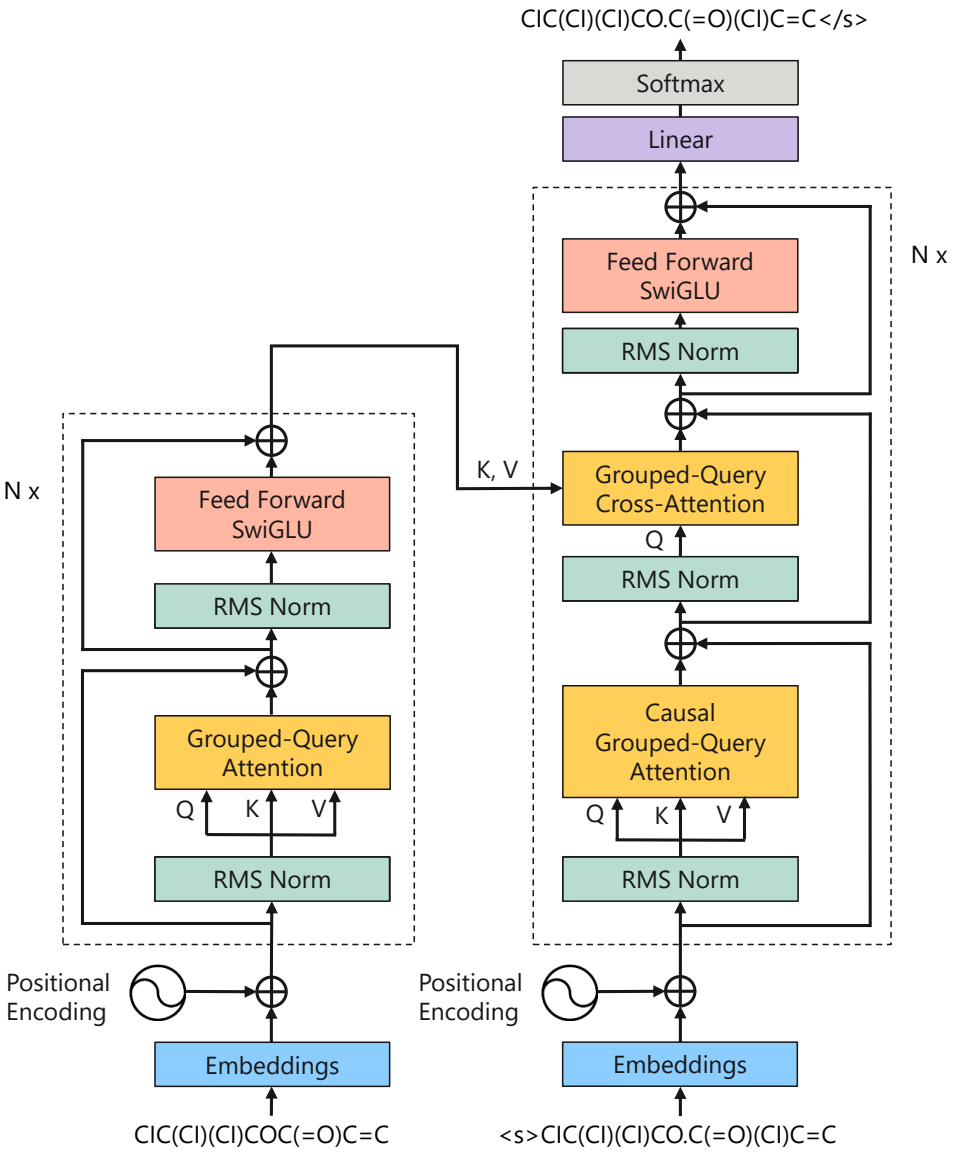
Supplementary Information for: Chemist-aligned retrosynthesis by ensembling diverse inductive bias models

