction types needed to cover 50%, 75%, 85% and 95% of the classified reactions per year is shown.

Evolution of major reaction classes over time

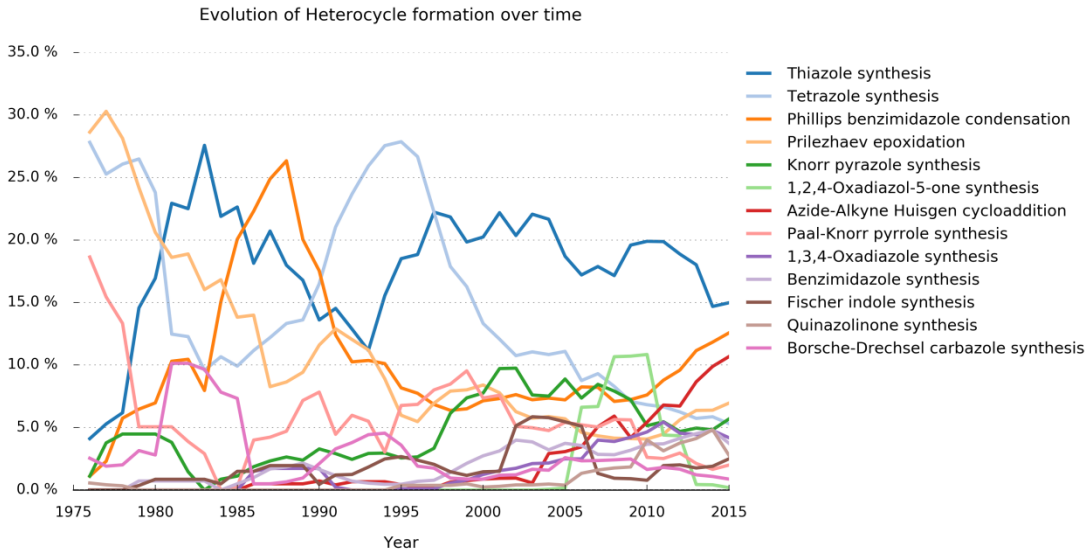


Figure S3: Evolution of *Heterocycle formation* over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

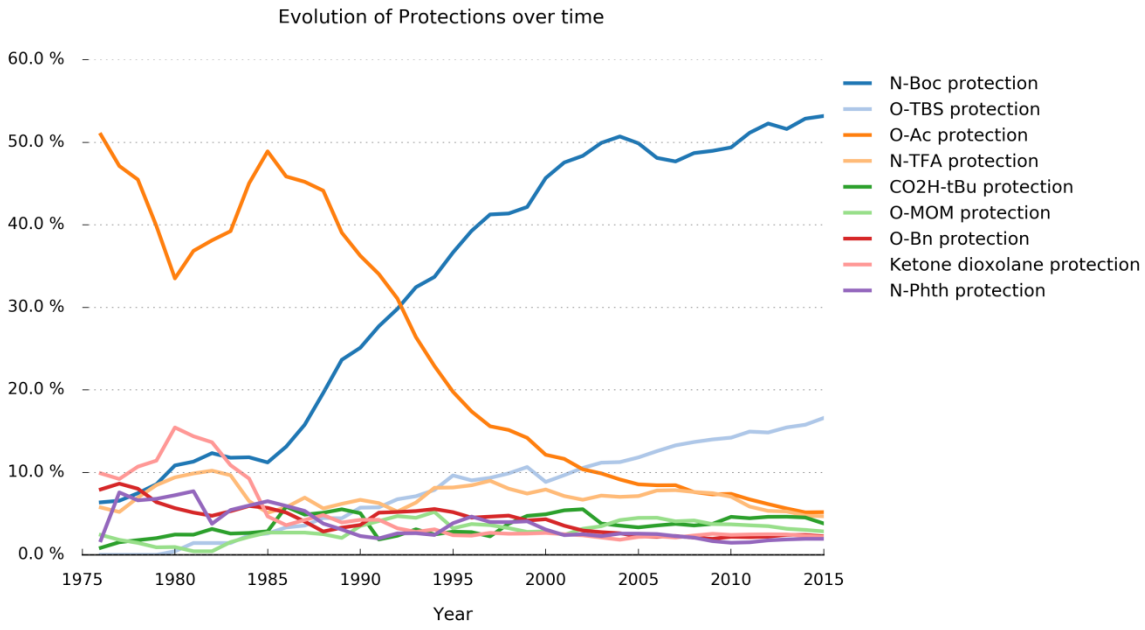


Figure S4: Evolution of *Protection reactions* over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

