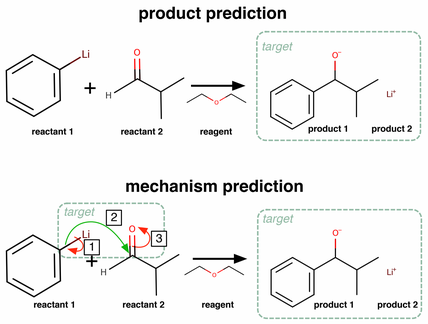
A Generative Model For Electron Paths

John Bradshaw[1,2], Matt J. Kusner[3,4], Brooks Paige[1,4], Marwin H. S. Segler[5], José Miguel Hernández-Lobato[1,4,6]

[1]University of Cambridge, [2]MPI for Intelligent Systems, [3]University of Oxford, [4]The Alan Turing Institute, [5]BenevolentAI, [6]Microsoft Research Cambridge.

Aim: predict reaction outcomes through predicting the movement of electrons (the reaction mechanism)

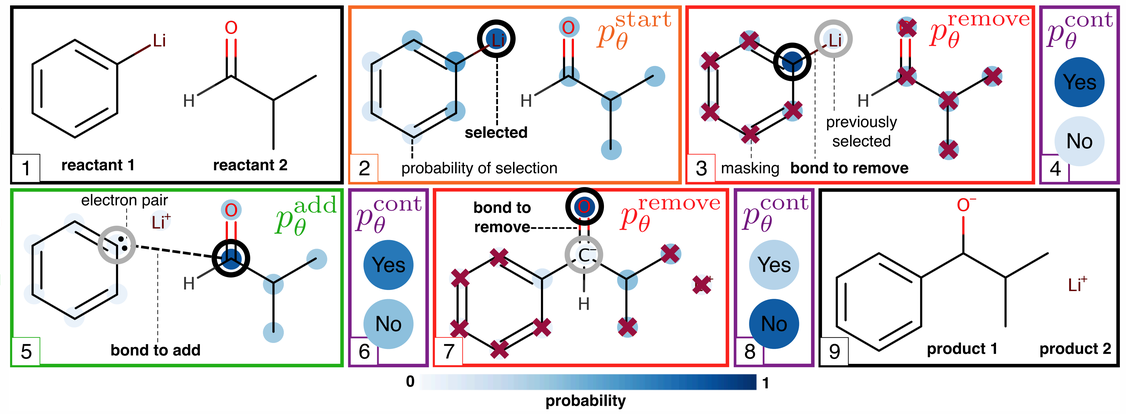


- Reactions involve the step wise movement of electrons along the atoms: this **removes** and **adds** bonds.
- Recent approaches of ML applied to reaction prediction [1, 2] have focused on product prediction.
- We also want to know how the reactants formed the products.
- Therefore, we instead approach the problem from the angle of mechanism prediction: this provides **interpretability**, takes advantage of the **sparsity** of reactions, and naturally **incorporates the constraints of chemistry** (eg balanced atom counts).
- We model reactions exhibiting linear electron flow (LEF) topology, the most common occurring of reactions [3].

Our model, ELECTRO, uses graph neural networks to parameterize the probability of each kind of action

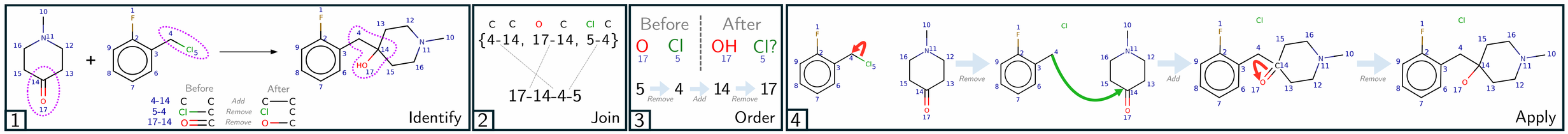
Reactions exhibiting LEF topology have a single electron path consisting of alternating **remove** and **add** bond steps.