Krzysztof Maziarz*[1], Guoqing Liu (刘国庆)*[1], Felix Pultar*[1], John Gardner[1], Jean Helie[1], Tobias Gensch[1], Hubert Misztela[2], Austin Tripp[3], Junren Li[1], Aleksei Kornev[2], Piotr Gaiński[1,4], Holger Hoefling[2], Mike Fortunato[2], Rishi Gupta[2], Andrew Baxter[5], Darren L. Poole[5], Jennifer M. Elward[5], Adrian Krzyzanowski[5], Peter Pogány[5], Stephen D. Pickett[5], Ian D. Wall[5], Christopher M. Bishop[1], Philip G. Humphreys[5], James A. Lumley[5], Mario P. Wiesenfeldt[6], Marwin Segler[1]

[1]Microsoft Research AI for Science; [2]Novartis Biomedical Research; [3]University of Cambridge; [4]Jagiellonian University; [5]GSK; [6]Bergische Universität Wuppertal;
*Equal core contributor
Correspondence to {krmaziar,guoqingliu,marwinsegler}@microsoft.com

Contents

1 De-Novo Transformer Architecture



Supplementary Figure 1: Architecture of the de-novo model (R-SMILES 2). The input product is converted to a SMILES string and tokenized into a sequence of tokens. Before the sequence is processed further, sinusoidal positional embeddings are incorporated to infuse positional information. The sequence then undergoes transformation through layers composed of grouped-query attention, RMS normalization, and feedforward layers with SwiGLU activations. The autoregressive decoder predicts the SMILES sequence of reactants utilizing self-attention over already produced tokens and cross-attention over encoder output. The model is trained using a cross-entropy loss.

2 Multistep Search Results

Model	Targets solved
RetroChimera	90.7%
RetroChimera ($T = 0.5$)	91.9%
R-SMILES 2	93.5%
R-SMILES 2 ($T = 0.09$)	**94.4%**
NeuralLoc	86.5%
NeuralSym	82.0%

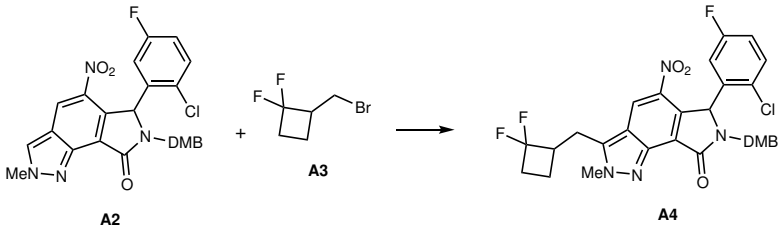
Supplementary Table 1: Solve rates on the $\text{Pistachio}_{\text{Hard}}$ target set (1784 molecules). For RetroChimera and R-SMILES 2, results at the default temperature as well as a tuned temperature setting are reported. Highest solve rate is shown in bold.

3 Extended analysis of routes towards $\text{Pistachio}_{\text{Hard}}$ targets

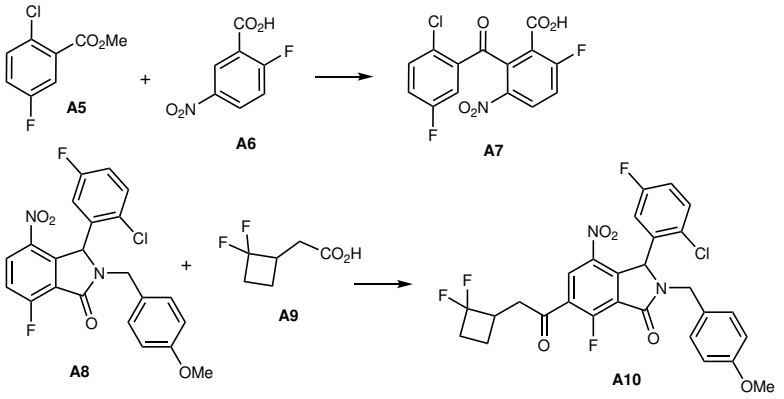
3.1 Route towards A1

Both RetroChimera and R-SMILES 2 produced reasonable routes toward **A1**. In contrast, NeuralLoc predicted routes containing Friedel-Crafts acylations of electron-deficient arenes **A5** and **A8**, which were rejected unanimously by all raters (Supplementary Figure 2). Noteably, in case of **A8**, the presence of a second, electron-rich arene in the same molecule also raises the concern for regioselectivity. The same carbon-carbon bond was introduced in the route designed by RetroChimera from primary bromide **A3** and heterocycle **A2**, which is precedented for related systems.[1,2] For the same target, NeuralSym predicted a cross-coupling reaction of chloride **A11** and Grignard **A12** towards **A13**, as well as multiple halogenation of **A14** and dehalogenation of **A16** without accounting for chemo- and regioselectivity, which renders the route unfeasible (Supplementary Figure 2).

