NextMove Software

Filling in the Gaps: Extracting Implied Reactions with RDKit

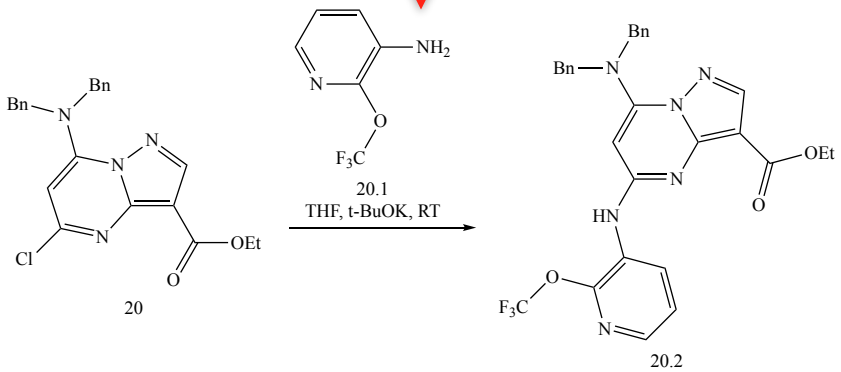
Rachael Pirie, John Mayfield, Roger Sayle

CURRENT REACTION EXTRACTION

"Synthesis of Compound 20.2.

To a cooled solution of 20 (0.380 g, 0.902 mmol, 1 eq), and 20.1 (0.144 g, 0.812 mmol, 0.9 eq) in tetrahydrofuran (5 mL) at 0° C. was added potassium ter-butoxide (1.80 mL, 1.80 mmol, 2.0 eq). The reaction was stirred at room temperature for 30 min. After completion of reaction, reaction mixture was transferred into saturated bicarbonate solution and product was extracted with ethyl acetate. Organic layer was combined and dried over sodium sulphate and concentrated under reduced pressure to obtain crude material. This was further purified by column chromatography and compound was eluted in 17% ethyl acetate in hexane to obtain pure 20.2. (0.270 g, 53.16%). MS (ES): m/z 563.55 $[M+H]^+$."

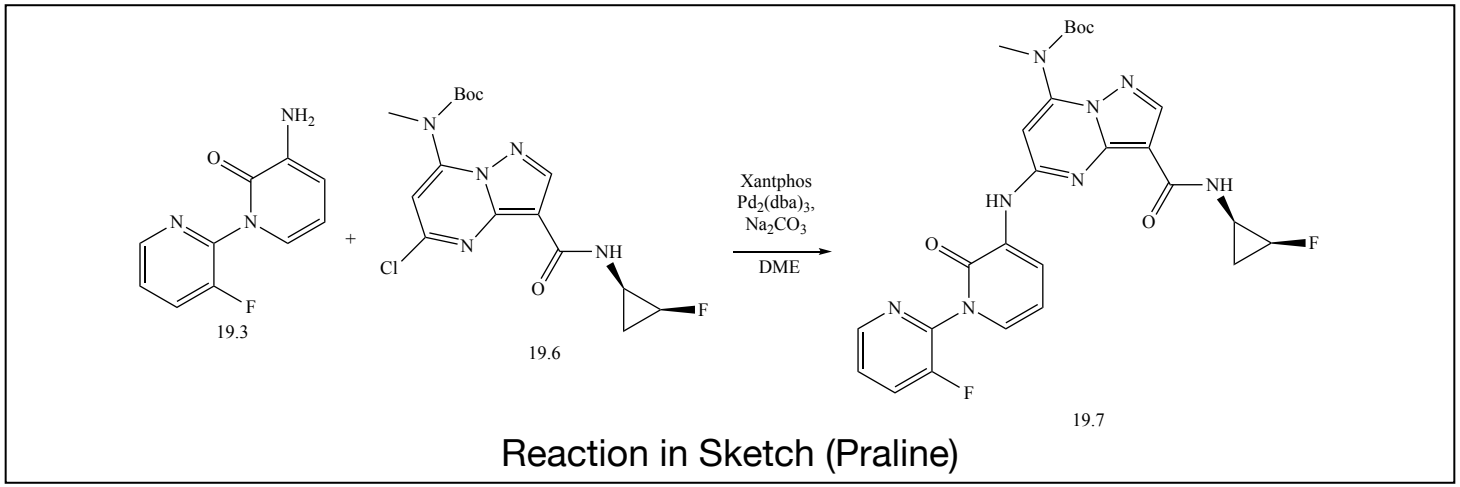
Reaction in text (LeadMine)