1: Select **starting** atom A using $p_\theta^{\text{start}}(a \mid \mathcal{M}_0, \mathcal{M}_e)$
2: **repeat**
3: Select atom B bonded to A using $p_\theta^{\text{remove}}(a_t|\mathcal{M}_t, a_{t-1})$ and **remove** 2 electrons from bond B-A.
4: Decide whether to **continue** using $p_\theta^{\text{cont}}(c_{t+1} \mid \mathcal{M}_{t+1})$.
5: Select new atom A using $p_\theta^{\text{add}}(a_t|\mathcal{M}_t, a_{t-1})$ and **add** bond B-A.
6: Decide whether to **continue** using $p_\theta^{\text{cont}}(c_{t+1} \mid \mathcal{M}_{t+1})$.

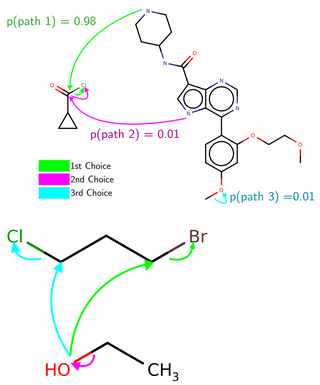


We can extract these approximate electron paths from large scale reaction databases (such as USPTO [5]) for training ELECTRO

Reaction datasets often store reactions as atom-mapped SMILES strings, e.g. `[CH3:10][N:11]1[CH2:12][CH2:13][C:14](=[O:17])[CH2:15][CH2:16]1.[F:1][c:2]1[c:3]([CH2:4][Cl:5])[cH:6][cH:7][cH:8][cH:9]1 >> [F:1][c:2]1[c:3]([CH2:4][C:14]2([OH:17])[CH2:13][CH2:12][N:11]([CH3:10])[CH2:16][CH2:15]2)[cH:6][cH:7][cH:8][cH:9]1`
From this we want to extract the reactants, the reagents, the products, and an approximate electron path for training our model.

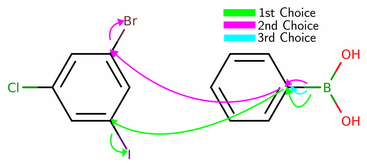


Assessing ELECTRO for mechanism prediction



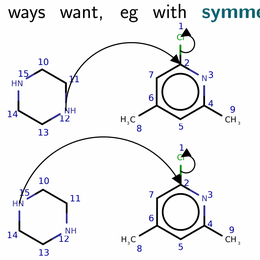
Model Name	Accuracies (%) Top-1	Top-2	Top-3	Top-5
ELECTRO-LITE	70.3	82.8	87.7	92.2
ELECTRO	77.8	89.2	92.4	94.7

Tab. 1: Results when using ELECTRO for *mechanism prediction*. Here a prediction is correct if the atom mapped action sequences predicted by our model match exactly those extracted from the USPTO dataset. ELECTRO-LITE is an ablation study for which we hide the reagents from our model.



Assessing ELECTRO for product prediction

Assessing only for predicting mechanisms means that predicting the correct product in an alternative way is labeled as incorrect. This may not be what we always want, eg with **symmetry**:



∴ We also evaluate ELECTRO for product prediction.

Model Name	Accuracies (%) Top-1	Top-2	Top-3	Top-5
WLDN FTS [1]	84.0	89.2	91.1	92.3
WLDN [1]	83.1	89.3	91.5	92.7
Seq2Seq FTS [2]	81.7	86.8	88.4	89.8
Seq2Seq [2]	82.6	87.3	88.8	90.1
ELECTRO-LITE	78.2	87.7	91.5	94.4
ELECTRO	**87.0**	**92.6**	**94.5**	**95.9**

Tab. 2: Results for *product prediction*. For the baselines we compare against models trained (a) on the full USPTO training set (marked FTS) and only tested on our subset of LEF reactions, and (b) those that are also trained on the same subset as our model. ELECTRO-LITE is an ablation study for which we hide the reagents from our model.

References

[1] Jin, Wengong, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. NeurIPS, 2017.

[2] Schwaller, Philippe, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem. Sci., 2018.

[3] Herges, Rainer. Coarctate Transition States: The Discovery of a Reaction Principle. Journal of chemical information and computer sciences, 1994.

[4] Kayala, Matthew A., and Pierre F. Baldi. A Machine Learning Approach to Predict Chemical Reactions. NeurIPS, 2011.

[5] Lowe, Daniel Mark. Extraction of chemical structures and reactions from the literature. PhD diss., University of Cambridge, 2012.

arXiv:1805.10970

jab255@cam.ac.uk, mkusner@turing.ac.uk, bpaige@turing.ac.uk, marwin.segler@benevolent.ai, jmh233@cam.ac.uk