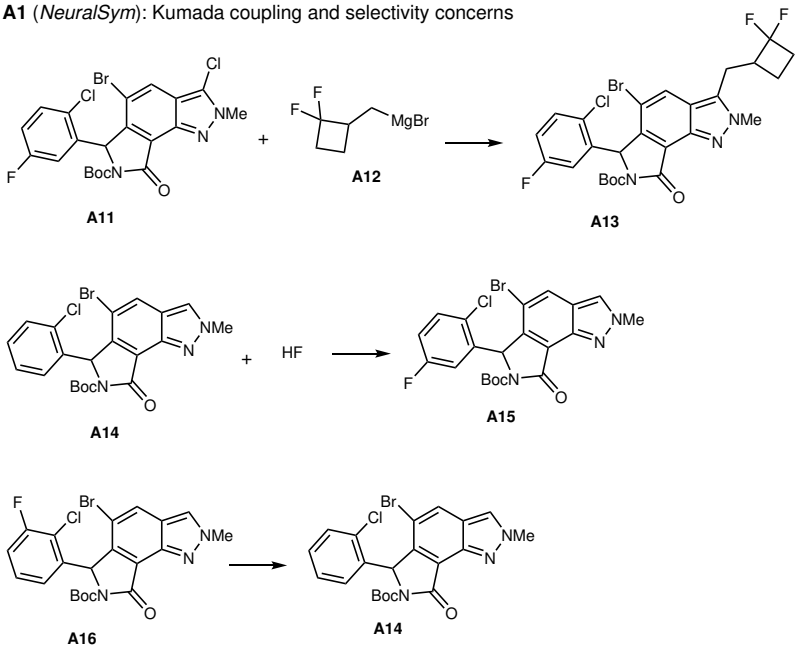
A1 (*RetroChimera*): Alkylation mediated via primary bromide



A1 (*NeuralLoc*): Friedel-Crafts acylations of electron-deficient arenes



A1 (*NeuralSym*): Kumada coupling and selectivity concerns

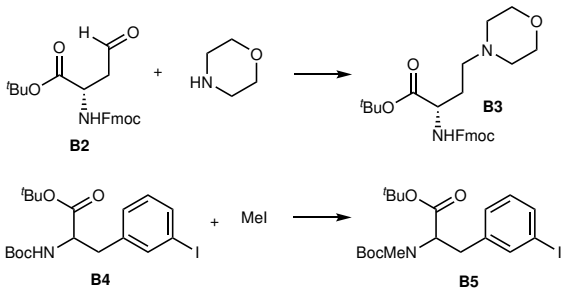


Supplementary Figure 2: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **A1**. DMB is 2,4-dimethoxybenzyl.

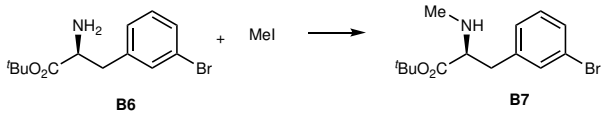
3.2 Route towards B1, C1, and D1

For **B1**, **C1**, and **D1**, models other than RetroChimera predicted in general routes with multiple *N*-alkylations of primary amines, which introduce the potential for overalkylation.[3] While chemically not implausible, those types of reactions typically require high excess of the substrate. In contrast, routes produced with RetroChimera circumvent this problem by substitution of these steps with either reductive aminations of **B2** and **D2** or *N*-alkylation of Boc protected amines **B4** or **C3** instead (Supplementary Figure 3).