Convert to Sketch

e.g. US20190241576A1

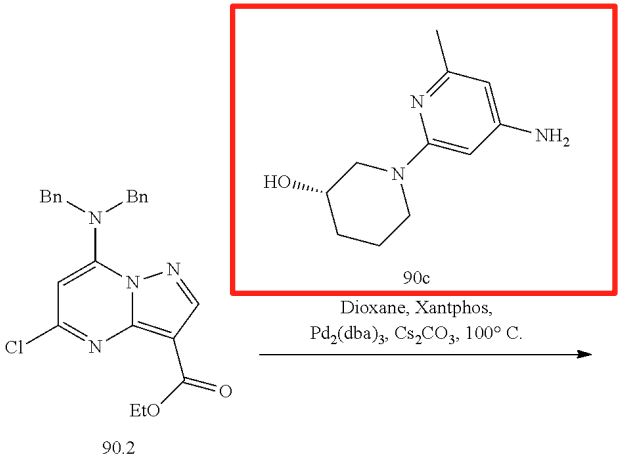
CURRENT REACTION EXTRACTION

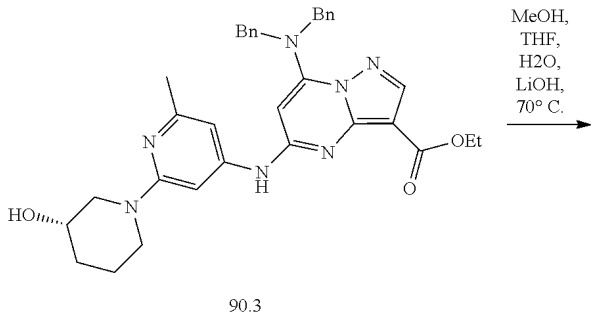


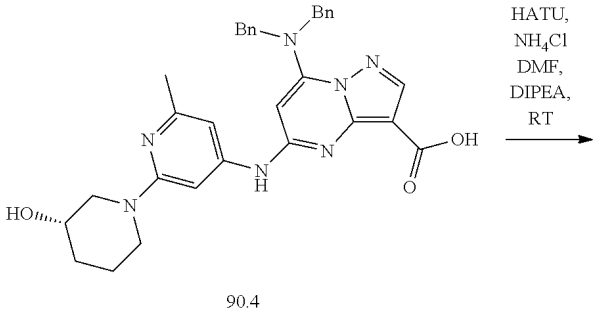
Reaction in Sketch (Praline)

e.g. US20190241576A1

MOTIVATION: IMPLIED REACTIONS



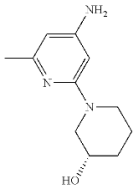
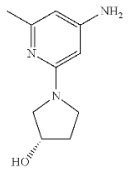
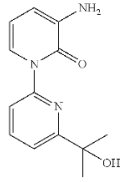
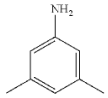




e.g US20190241576A1: *"Compounds in Table 15 were prepared by methods substantially similar to those described to prepare I-62, where 90c was replaced with the reagent as indicated in Table 15."*