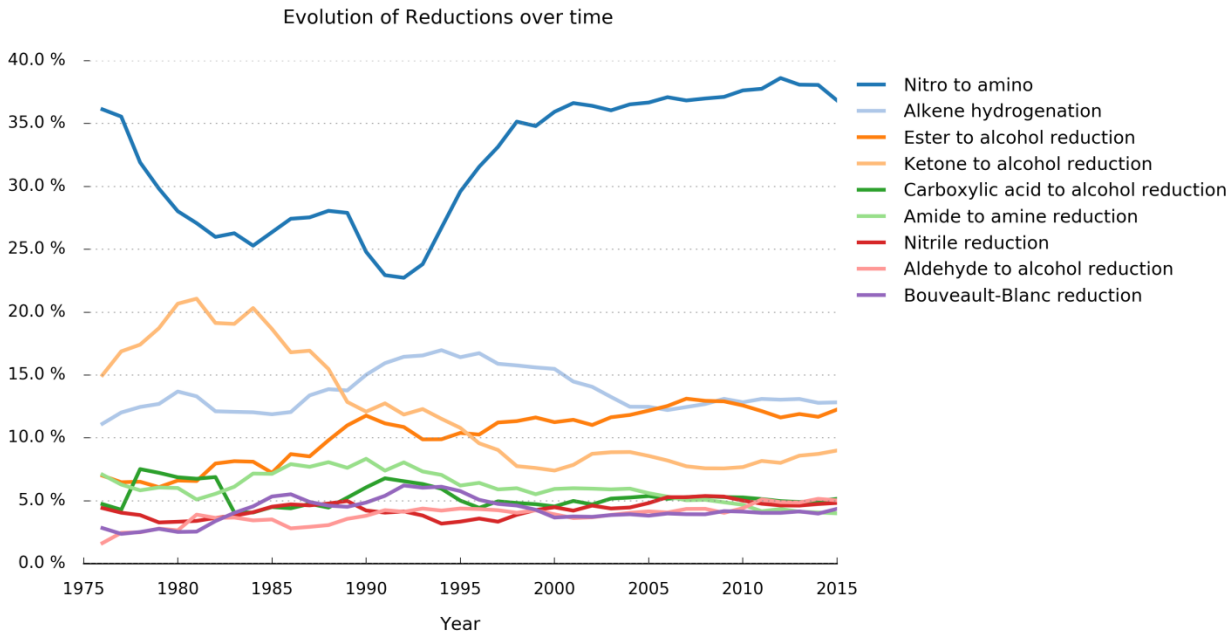


Figure S5: Evolution of *Reduction reactions* over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

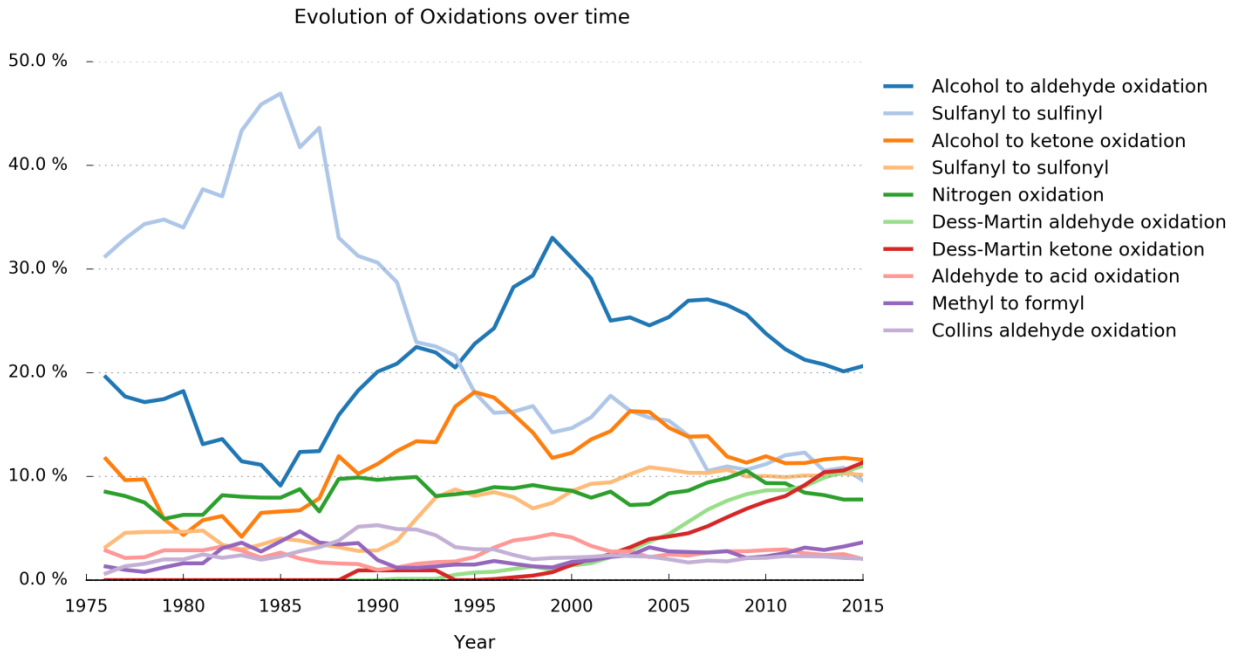


Figure S6: Evolution of *Oxidation reactions* formation over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

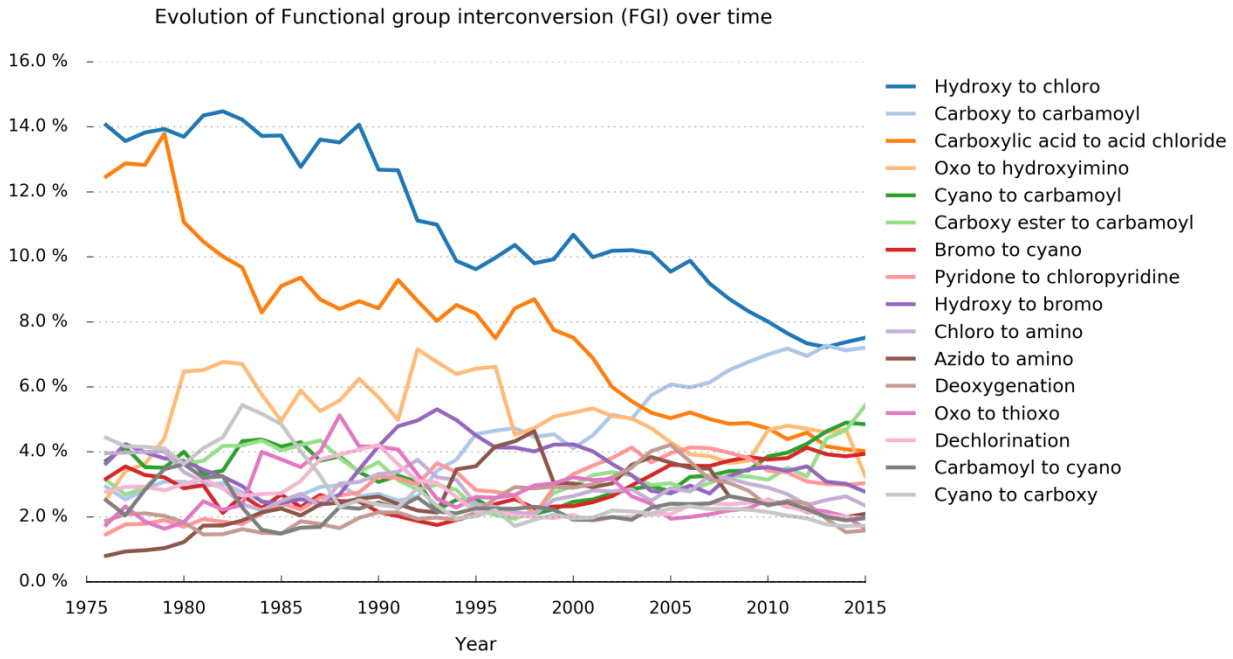


Figure S7: Evolution of *Functional group interconversions* over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