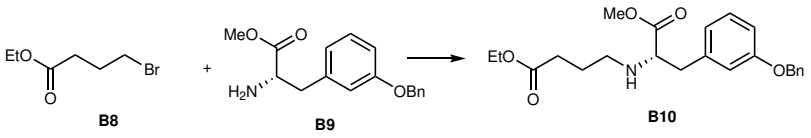
B1 (*RetroChimera*): Avoiding overalkylation



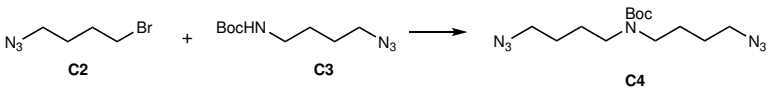
B1 (*NeuralLoc*): Potential overalkylation



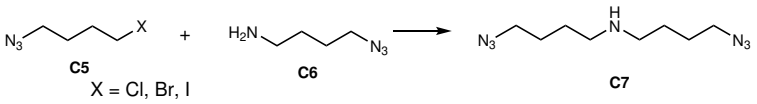
B1 (*NeuralSym*): Potential overalkylation



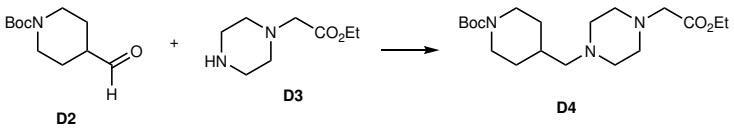
C1 (*RetroChimera*): Avoiding overalkylation



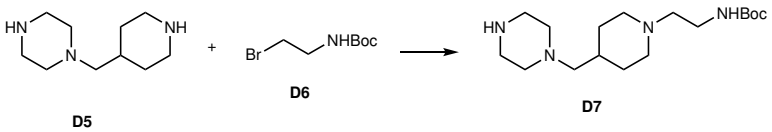
C1 (*Other models*): Potential overalkylation



D1 (*RetroChimera*): Avoiding overalkylation



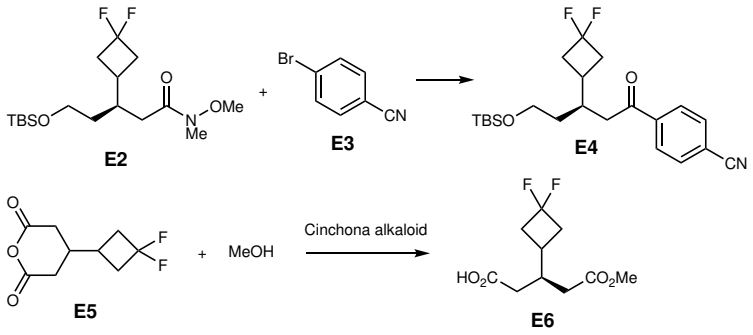
D1 (*NeuralSym*): Potential overalkylation and regioselectivity concerns



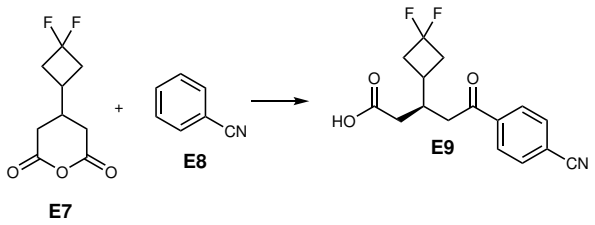
Supplementary Figure 3: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of targets **B1**, **C1**, and **D1**.

3.3 Route towards E1

E1 (*RetroChimera*): Addition to Weinreb amide and cyclic anhydride desymmetrization



E1 (*NeuralSym*): Friedel-Crafts acylations of electron-deficient arenes

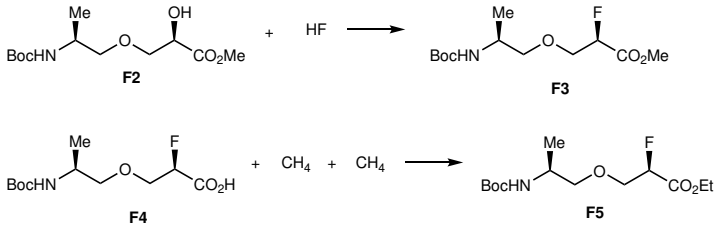


Supplementary Figure 4: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **E1**.

