

文献質問回答(6/2,笹本)

Q1.メチルシフトを伴う置換と脱離の例。この時成績の悪い理由について

メチルシフトを伴う置換の例についてはデータセットを見てもどのような反応を指しているのか不明でした。Reaction Type3 と Reaction Type4 はどちらと判断されるかランダム(50%)になるような教師シグナルを用いているので成績が悪くなっています。

Q2.機械学習の合成ルートの方が現実性があるため、文献のルートよりも学生が好んだというのは当然ではないか。

1970~1990 年代に見られたような、明らかに不合理なステップを含まないという点について確認していると考えています。現実性という点でいうと、そのとおりだと思います。

Q3.どのようなパラメータに報酬を与えると合成ルートが短くなるのか。

A3.ノードの評価値は、Rollout でビルディングブロックまで展開するのにかかったステップ数を考慮しています。そのため、短いステップ数が実現できそうなノードは高く評価されることとなります。

Q4.誰がこのような分野に取り組んでいるのか。他の分野に比べ儲からないように思える。

A4.メルクが Chematica を開発した企業の買収を行った例もありますし、深層学習を用いて新薬設計を行うソフトウェアを開発している企業が 50 億の調達に成功した例もあるので、決して儲からないということはないと思います。

Q5.合成経路探索では教師付き学習が主のようだが、deep learningなどで評価関数なども自動生成できるようになるまでには何がハードルとなるのか。

A5.合成経路の評価関数については、強化学習の活用により解決できるのではないかと考えています(今回紹介した MCTS 法も評価関数を用いていません)。反応予測という問題になると、テンプレートフリーの手法は何個か出てきていますが、完全に教師データを必要としない、とすると物理計算を用いる必要がありかなり計算コストを要求される点がハードルとなるのかなと思います。

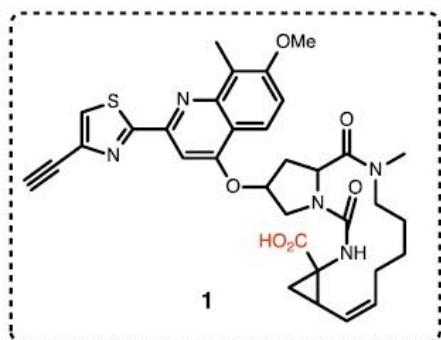
Q6.合成経路に対する学生の好みはなんのために測定しているのか。そもそも必要か。

A6.正しいとか、実際に使えるかどうかという点については別に検証する必要があります。明らかに不合理なステップ(複数の反応点を含む時に、より反応しにくい方で反応させようとするなど)を含まないという点を確認しているのではないかと思います。

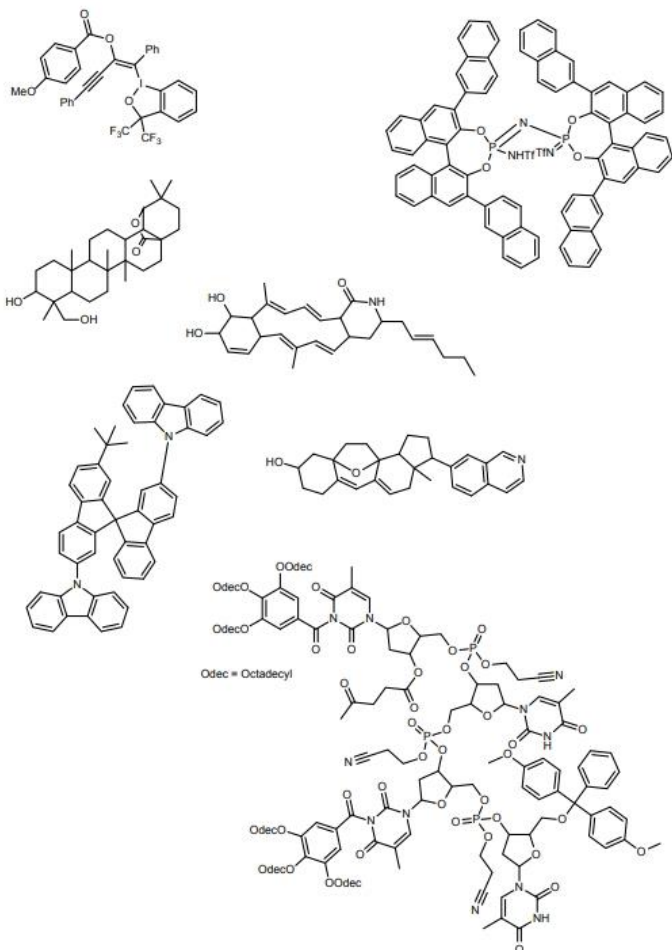
Q7.実際にどのくらいの複雑な分子がデザインされているのか。

A7.

実際にデザインできた例



できなかった例



学習に用いているデータの不足がデザインの可否を分けているのではないかと考えています。データの不足は、ヒューリスティックや物理化学計算などと組み合わせることである程度解決できるかもしれません。

Q8.Chematica は新規化合物に使えないというのはどういうことか。

A8.Chematica は既に合成ルートが報告されている化合物同士を反応ネットワークでつないでいるので、ネットワークが存在しない新規化合物については合成ルートを提示することができません。そのため、Syntaurus を用いてその化合物に反応が適用できるか・できないかを確認しつつ合成ルートをデザインしています。

Q9.人が分類しているのはどこか。ケミストが入っていけるのはどの点か。

A9.基本的にデータは機械を用いて自動的に分類させています。現在のプログラムでは扱うことのできない複雑な分子(=データが少ない)に対して、今まで化学者が蓄積してきたヒューリスティックと組み合わせることで対応することは可能であると考えています。将来的に優れたプログラムが出てきた時には、機械の導き出したパラメータを解釈する際に化学者が必要となってくるのではないかと思います。

Q10.精度はどれくらいなのか。

A10.MCTS を用いた場合、逆合成の各ステップが文献と一致するのは 30~50%程度です。文献と必ずしも一致する必要はありませんが、別のアプローチを用いるなど改善の余地はあると考えています。

Q11.どれくらいのデータ数が必要か。

A11.化学反応の数はとても多いので、反応予測に関しては 10~100 万程度、逆合成という課題に取り組む場合にはより多くのデータが必要となると思われます。

Q12.MCTS だと p.31 のように A,B の収率がいいがその前はダメといった問題はでないのか。

A12.Rollout で SM まで展開して、ノードの評価を行っているので問題ありません。もっとも、収率は評価のパラメータに用いられていないので、実際にやってみたらダメだったという可能性は十分ありえます。

Q13.LHASA の時点で難しい化合物の経路設計はできているのか。

LHASA の場合、原理的には複雑な化合物の設計も行えるはずですが、実際には用いることのできるルールセットは限られているため、適用できる分子は限られています。タキソールの一部の合成経路デザインをした例が報告されていますが、その程度が限界でしょうか。
(Red. Trav. Chim. Pays-Bas, 1994, 113, 297-303)