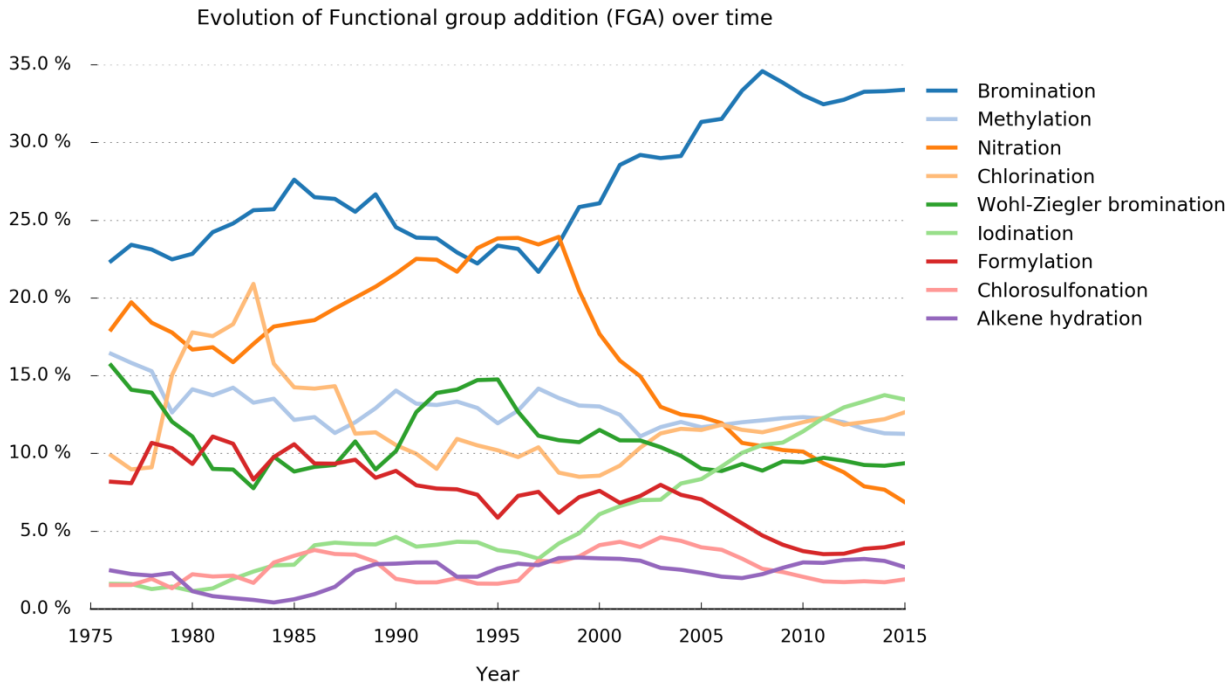


Figure S8: Evolution of *Functional group additions* over time. In the plot reaction types which at least represent 2% of the data were shown. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

The following plots also show rarely used reaction types of the major reaction classes.

