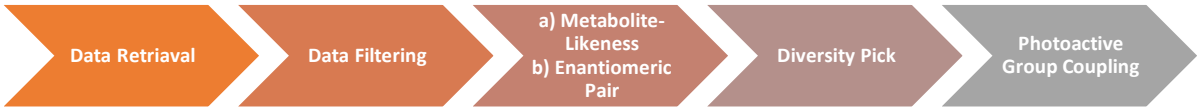
Fabian Kaiser
University of Vienna, 2019

Summary: High Diversity Fragment Datasets for Proteomics Studies

The aim of the project was to create high diversity fragment datasets, with commercially available compounds, for in vitro testing. The fragments will be used to describe the "ligand-able proteome" by checking which proteins are able to interact with the fragments. The bioassay requires a photoactive group which needs to be coupled to the fragment.

Two different subsets were created, one representing "metabolite-like" fragments and the other one with enantiomeric pairs.



The project was performed using KNIME Analytics Platform 4.0.2.

Data retrieval and filtering

As starting point we used the *Enamine fragment collection* (https://enamine.net/fragments/fragment-collection) which is the biggest database of that kind, with 172k synthetically available fragments.

The database was filtered by the molar weight, keeping structures with 200-350 g/mol and filtering out fragments with a logP below -0,4 (Rule of 5 addition by Ghose et al.[1]).

To avoid false positive results in the high throughput screening a *Pan Assay Interference* (PAIN) filter was applied. We used the three subsets (A, B & C) from the RDKit PAIN filter implementation of Baell et al.[2].

The remaining data was filtered by functional groups. Only fragments with a carboxylic acid group or a primary amine or a secondary amine group were kept. This was done due to the synthetic coupling of the photoactive group which was described by Parker et al.[3]. Structures with multiple of those groups were filtered out due to eventual problems with the synthesis.

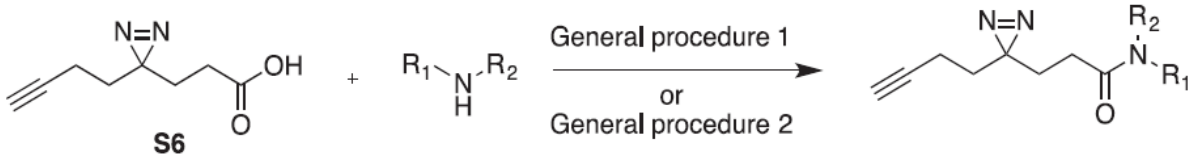


Figure 1 Coupling of the photoactive group (left) with an amine [4]

Next, we standardized the molecules with a company intern standardization protocol, the main goal of this step is to annotate all functional groups in the same manner. This is important for the coupling of the photoactive group. The standardization tool also computes the *International Chemical Identifier* (InChI) key which we used to remove duplicate compounds.

[1] Arup K. Ghose, Vellarkad N. Viswanadhan, and John J. Wendoloski, 'A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases', *Journal of Combinatorial Chemistry* 1, no. 1 (12 January 1999): 55–68, https://doi.org/10.1021/cc9800071.
[2] Jonathan B. Baell and Georgina A. Holloway, 'New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays', *Journal of Medicinal Chemistry* 53, no. 7 (8 April 2010): 2719–40, https://doi.org/10.1021/jm901137j.
[3] Christopher G. Parker et al., 'Ligand and Target Discovery by Fragment-Based Screening in Human Cells', *Cell* 168, no. 3 (26 January 2017): 527-541.e29, https://doi.org/10.1016/j.cell.2016.12.029.
[4] Parker et al.

Ad Metabolite-Like Dataset

The preprocessed Enamine dataset was cross-checked for metabolite-like substructures using the *human metabolome database* (https://hmdb.ca/), as reference. Before that, the dataset was filtered for "endogenous human metabolites", a molar weight oft 150-400 g/mol and linear molecules were removed. Fragments with zero hits were filtered out.

This resulted in a metabolite-hit list with approximately 3,400 datapoints. The similarity of the fragment and the corresponding metabolite were calculated using the Tanimoto coefficient and the MACCS fingerprint.

Diversity Pick

The customer asked for a dataset of 100 fragments with a generally even distribution of the molar weight. First, we sorted the table for the 1000 most similar hits. Next, we split the data 50:50 by their molar weight, therefore we achieved a better distribution after the final pick.

