Bayesian Algorithm for Retrosynthesis

Zhongliang Guo　Data Science Center for Creative Design and Manufacturing

Motivations and Objectives

Difficulties in retrosynthetic prediction

Most machine learning models for retrosynthetic analysis build a direct transformation from a given product to its reactants. There are two drawbacks for the backward prediction approach.

- **The unavailable molecules in the outcomes**
- **The low accuracy of backward prediction models**

Task	Model	top-1	top-3	top-5	top-10
Backward	Similarity (Coley et al. 2017)	37.3	54.7	63.3	74.1
	SCROP (Zheng et al. 2019)	43.7	60.0	65.2	68.7
	Lin et al. 2019	43.1	64.6	71.8	78.7
Forward	Template-based (Coley et al. 2017)	71.8	86.7	90.8	94.6
	WLDN (Jin et al. 2017)	79.6	87.7	89.2	-
	Molecular Transformer (Schwaller et al. 2019)	90.4	94.6	95.3	-

Bayesian retrosynthesis

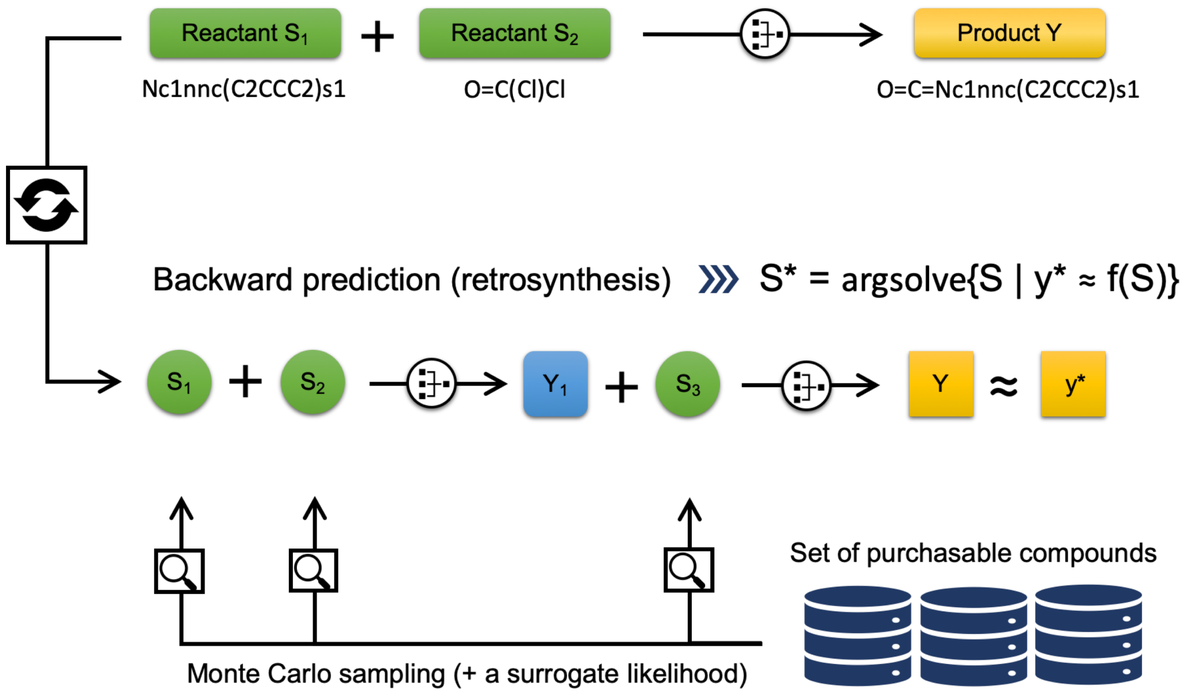
The retrosynthetic analysis can be reduced to a combinatorial optimization task whose solution space is subject to the combinatorial complexity of all possible paires of purchasable reactants in the catalog. We address this issue within the framework of Bayesian inference and computation.

Bayes' law of conditional probability

$$p(S|Y=y^*) \propto p(Y=y^*,S) = p(Y=y^*|S)p(S)$$

The workflow consists of two steps

1. Predict the product of given reactants using an accurate model (forward prediction)
2. Invert the forward model to the backward model