For **E1**, NeuralSym suggests Friedel-Crafts acylation to functionalize benzonitrile **E8** and introduce a stereocenter, which the raters flagged as problematic due to the low reactivity of electron-deficient arenes (Supplementary Figure 4). Furthermore, it is unclear whether such a transformation would be stereoselective or if separation of enantiomers would be required, which would decrease the overall yield further. In contrast, RetroChimera suggests introduction of this motif via Grignard addition to Weinreb amide **E2**, a much more reliable transformation. Interestingly, formation of one of the two stereocenters is suggested to proceed via enantioselective methanolysis of symmetric anhydride **E5**. This reaction is indeed precedented via catalysis of cinchona alkaloids[4] and is another example of a rare reaction identified by RetroChimera.

3.4 Route towards F1

F1 (*NeuralSym*): Poor reagent choice and hallucination example

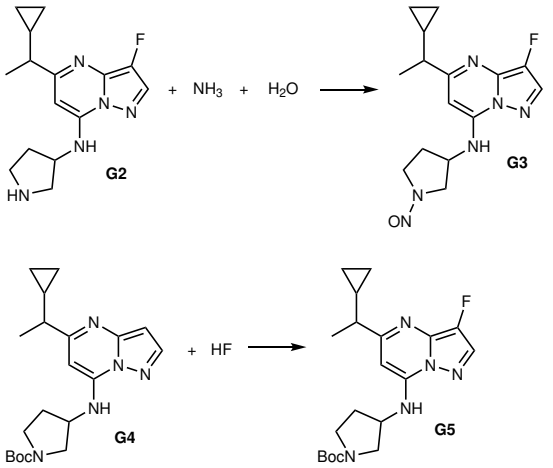


Supplementary Figure 5: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **F1**.

In the synthesis plan suggested for **F1** by NeuralSym, HF was suggested as reagent for the substitution reaction of secondary alcohol **F2** to fluoride **F3** under retention. The plan also features other hallucinations, such as methane molecules as reagents to form ethyl ester **F5** from carboxylic acid **F4** (Supplementary Figure 5).