The diversity pick (for both datasets) was performed using the physicochemical RDKit Descriptors and the *MaxMin algorithm*[5] implementation of the *RDKit Diversity Picker* node. The results were validated with a principal component analysis based on the RDKit Descriptors, comparing the final dataset to the source Enamine database. It could be shown that we achieved an even distribution over the Enamine chemical space. Additionally, we checked the chemical diversity with a scaffold analysis which showed a high rate of approx. 85% unique *murcko scaffolds*.

Ad Enantiomeric Pair Dataset: Various combinations of functional group (COOH, RNH_2, RNHR) ratios were tested. Therefore, before doing the diversity pick the dataset was split and the diversity pick was performed three separate times.

Ad Enantiomeric Pair Dataset

The final dataset should include commercially available pairs of enantiomeric fragments, no racemic mixtures. The preprocessed Enamine dataset was filtered for explicit (annotated) chiral centers with the *Speedy SMILES Explicit Chirality Splitter* node from Vernalis. If no chiral center was annotated, the commercially available substance could be a racemate and therefore, these substances have been filtered out.

All substances with exactly one stereo center were filtered and their stereo information was removed with the *Speedy SMILES Strip Stereochemistry* node. Then the dataset was checked for "duplicate" molecules which would represent an enantiomeric pair.

[5] Mark Ashton et al., 'Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions', *Quantitative Structure-Activity Relationships* 21, no. 6 (2002): 598–604, https://doi.org/10.1002/qsar.200290002.

Photoactive Group Coupling

Both final datasets should include the fully functional (FF, fragment + photoactive group) and the parental (PAR, fragment + amide) structure.

The dataset had to be split into the three functional groups: carboxylic acids, primary amines & secondary amines. The coupling reaction was performed with the *RDKit Two Component Reaction* node. The reaction for both the PAR and FF was defined by the following SMART pattern:

[#7;!H0;!$([#7]-C=[O,N,S]);!$([#7]~[!#6]):1].[C:2](=[O:3])-[O;H,-]>>[#7:1]-[C:2]=[O:3]

See figure 2 and 3.

The parental structure was created by the following reaction:

Educ 1	**Educ 2**	**Parental structure**
R-COOH	NH_2CH_3	$R\text{-}CONHCH_3$
$R\text{-}NH_2$	CH_3COOH	$R\text{-}NH\text{-}COCH_3$
R-NH-R	CH_3COOH	$R\text{-}NR\text{-}COCH_3$

The FF reaction can be seen in figure 1.

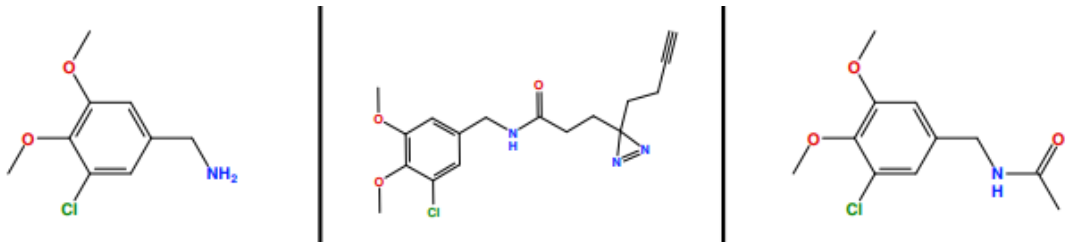


Figure 2 Photoactive Group coupling example: amine. fragment (left), FF (mid), PAR (right)

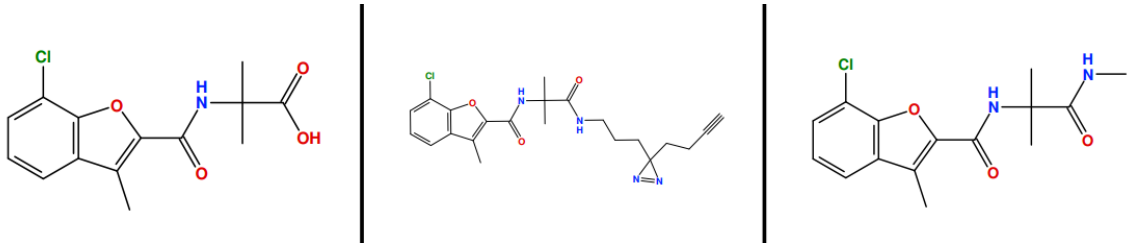


Figure 3 Photoactive Group coupling example: carboxylic acid. fragment (left), FF (mid), PAR (right)