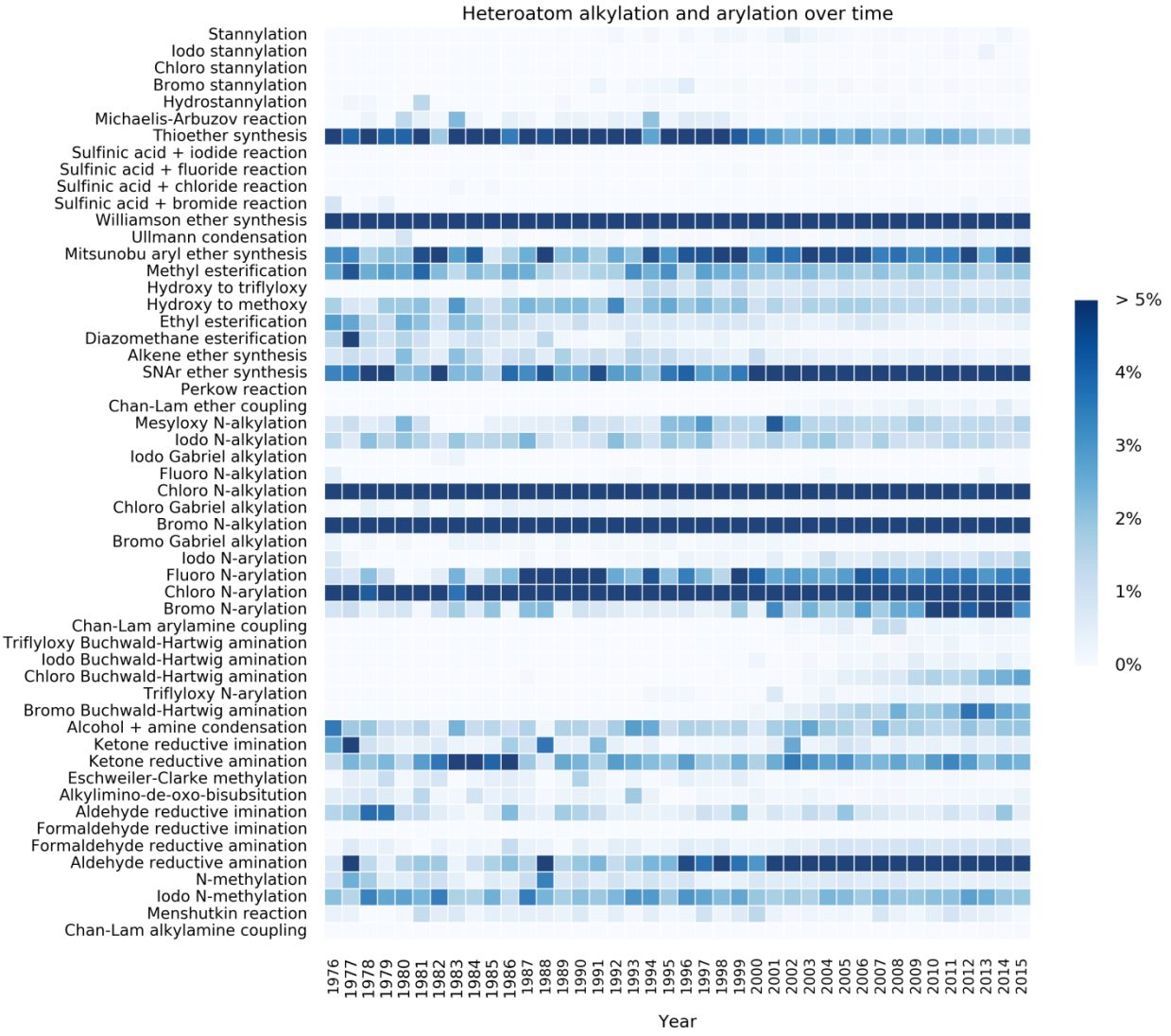


Figure S9: Evolution of *Heteroatom alkylations and arylations* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

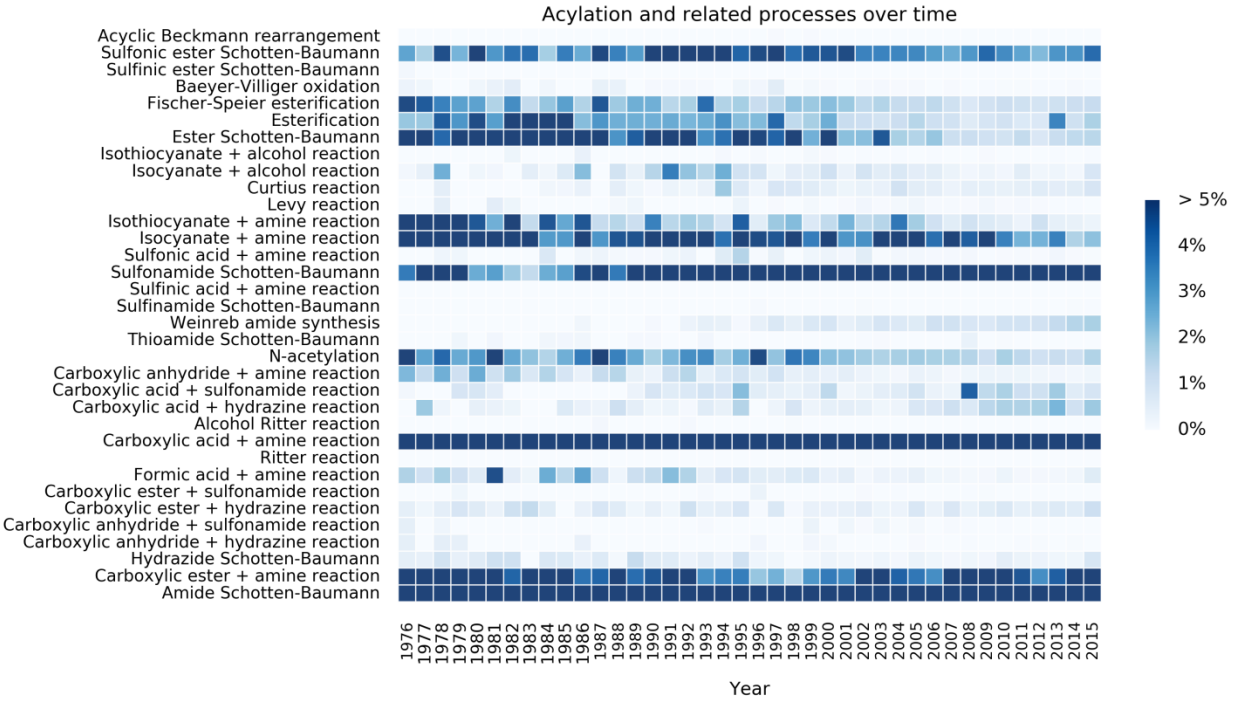


Figure S10: Evolution of *Acylation and related processes* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

Figure S11: Evolution of *C-C bond formations* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