3.5 Route towards G1

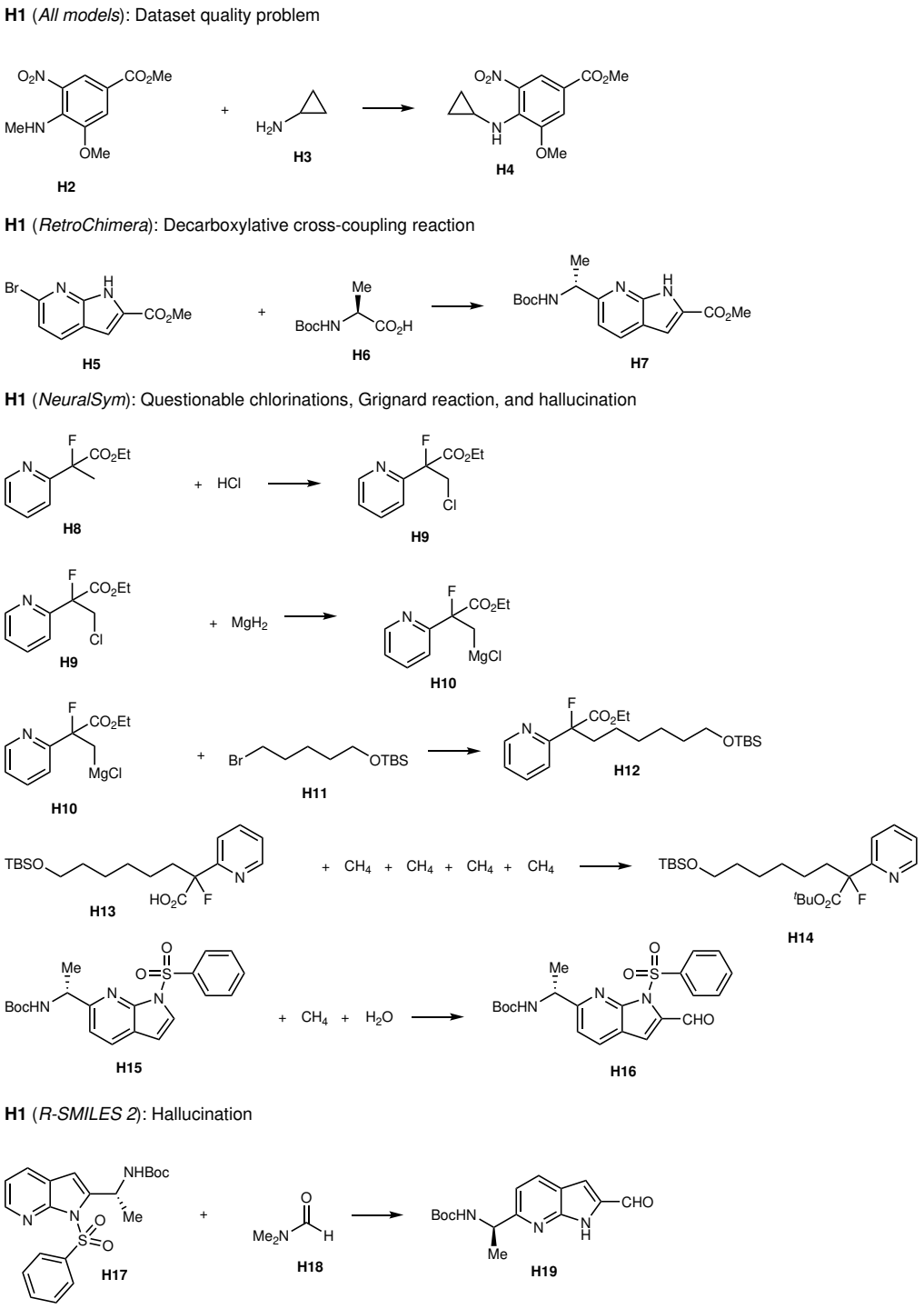
G1 (*NeuralSym*): Poor reagent choice and protecting group conflicts



Supplementary Figure 6: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **G1**.

For **G1**, NeuralSym suggested HF as reagent to introduce fluoride into arene **G4** in the presence of an acid labile Boc protecting group (Supplementary Figure 6). In the same route, ammonia and water were suggested as reagents for the nitrosylation of secondary amine **G2**. Conversely, RetroChimera correctly predicted sodium nitrite as nitrosylation agent.

3.6 Route towards H1

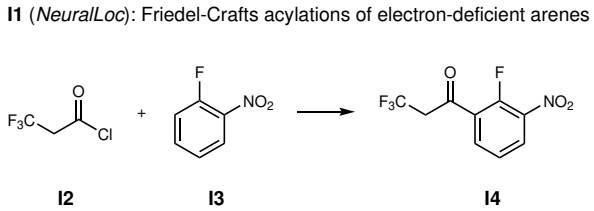


Supplementary Figure 7: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **H1**.

For **H1**, all models suffered from a nucleophilic aromatic substitution reaction of **H2** with amine **H3** predicted erroneously due to limited dataset quality (Supplementary Figure 7). While nucleophilic aromatic substitution with amines as leaving groups has minimal precedence,[5] the Pistachio training set gives many examples for this transformation, suggesting errors in the data pipeline. RetroChimera predicted otherwise a reasonable route and even suggested a decarboxylative cross-coupling reaction[6,7] to functionalize the heterocyclic core of **H5** (Supplementary Figure 7). We note that precedence of this reaction type is scarce (≈ 10 examples), which highlights again RetroChimera's strength to recall rare reaction types. It is noteworthy, however, that RetroChimera suggests inversion of the stereocenter, which is precedented via the absolute configuration of the nickel ligand.[7] Given the target only features a single stereocenter, however, scrambling of the configuration is acceptable and the correct enantiomer could be isolated via chiral separation.

