Supporting Information for: Single-step Retrosynthesis Prediction based on the Identification of Potential Disconnection Sites using Molecular Substructure Fingerprints

Haris Hasic and Takashi Ishida*

Department of Computer Science, School of Computing, Tokyo Institute of Technology, W8-85, 2-12-1 Ookayama, Meguro, Tokyo, Japan

E-mail: ishida@c.titech.ac.jp

Data Pre-processing Details

The chemical reaction data used in this research is derived from the USPTO dataset, which is currently the largest chemical reaction dataset available for free. The raw dataset consists of reactions that were text mined from U.S. patent documents between 1976 and 2016 from fields like chemistry, physics, or human necessities. The two parts of the dataset are patent grants (1976 - 2016) and patent applications (2001 - 2016) which contain 1,808,937 and 1,939,253 entries, respectively. A pharmacy-related subset of these reactions has been extracted, cleaned and classified into 10 different reaction classes, resulting in a dataset of around 50,000 reactions, often referred to as USPTO-50k. The overview of the distribution of the reaction classes is depicted in Figure S1. The data has been further processed by split-

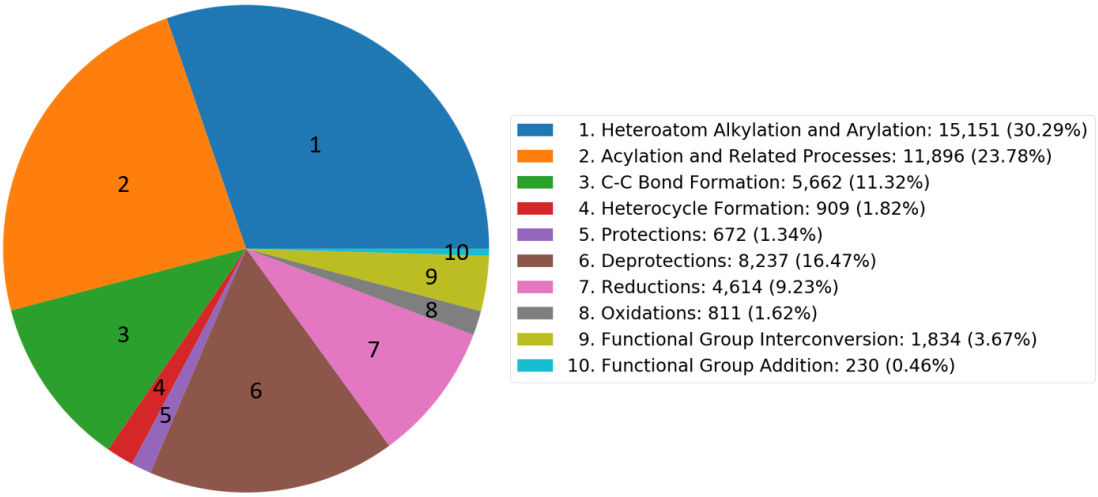

Figure S1: Distribution of the reaction classes in the refined USPTO-50k dataset.

ting reactions with multiple products into multiple reactions and removing smaller molecules from the product side (*i.e.*, removing byproduct salts). This version of the dataset (which can be found at `https://github.com/connorcoley/retrosim/tree/master/retrosim/data/data_processed.csv`) is used as a starting point in this research.

As the first step to create a suitable dataset for the prediction model, all of the reactant and product molecules are parsed from the reaction SMILES strings in the dataset. These molecules are stripped of their mapping and converted to the canonical SMILES form by using RDKit. Duplicate compounds are filtered out and pools of unique reactants (52,838 entries) and unique products (49,594 entries) are created. Based on these compound pools, the dataset is expanded with the unique ID mapping of the reaction in the form of:

$$\{uq\_reac\_id\_0, \ldots, uq\_reac\_id\_n\} \rightarrow \{uq\_prod\_id\_0, \ldots, uq\_prod\_id\_n\}$$