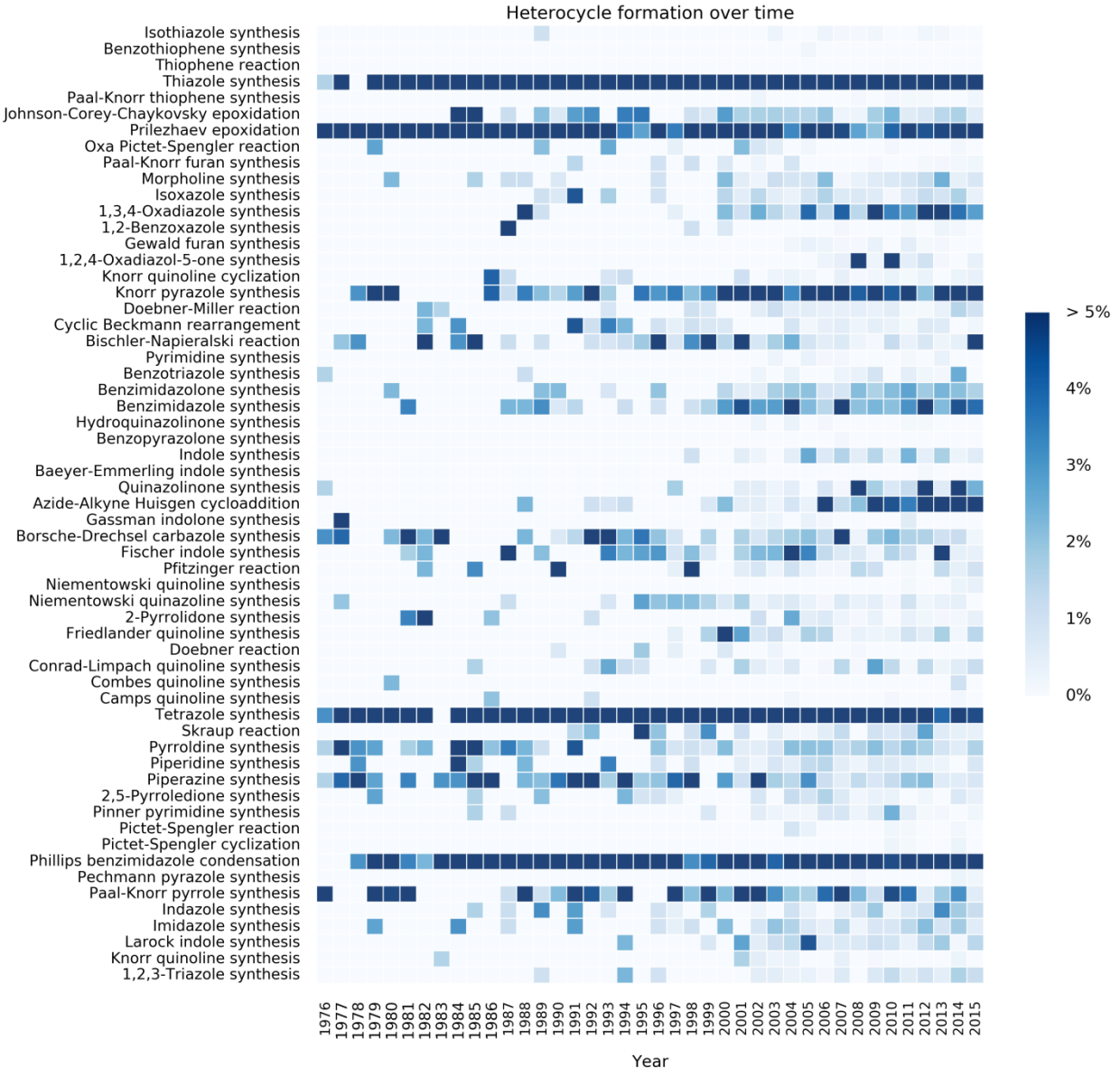


Figure S12: Evolution of *Heterocycle formation* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

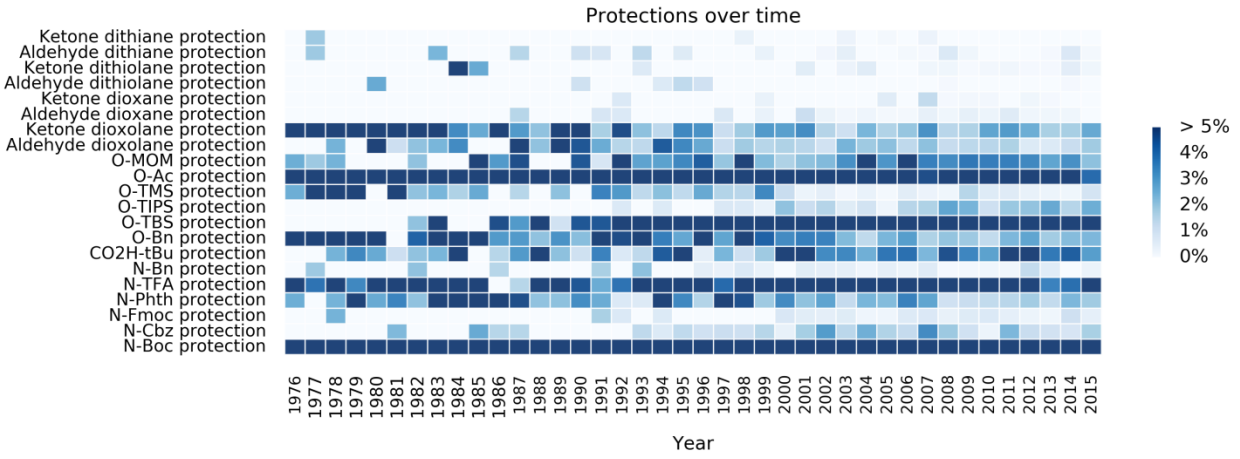


Figure S13: Evolution of *Protections* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