R-SMILES 2 produced a hallucinated step during its route toward **H1** suggesting migration of a side chain, which we hypothesize may be introduced by subtle mistakes in *de novo* SMILES generation. For the same molecule, NeuralSym proposed multiple chlorinations under rather harsh conditions, Grignard formation to yield **H10** in the presence of an ester group, and multiple cases of hallucinations, which all invalidate the proposed route (Supplementary Figure 7).

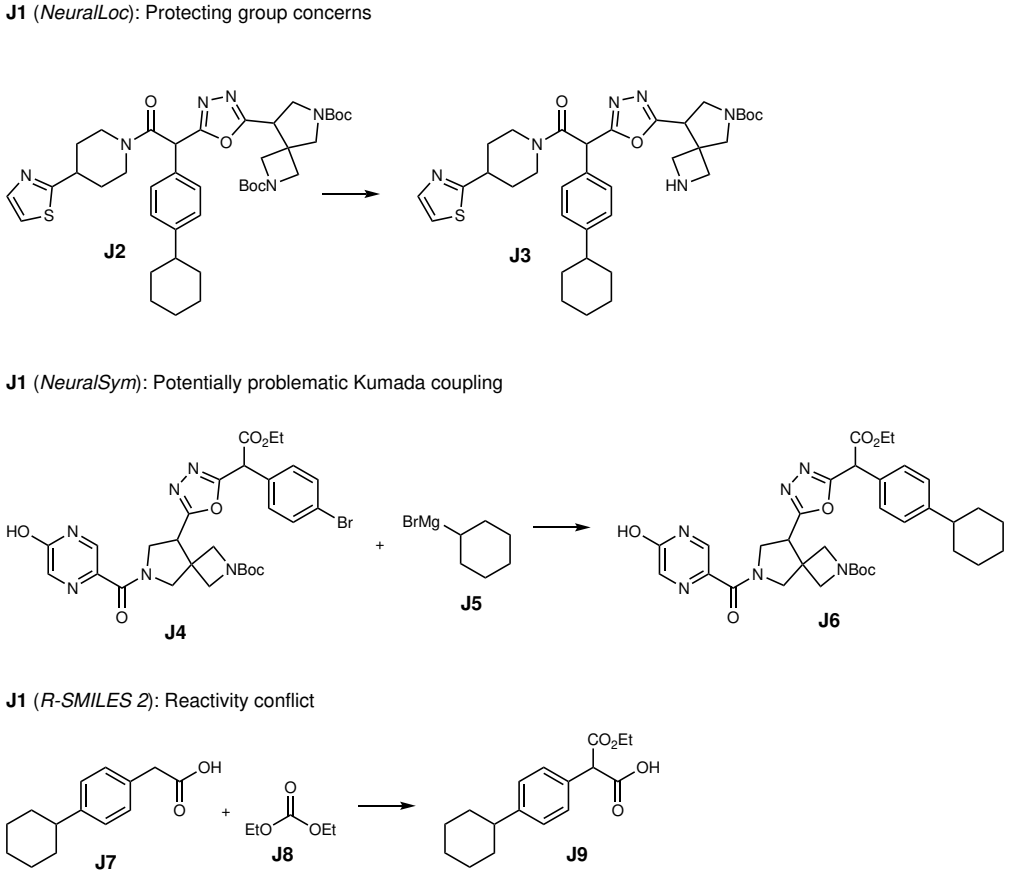
3.7 Route towards I1



Supplementary Figure 8: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **I1**.

For **I1**, all models proposed reasonable routes with the exception of NeuralLoc, which proposed Friedel-Crafts acylation of electron-poor arene **I3**, which all raters agreed is unlikely to proceed (Supplementary Figure 8).