In the next step, a total of 50,016 raw chemical reaction entries is split into training, validation and test sets in 5 different ways according to the 5-fold cross-validation approach. These splits are performed per class so that every reaction type is equally represented in each fold. The value for n was set to 5 because some classes (*e.g.*, 'Reduction' or 'Functional Group Addition') have a low number of entries and a high number of folds would cause an insufficient representation of these reaction classes in each fold. The dataset splits were done following the 70:10:20 ratio for the training, validation and test set, respectively. Next,

for all of the generated splits, the substructures from the product molecules are extracted and labelled as potential disconnection or non-disconnection sites, as described in the main manuscript. These substructures are then converted to fingerprints according to a specified configuration which dictates the type, radius, number of bits and neighbourhood extension. Entries processed this way result in significantly more negative than positive substructure samples. The number of samples that is kept is equal to the number of samples of the most represented reactive substructure class. It is noticeable from Table S1 and Table S2 that the total number of extracted substructures is slightly higher than the total number of reactions. This is due to the occurrence of reaction cores which consist out of two or more separated atom groups which are considered as separate reaction cores. After the process is finished, the positive and negative samples are combined and labelled with classes [0, 10] for multi-class classification purposes. The label 0 represents a non-disconnection site and labels from 1 to 10 represent potential disconnection sites expanded with additional information about the established classes. This concludes the dataset pre-processing procedure.

Table S1: The overall number of reaction entries, number of extracted substructures and the final dataset split sizes per each fold.

Fold	**Dataset**	**Number of Reactions**	**Number of Substructures**	**Final Size (w/o Augm.)**	**Final Size (w/ Augm.)**
1	Training	35,006	35,018	45,651	116,963
	Validation	5,002	5,002	6,490	6,490
	Testing	10,008	10,011	13,042	13,042
2	Training	35,009	35,020	45,652	116,952
	Validation	5,002	5,003	6,492	6,492
	Testing	10,005	10,008	13,039	13,039
3	Training	35,012	35,020	45,600	116,380
	Validation	5,002	5,008	6,550	6,550
	Testing	10,002	10,003	13,033	13,033
4	Training	35,012	35,023	45,610	116,457
	Validation	5,002	5,003	6,538	6,538
	Testing	10,002	10,005	13,035	13,035
5	Training	35,015	35,025	45,624	116,589
	Validation	5,002	5,002	6,525	6,525
	Testing	9,999	10,004	13,034	13,034

Table S2: The average number of reactions and the average difference of extracted substructures compared to the number of reactions for each dataset split per reaction class.

Reaction Class	**Training Dataset**		**Validation Dataset**		**Test Dataset**	
	Avg. Rxns.	**Avg. Substrs.**	**Avg. Rxns.**	**Avg. Substrs.**	**Avg. Rxns.**	**Avg. Substrs.**
No Disconnection	10,606		1,515		3,030	
Het. Alk. and Ary.	10,606	+0.2	1,515	+0.4	3,030	+0.4
Acyl. and Rel. Pr.	8,320	+0.8	1,196	+0.0	2,379	+0.2
C-C Bond Form.	3,957	**+4.2**	573	+0.2	1,132	**+1.4**
Heterocy. Form.	633	+0.4	95	+0.4	182	+0.2
Protections	474	+0.2	64	+0.2	134	+0.4
Deprotections	5,769	+0.6	821	+0.2	1,647	+0.6
Reductions	3,233	**+1.2**	458	**+0.6**	923	**+1.2**
Oxidations	571	+0.4	77	+0.4	162	+0.2
FGI	1,289	**+3.2**	179	+0.0	367	+0.3
FGA	159	+0.2	25	+0.2	46	+0.0

The total number of negative ECFP samples before the filtration process is 109,537, 31,113, and 49,978 for training, validation and testing, respectively. The total number of negative HSFP samples is 302,220, 38,042, and 106,739 for training, validation and testing, respectively. The filtering is necessary because some of the generated negative samples are misleading and might cause the model to be heavily biased towards negative labels.

Model Construction Details

The parameter optimization process for the neural network is split into two stages. In the first stage, the molecular descriptor parameters are analyzed and the two best performing substructure fingerprint configurations are selected. In the second stage, the hyperparameters of the model are optimized to achieve the best possible performance and avoid overfitting.