~ 5000 patents with ~15000 tables between 2019-2023

TABLE 15

Compound #	Reagent	Characterization Data
I-62		MS (ES): m/z 383.53 [M + H]+, LCMS purity: 98.76%, HPLC purity: 98.31%, CHIRAL HPLC purity: 100%, ^{1}H NMR (DMSO-d_6, 400 MHZ): 9.47 (s, 1H), 8.19 (s, 1H), 7.70 (s, 2H), 7.43 (s, 1H), 7.30 (s, 1H), 6.86 (s, 1H), 6.67 (s, 1H), 5.76 (s, 1H), 4.89 (s, 1H), 4.16-4.13 (d, J = 9.6 Hz, 1H), 3.95- 3.92 (d, J = 12.8 Hz, 1H), 3.47 (s, 1H), 2.82- 2.72 (m, 1H), 2.24 (s, 3H), 1.89 (bs, 1H), 1.76 (bs, 1H), 1.44-1.28 (m, 2H), 1.10-1.08 (m, 1H).
I-56		MS (ES): m/z 369.52 [M + H]+, LCMS purity: 98.14%, HPLC purity: 97.13%, ^{1}H NMR (DMSO-d_6, 400 MHZ): 9.46 (s, 1H), 8.20 (s, 1H), 7.71 (s, 2H), 7.43 (s, 2H), 6.52 (s, 1H), 5.78 (s, 1H), 4.90 (s, 1H), 4.38 (s, 1H), 3.44-3.40 (m, 4H), 3.30-3.24 (m, 1H), 2.25 (s, 3H), 2.01-1.91 (m, 2H).
I-41		MS (ES): m/z 421.23 [M + H]+, LCMS purity: 95.01%, HPLC purity: 95.00%, ^{1}H NMR (DMSO-d_6, 400 MHZ): 9.00 (s, 1H), 8.19 (s, 1H), 8.02-8.00 (t, J = 8 Hz, 1H), 7.76- 7.74 (d, J = 8 Hz, 1H), 7.68-7.66 (d, J = 8 Hz, 2H), 7.59-7.57 (d, J = 8 Hz, 1H), 7.39 (s, 2H), 7.22 (s, 2H), 6.43-6.42 (t, J = 4 Hz, 1H), 6.11 (s, 1H), 5.39 (s, 1H), 1.48 (s, 6H).
I-21		MS (ES): m/z 297.48 [M + H]+, LCMS purity: 95.04%, HPLC purity: 95.37%, ^{1}H NMR (DMSO-d_6, 400 MHZ): 9.27 (s, 1H), 8.13-8.12 (d, J = 4.4 Hz, 1H), 7.548 (s, 3H), 7.305 (s, 1H), 7.184 (s, 2H), 6.66 (s, 1H), 5.67 (s, 1 H), 2.25 (s, 6H).

CANONICAL ATOM MAP IDXS

- Option 1: invariant atom property (RDKit default)
  - **Different atom maps == different order**

```
[CH3:1][CH2:2][OH:3].[Cl:4]>>[CH3:1][CH2:2][Cl:4]
[Cl:1].[OH:1][CH2:2][CH3:3]>>[Cl:1][CH2:2][CH3:3]
```

- Option 2: ignore
  - Useful for **tracking atoms through reordering**

```
[CH3:1][CH2:2][OH:3].[Cl:4]>>[CH3:1][CH2:2][Cl:4]
[CH3:3][CH2:2][OH:1].[Cl:1]>>[CH3:3][CH2:2][Cl:1]
```