Methods

Simple SMC

Joint Probability

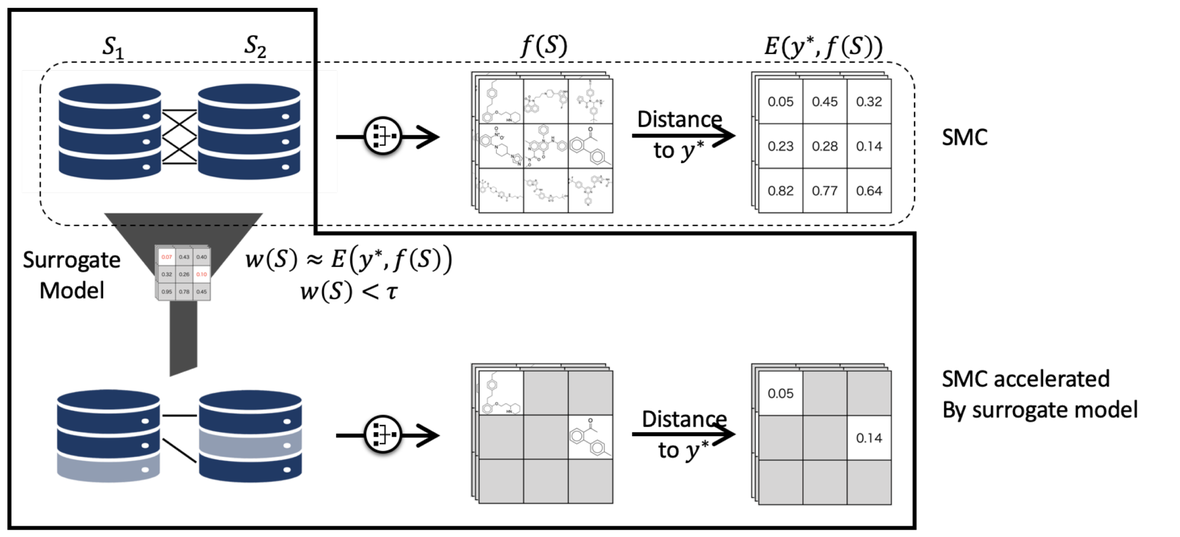
$$p(Y=y^*,S) \propto \exp\Big(-\beta E\big(y^*,f(S)\big)\Big)$$

The difficulties in the simple SMC

1. The diversity of the solutions and the problem of particle impoverishment in SMC.
2. The cost of forward prediction.

We propose an surrogate-accelerate Bayesian retrosynthesis.

Surrogate-accelerate SMC



Experiments and Results

Data and forward model

- Dataset: 50K single-step reactions (Schneider et al. 2016). 80% for training, 10% for validation, 10% for test.
- Solution space: all possible combinations of 600K reactants in USPTO dataset.
- Forward prediction model: Molecular Transformer (fine-tuned on the training and validation data). Top-1 accuracy 86.9%; top-5 accuracy 95.5%.

One-step retrosynthesis

- 100 randomly selected reactions from the test set
- 87 Molecular Transformer-predictable reactions
- 600,000 (p=1000, t=600) searches for each test case, aroud 0.0001% of the complete search space (3.6×10^{11})

Table 1: Performance of the surrogate-accelerated SMC

	# of reactions	Detection of reactants ending with target product [%]	Inclusion of ground-truth reactants [%]
Random100	100	98.4	88.3
MT-predictable	87	99.1	94.5

Table 2: Performance of various retrosynthesis prediction methods with or without the reaction-class labels

Model	top-1	top-3	top-5	top-10
Similarity	37.3	54.7	63.3	74.1
SCROP	43.7	60.0	65.2	68.7
Lin et al. 2019	43.1	64.6	71.8	78.7
Bayesian-Retro	**47.5**	**67.2**	**77.0**	**80.3**
Bayesian-Retro (MT-predictable)	54.6	74.9	86.0	89.4

Multi-step retrosynthesis

- 11 two-step reactions generated by connecting two one-step reactions
- 10 valid reactions evaluated by expert chemists
- 2,000,000 (p=2000, t=1000) searches for each test case, around 10^{-11} of the complete search space (2.2×10^{17})

In 9 of the 11 reactions, the recorded synthetic routes were identified.

Figure 1: The recorded reaction 9.

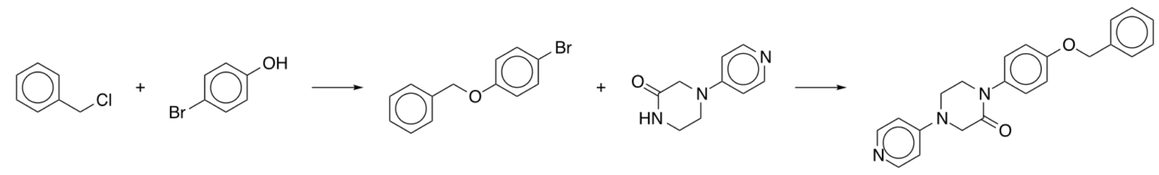


Figure 2: Distribution of detected synthetic routes to the target product in reaction 9.

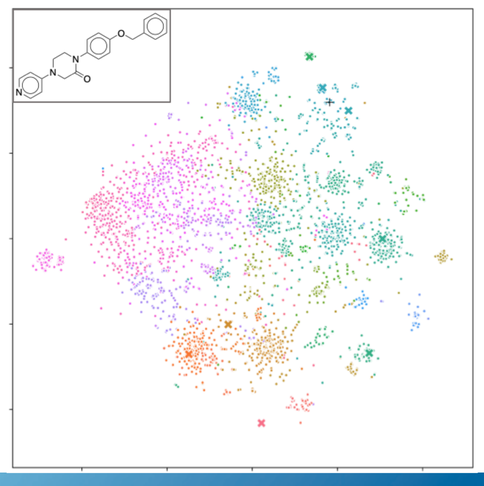


Figure 3: Five detected synthetic routes to the target product in reaction 9.