Pre-selection of Input Data Configurations

To judge how different molecular descriptor configurations affect the performance of the model, a simple prediction model was constructed. This architecture is shown in Figure S2. It consists of only one fully connected layer. The input and hidden layer had the number of neurons set to the bit size of the current fingerprint, used the ReLU activation function and had dropout set to 0.2. The output layer size was 11 for each of the established classes and Softmax activation function was used. The monitored loss was categorical cross-entropy and it was minimized using the ADAM optimizer, with the application of early stopping on validation loss after 10 epochs with no improvement. The effect on model performance was evaluated through the variation of the following configuration parameters:

- Fingerprint Type: $\{SubstructureECFP, HSFP\}$
- Fingerprint Radius: $\{1, 2, 3\}$
- Fingerprint Bits: $\{1024, 2048, 4096\}$
- Neighbourhood Extension for HSFP: $\{1, 2\}$
- Data Augmentation: $\{None, SMOTE\}$

The performance of models like this is usually judged based on the average test loss or test accuracy, but in this case, the discriminative power of the model is more important. How much of the relevant data (disconnection sites) is the model able to recognize and correctly label can be assessed from the Precision-Recall curve by calculating the average precision score for each class according to the formula:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{1}$$

where P_n and R_n are the precision and recall scores at the n-th threshold. Given that this is a heavily imbalanced multi-class classification problem, the average precision score of the

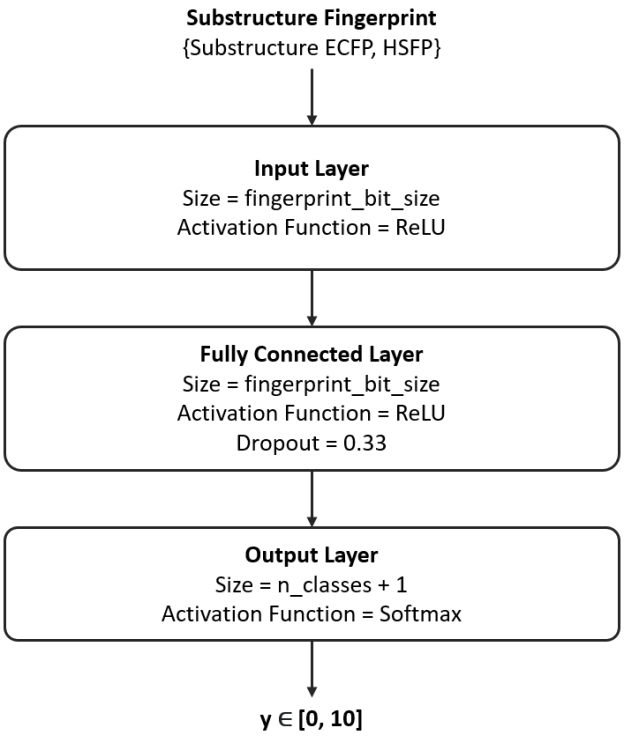

Figure S2: Architecture of the initial model used for evaluating the effects of molecular descriptor configuration parameters on overall model performance.

whole model is determined by calculating the micro-average value for all classes first and then calculating the average precision score. The class imbalance problem is handled by applying the SMOTE approach, which works well with numerical vectors and helps avoid losing valuable reaction data to construct an unbiased prediction model. Taking all these factors into account, the pre-selection of models based on input data configurations was done in two phases. The focus of the first phase is the fingerprint bit size and the focus of the second phase is the amount of information contained in the fingerprint. During the evaluation, the focus parameter is modified while the other parameters remain fixed. In this way, the effect of the parameter in focus on the overall performance of the model can be properly monitored. The best models for each of the configurations that were chosen for further investigation through network hyperparameter tuning were:

1. Substructure ECFP: $\{radius = 2, bit\_size = 1024\}$

2. HSFP: $\{radius = 2, bit\_size = 1024, neighbourhood\_extension = 2\}$

Hyperparameter Tuning

The hyperparameter tuning was done through through a guided grid search approach because of the small amount of data available. Namely, model architectures in terms of number of layers and layer types have been predetermined based on standard model construction practices in this field. The following six fixed network architectures were considered:

1. **Fixed Model 1**: $\{Input\} \rightarrow \{FullyConnectedLayer\} \rightarrow \{Output\}$
2. **Fixed Model 2**: $\{Input\} \rightarrow \{FCL\}x2 \rightarrow \{Output\}$
3. **Fixed Model 3**: $\{Input\} \rightarrow \{FCL\}x3 \rightarrow \{Output\}$
4. **Fixed Model 4**: $\{Input\} \rightarrow \{FCL\} \rightarrow \{HighwayLayer\} \rightarrow \{Output\}$
5. **Fixed Model 5**: $\{Input\} \rightarrow \{FCL\} \rightarrow \{HL\}x3 \rightarrow \{Output\}$
6. **Fixed Model 6**: $\{Input\} \rightarrow \{FCL\} \rightarrow \{HL\}x5 \rightarrow \{Output\}$

To test the whether the model is throttled by capacity issues the number of layers is gradually increased. After the basic architectures have been fixed, the variable parameters values were defined. Each configuration that was constructed during this grid search was considered as an independent model and the evaluation was done through cross-validation of the inputs and the parameters. The following hyperparameter values were used:

- Hidden Layer Size: $\{2/3 * input\_size, \mathbf{input\_size}, 3/2 * input\_size\}$
- Dropout Values: $\{0.2, \mathbf{0.33}, 0.5\}$
- Activation Function: $\{eLU, \mathbf{ReLU}\}$
- Learning Rate: $\{\mathbf{0.0005}, 0.001, 0.00146\}$
- Batch Size: $\{16, 32, 64, \mathbf{128}, 256\}$

The best performing model was the first fixed model with only one fully connected layer which is expected given the small amount of data that is available.

Reactant Retrieval and Scoring Details

The reaction core is marked and the synthons are extracted from every reaction in the USPTO-50k dataset. The synthons are sorted based on their total number of atoms in descending order and the candidates with the highest similarity values are retrieved. The resulting Top-N accuracy is shown in Table S3.

Table S3: Top-50 accuracy chart of correctly retrieved reactant compounds based on the size the synthon substructures in terms of the number of atoms.

Position	**Top-1**	**Top-3**	**Top-5**	**Top-10**	**Top-20**	**Top-30**	**Top-50**	**Entries**
1	79.9%	89.2%	91.2%	92.9%	93.8%	94.0%	94.2%	50,016
2	35.7%	55.6%	64.6%	73.9%	78.9%	80.6%	81.8%	32,977

Around 79.9% of the correct compounds for the largest synthon in the first position are fetched as the Top-1 candidate. In the case of the smaller synthons, only 35.7% of the correct compounds are fetched as the Top-1 candidate. This number rises to 73.9% if the Top-10 candidates are considered. To get a better insight on how the size of the synthon substructures affects the accuracy of the reactant retrieval process, the synthons are further sorted into five categories based on the total number of atoms: [0, 5], (5, 10], (10, 25], (25, 50], and >50. The upper bound is set to 50 because the highest number of atoms in a single molecule in the USPTO-50k dataset is 97 and such molecules are not frequent. The Top-N accuracies for these categories are shown in Table S4.

Table S4: Top-50 accuracy chart of correctly retrieved reactant compounds based on the size the synthon substructures in terms of the number of atoms.

Atom Count	**Top-1**	**Top-3**	**Top-5**	**Top-10**	**Top-20**	**Top-30**	**Top-50**	**Entries**
[0, 5]	13.1%	30.4%	43.7%	55.9%	60.7%	62.8%	64.8%	2,914
(5, 10]	32.2%	55.4%	64.6%	75.6%	81.7%	83.7%	85.3%	21,502
(10, 25]	78.2%	88.7%	90.8%	92.2%	93.0%	93.2%	93.3%	41,179
(25, 50]	95.9%	99.6%	99.7%	99.8%	99.8%	99.8%	99.8%	11,857
>50	96.1%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	180