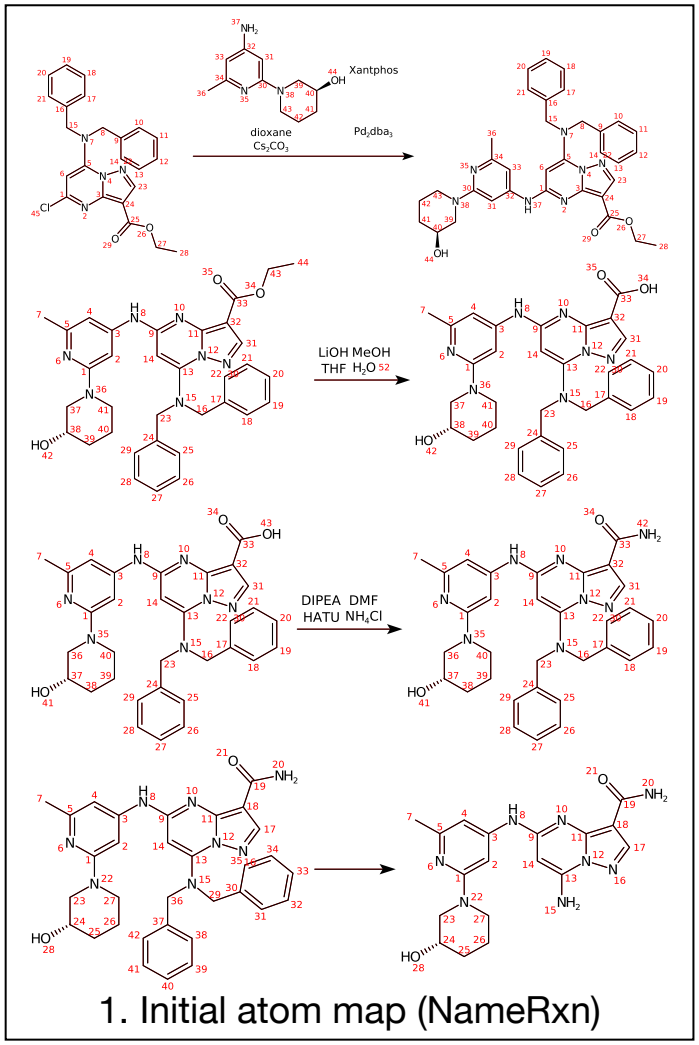
- Option 3: ignore and renumber (1-N)
  - Useful for: **“are these two reactions the same?”**

```
[CH3:1][CH2:2][OH:3].[Cl:4]>>[CH3:1][CH2:2][Cl:4]
[CH3:1][CH2:2][OH:3].[Cl:4]>>[CH3:1][CH2:2][Cl:4]
```

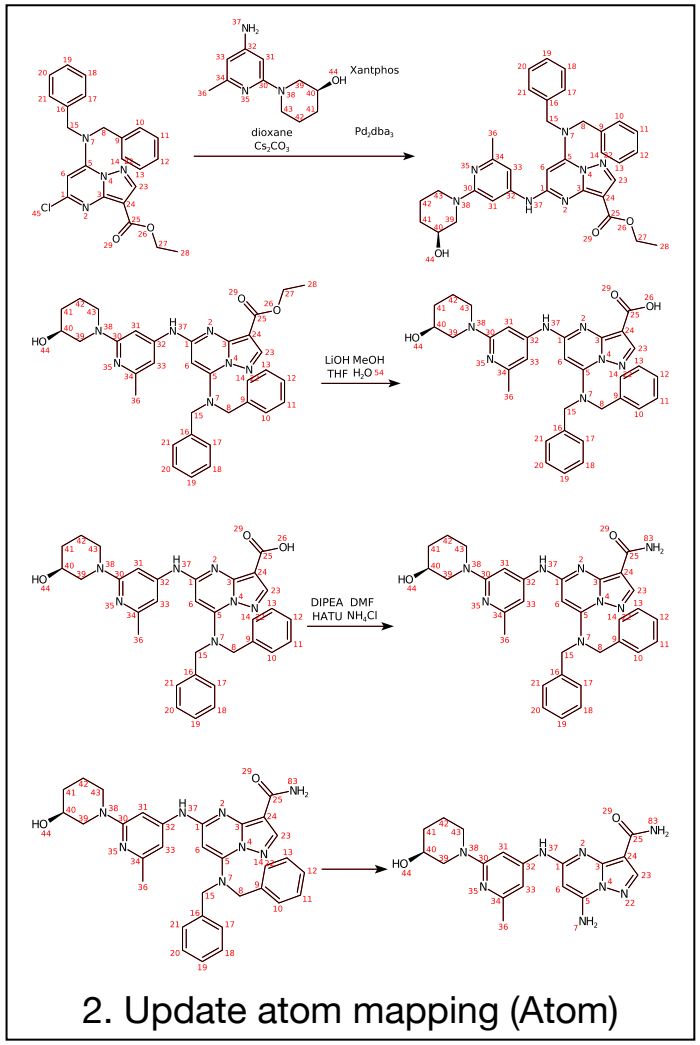
RDKIT WORK AROUND

```python
from rdkit import Chem

def mapidx_order(smi):

    mol = Chem.MolFromSmiles(smi)

    # backup and clear atom mapping
    mapidxs = [0] * mol.GetNumAtoms()
    for atm in mol.GetAtoms():
        mapidxs[atm.GetIdx()] = atm.GetAtomMapNum()
        atm.SetAtomMapNum(0)

    cansmi = Chem.MolToSmiles(mol)
    mapout = [mapidxs[x] for x in eval(mol.GetProp('_smilesAtomOutputOrder'))]

    # put them back on
    for atm in mol.GetAtoms():
        atm.SetAtomMapNum(mapidxs[atm.GetIdx()])

    return cansmi, mapout


print(mapidx_order("[CH3:1][CH2:2][OH:3]"))
print(mapidx_order("[CH3:3][CH2:2][OH:1]"))
print(mapidx_order("[OH:3][CH2:2][CH3:1]"))
```

