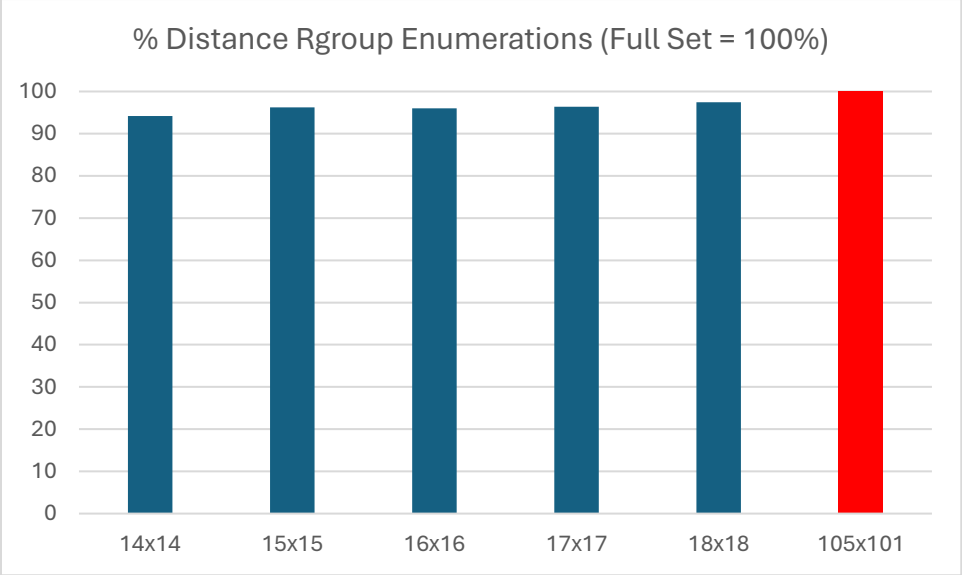
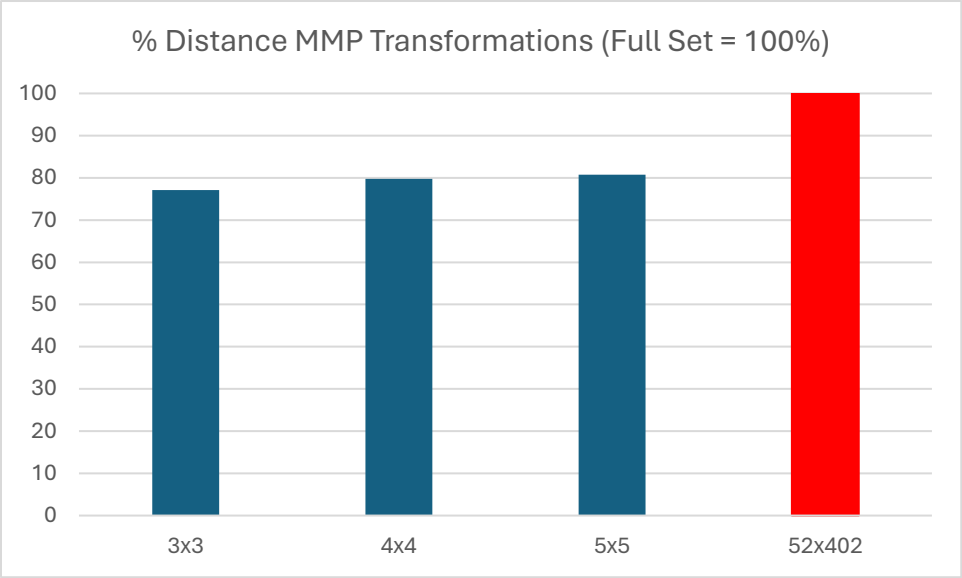
Electronic Supplementary Material (ESI) for Reaction Chemistry & Engineering.

Suppl. 3: Distance Calculations

We analyzed the average pairwise distances of the 25 selected target compounds as a function of the number of Rgroups defined as diversity requirements (see Section 3.3 and Table 1) and compared them to the average pairwise distances for the full test sets with 423 and 452 target compounds, respectively.
Pairwise distances were determined with the well established ECFP_4 fingerprints and calculated as 1 – Tanimoto similarity coefficient. As a consequence, the higher the avg. pairwise distance, the more diverse the set. In the graphs below, avg. pairwise distances of the selected target sets (blue bars) are expressed as percentage of the respective full sets (red bar).

As can be deduced, the selected 25-compound subsets, though representing only ~5% of the full sets, still feature between 77% and 97% of the diversity of the respective full sets, showing that our method can truly select highly diverse subsets.




% Distance MMP Transformations (Full Set = 100%)
100
90
80
70
60
50
40
30
20
10
0
3x3
4x4
5x5
52x402

Reactions Rgroup Enumeration 18x18	Frequency
Imidazole synthesis	23
Fluoro N-arylation	10
Bromo N-arylation	9
SNAr ether synthesis	9
Fluorination	5
Unrecognized	5
Carboxylic ester + amine reaction	3
Friedel-Crafts acylation	2
Bromo Sonogashira coupling	1
Carboxylic acid to acid chloride	1
Ullmann-type biaryl coupling	1
Bromo N-arylation	1
Wurtz-Fittig coupling	1
Carboxylic acid to acid chloride	1
Aldehyde reductive amination	1

Reactions MMP Transformation 5x5	Frequency
Carboxylic acid + amine condensation	24
Methoxy to hydroxy	11
Williamson ether synthesis	8
Mitsunobu aryl ether synthesis	3
Unrecognized	2