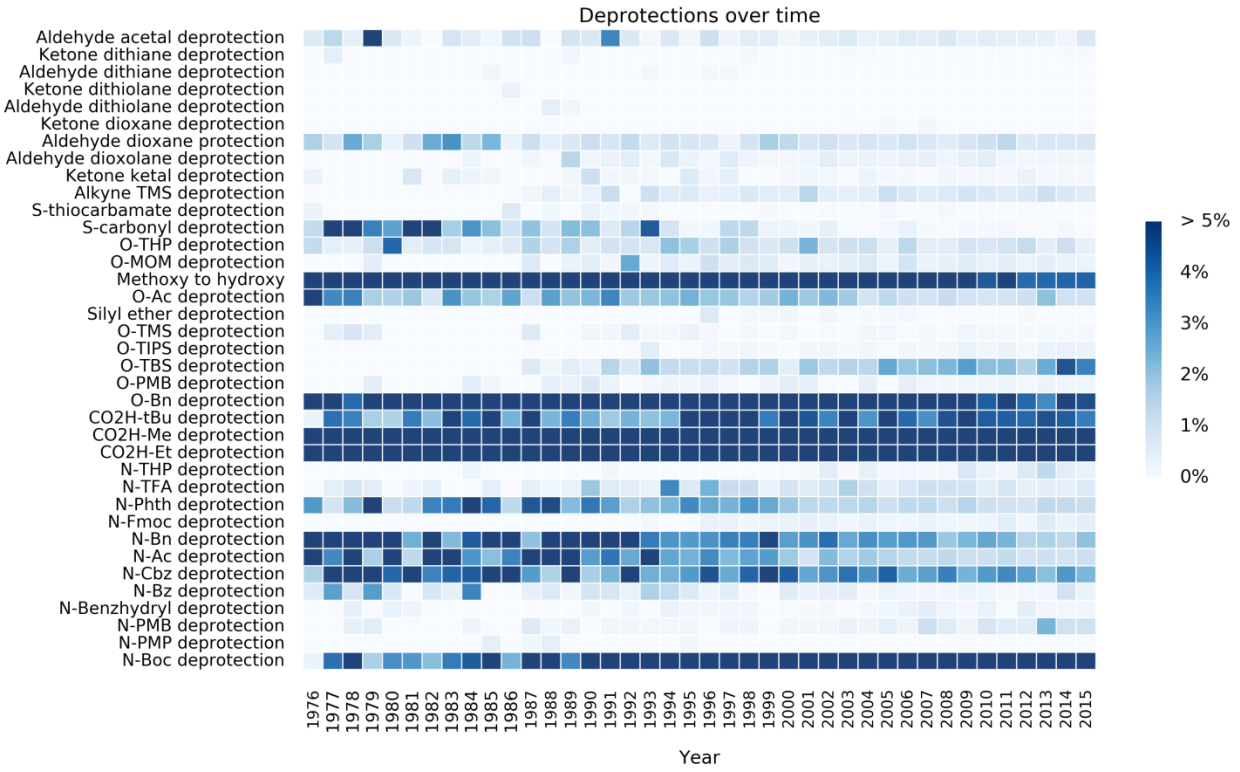


Figure S14: Evolution of *Deprotections* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

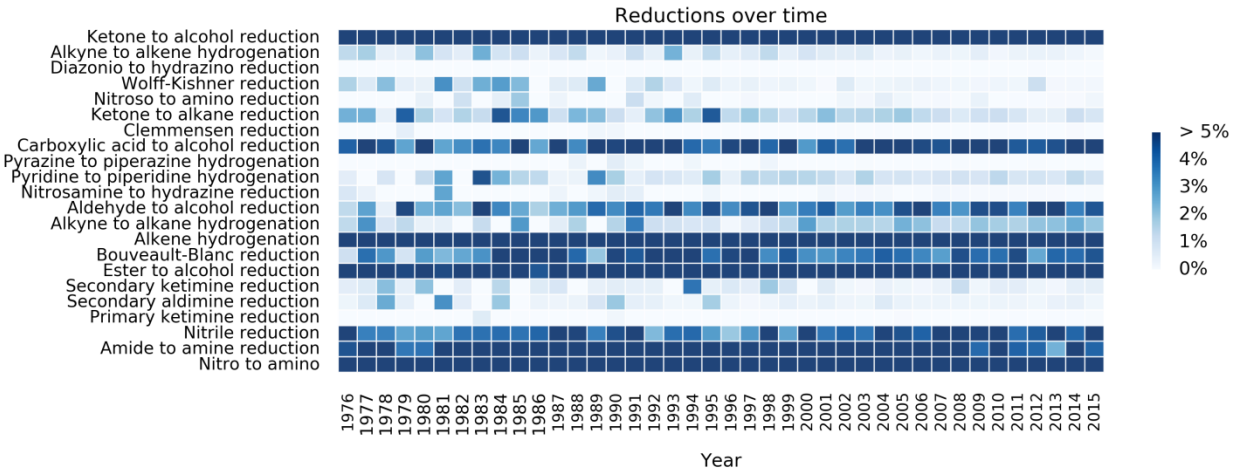


Figure S15: Evolution of *Reduction reactions* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

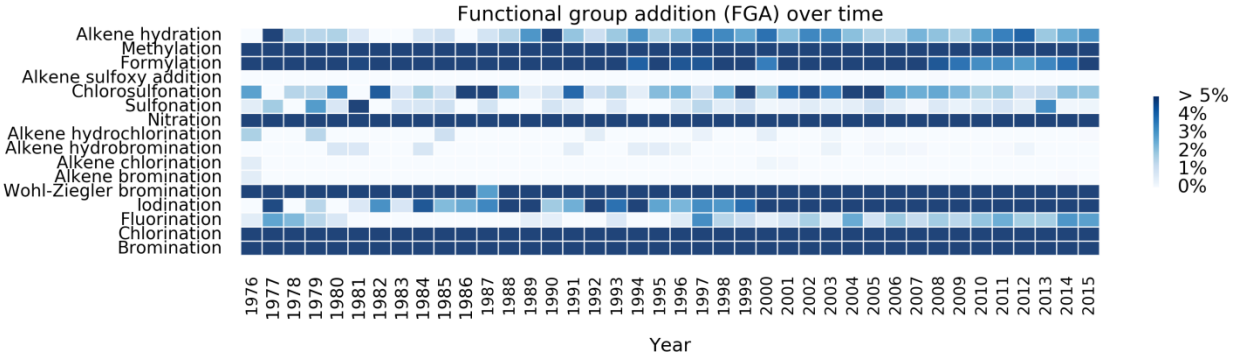
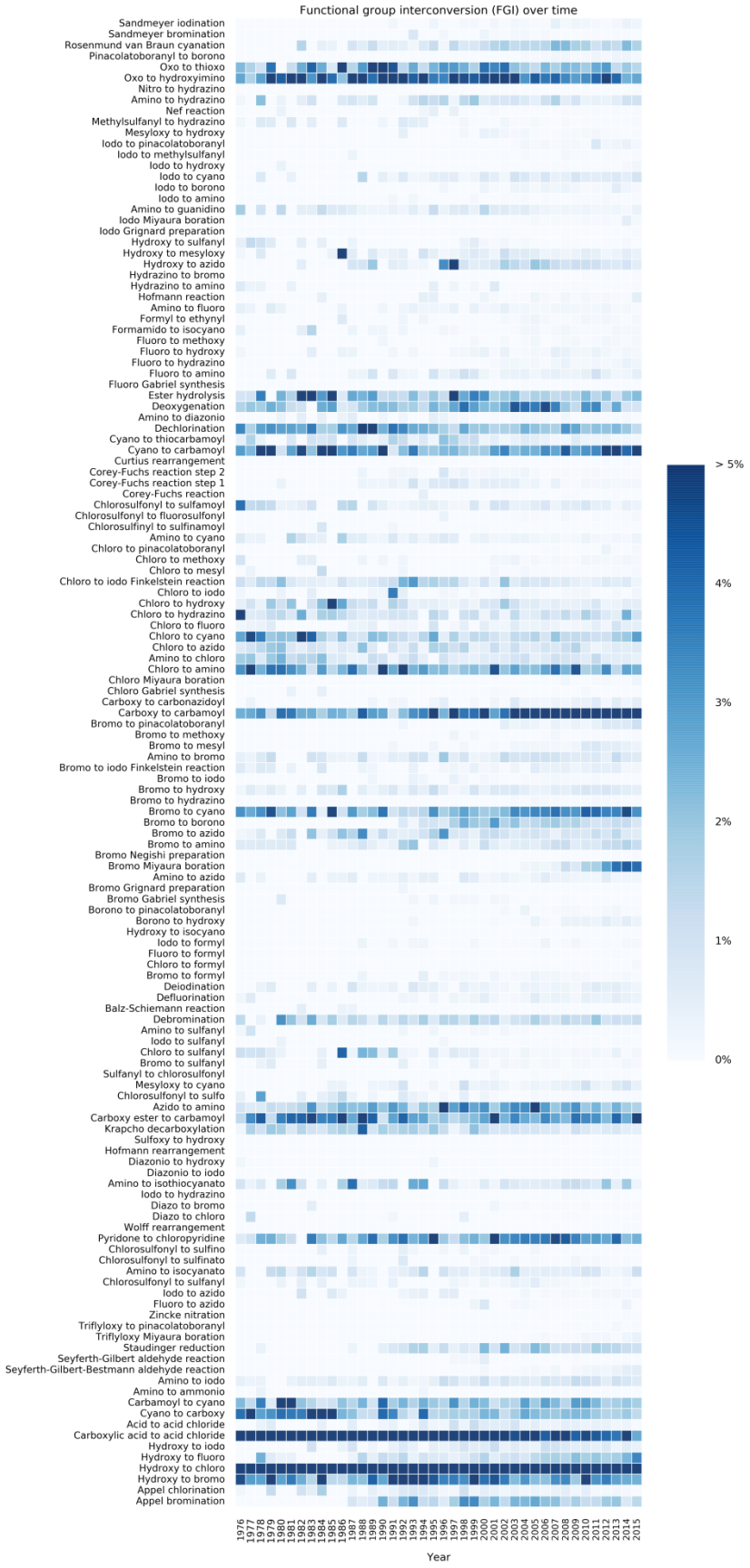