EXTRACTING IMPLIED REACTIONS

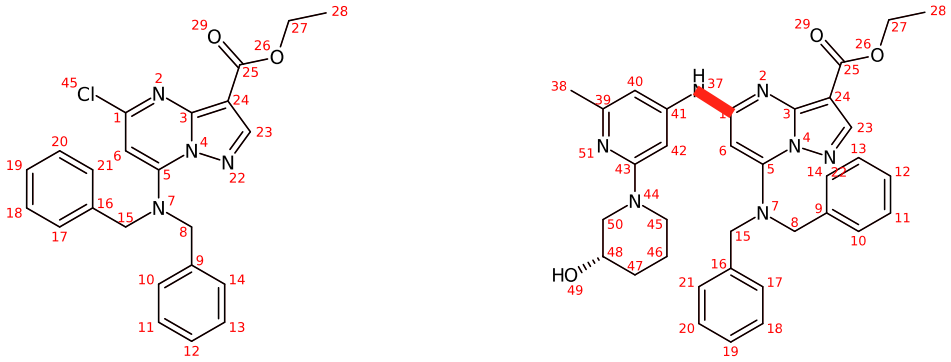


1. Initial atom map (NameRxn)

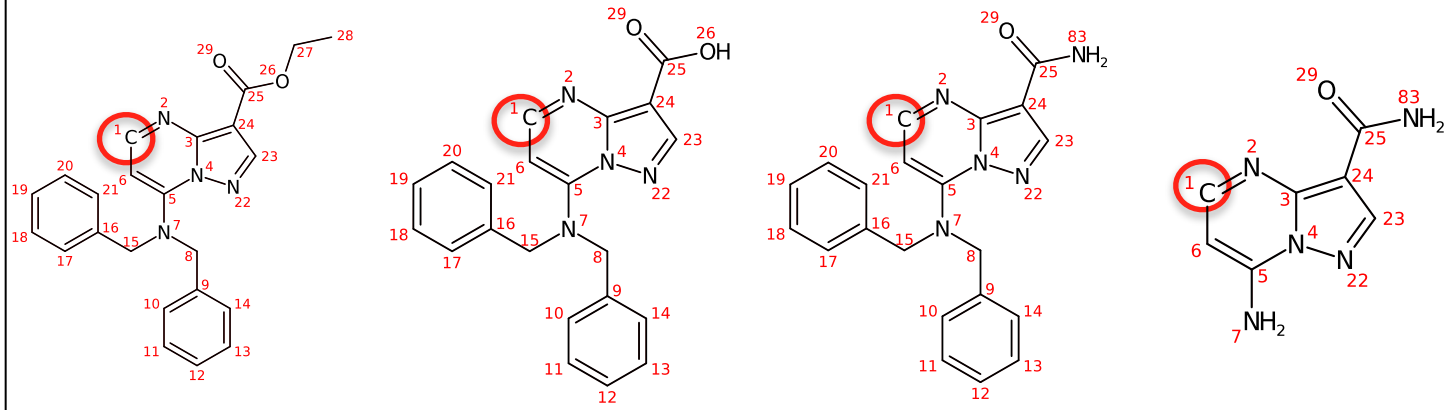


2. Update atom mapping (Atom)

EXTRACTING IMPLIED REACTIONS

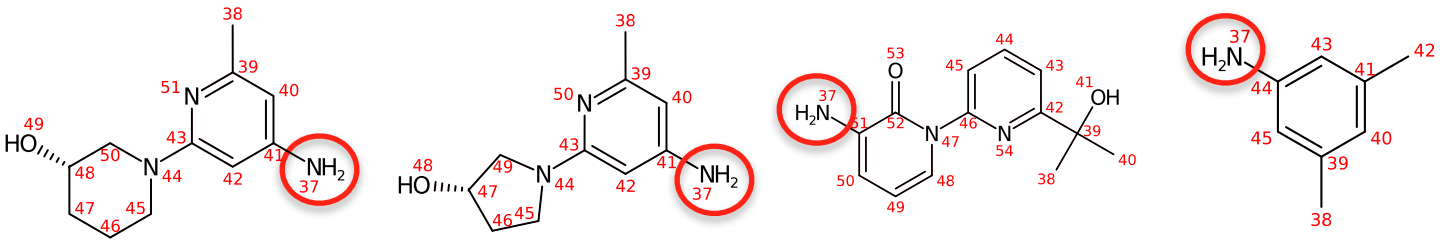


3. Find attachment position (Bond)

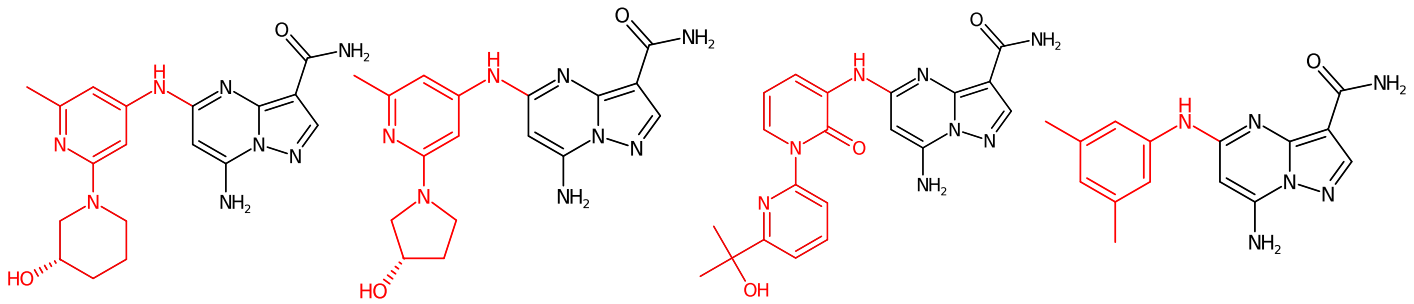


4. Get core molecules (Editable Mol)

EXTRACTING IMPLIED REACTIONS

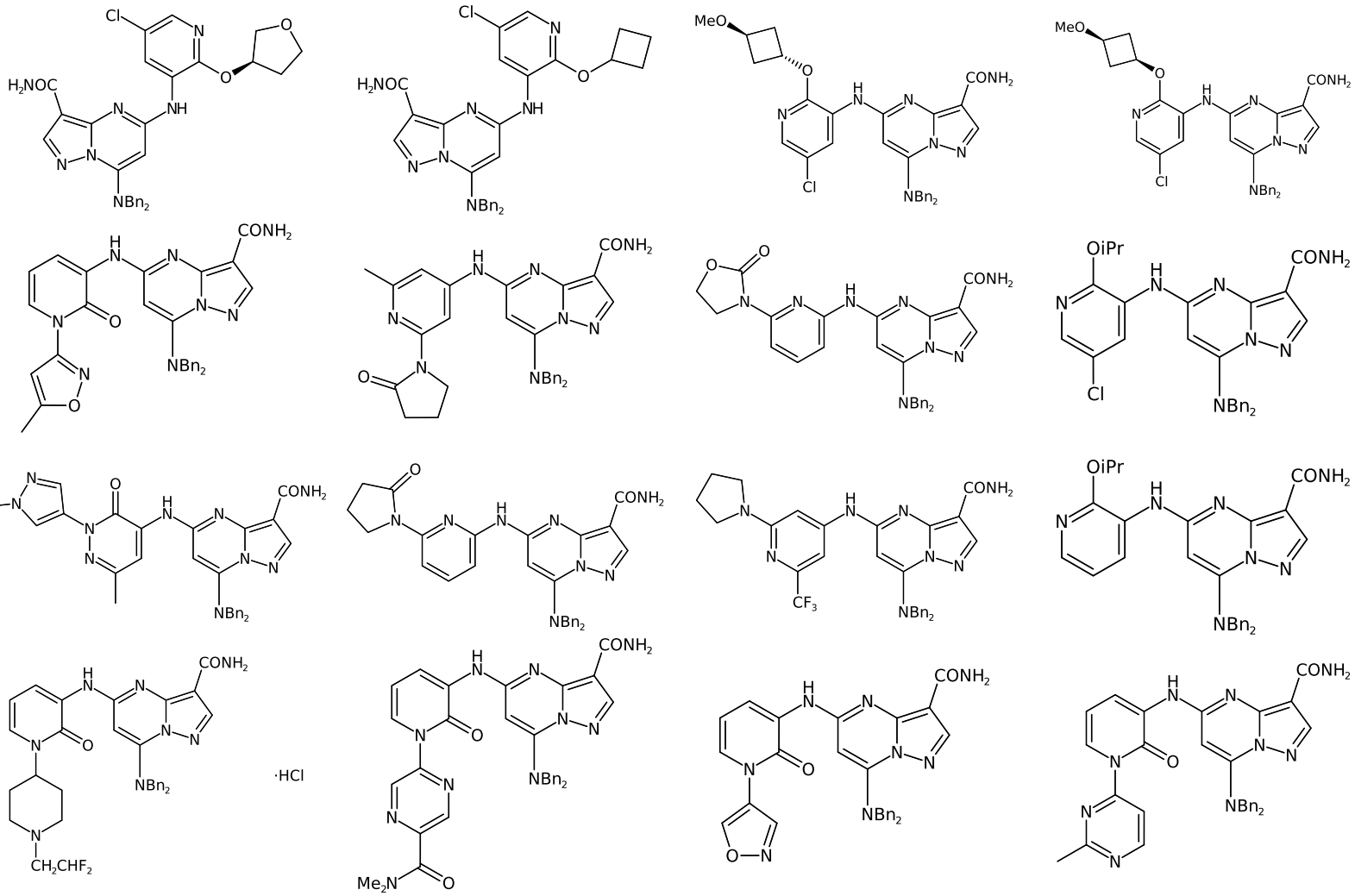


5. Find rxn centre in replacements with SMARTS and generate atom map (Atom)



6. Join replacements + cores (Combine Mols)

NOVEL TO PUBCHEM



e.g US20190241576A1

RECOMMENDATIONS

- **ReactionFromSmiles** instead of **ReactionFromSmarts(useSmiles=True)**
  - There is a ReactionToSmiles!!!
- Additional options for handling atom-maps in canonical SMILES
  - Ignore perhaps should be default (John and Roger's opinion)
    - SmilesWriteParams.atomMapInvariant // Option 1
    - SmilesWriteParams.atomMapRenumber // Option 3

ACKNOWLEDGEMENTS

Team @ NextMove
John Mayfield
Roger Sayle
Delia Sayle
Ingvar Lagerstedt

Contact:
rachael@nextmovesoftware.com
@rachaelpirie203