Figure S16: Evolution of *Functional group additions* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.

Figure S17: Evolution of *Functional group interconversions* over time. In the plot all reaction types found at least once over the whole time period were shown. Values were normalized by the yearly number of reactions in this reaction class.



Evolution of yield over time

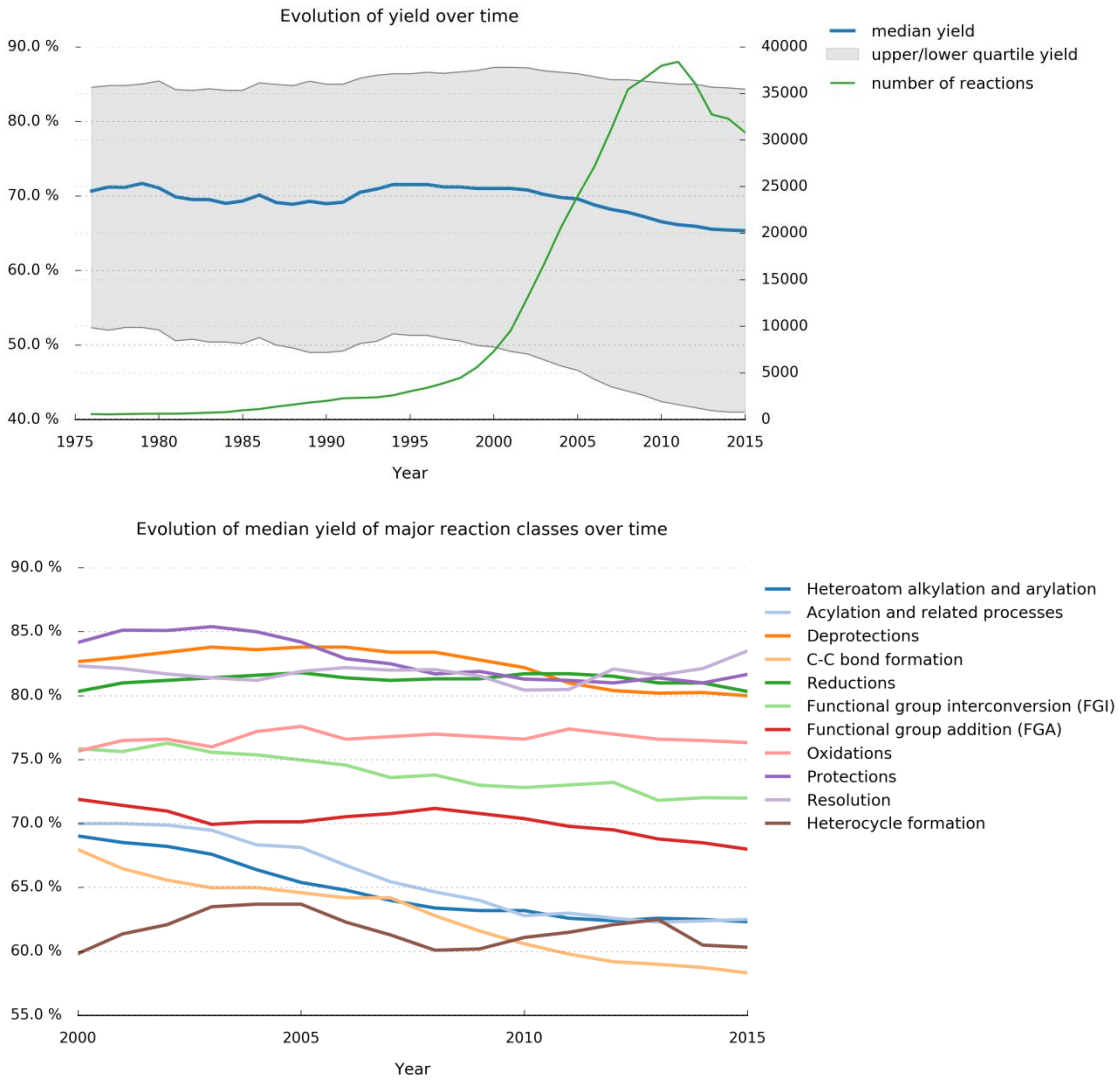


Figure S18: Top: Evolution of the median yield of all reactions over time (blue line). In gray the upper and lower quartile are drawn. In green the number of reactions from which the median yield value was calculated is plotted. Bottom: Evolution of the median yield of the major reaction classes over time. In this plot a time period between the year 2000 and the year 2015 (October) was considered. Please note that the values in both plots were smoothed/averaged over 5 years for better visual clarity.

Evolution of reaction products and their properties over time

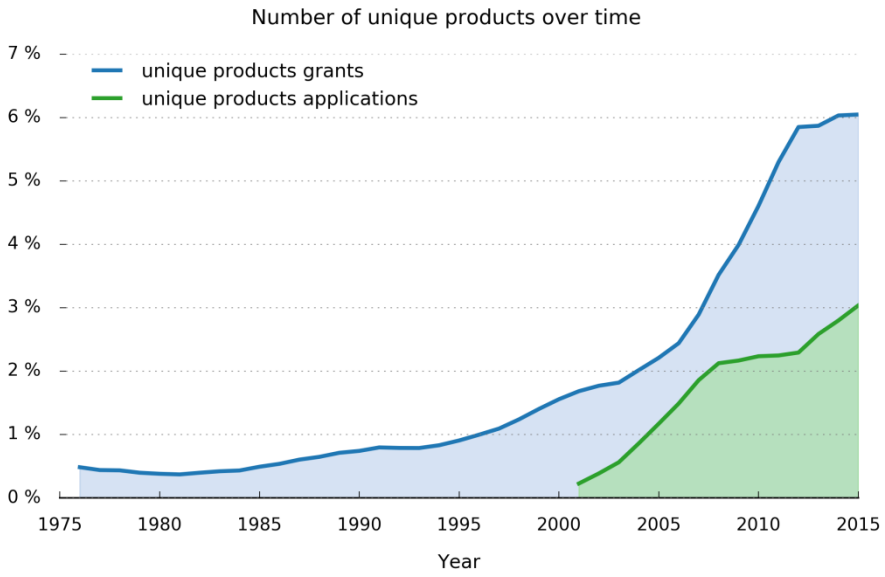


Figure S19: Evolution of the number of unique reaction products over time. In blue products extracted from granted US-patents are shown. In green products from US-patent applications were plotted. In this plot a time period between the year 1976 and the year 2015 (October) was considered. Please note that the values in the plot were smoothed/averaged over 5 years for better visual clarity.

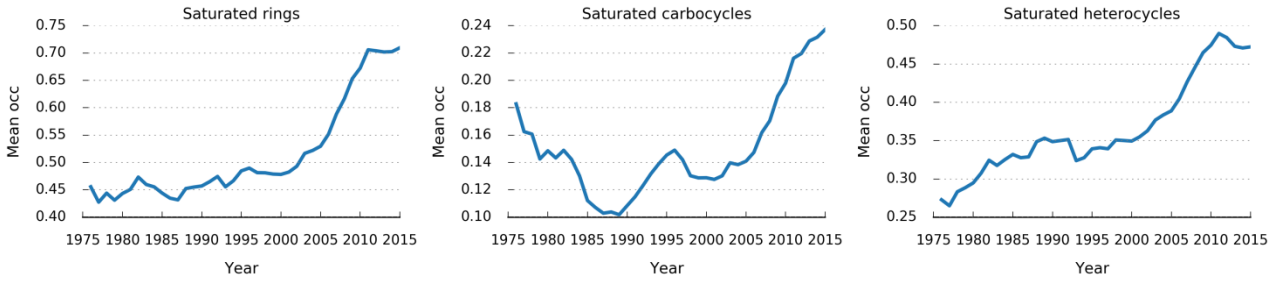


Figure S20: Evolution of the mean number of saturated rings in reaction products over time. In these plots a time period between the year 1976 and the year 2015 (October) was considered. Please note that the values in the plots were smoothed/averaged over 5 years for better visual clarity.

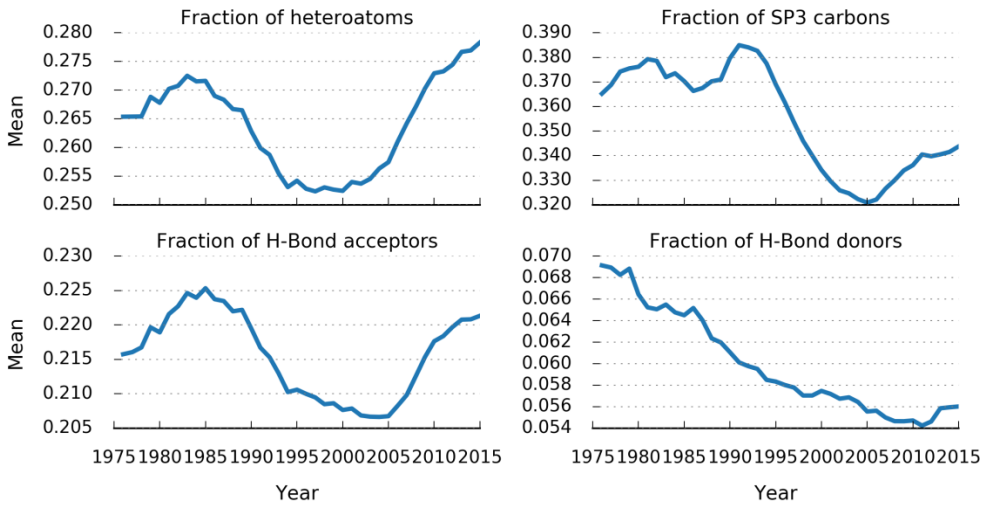


Figure S21: Evolution of the mean number of different structural properties of reaction products over time. These atom-based structural properties were normalized by the number of heavy atoms. In these plots a time period between the year 1976 and the year 2015 (October) was considered. Please note that the values in the plots were smoothed/averaged over 5 years for better visual clarity.