3.8 Route towards J1

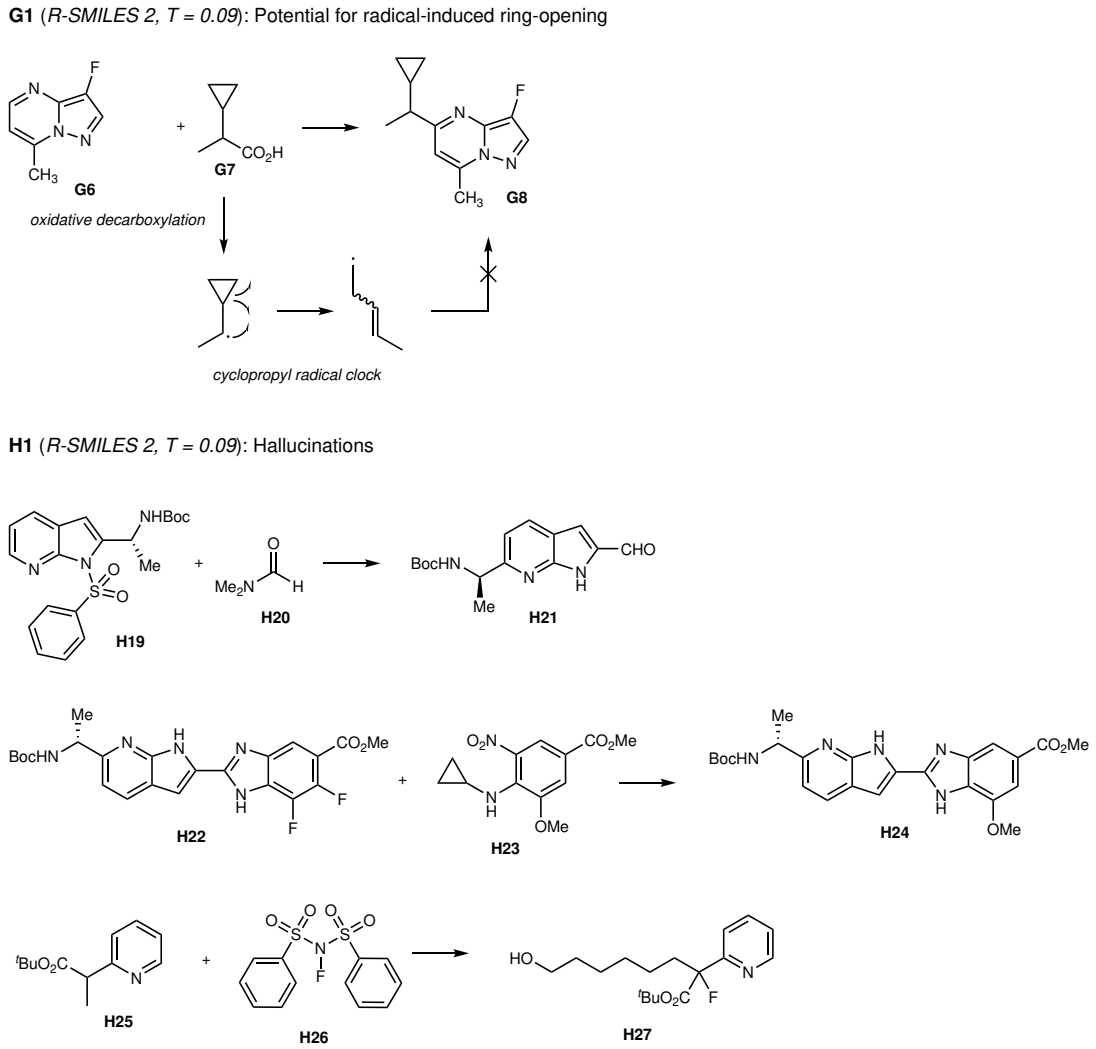


Supplementary Figure 9: Extended examples of accepted and rejected steps predicted by RetroChimera and other models for the synthesis of target **J1**.

For **J1**, NeuralLoc suggests selective removal of one out of two Boc protecting groups in **J2**, which is unlikely to proceed. For the same target, NeuralSym suggests a Kumada coupling between **J4** and **J5** in the presence of an ester and a phenol (Supplementary Figure 9). We note that there is some precedence of

Kumada coupling in the presence of such functional groups and that, as models do not give conditions, it is conceivable that the model actually predicts a Negishi cross-coupling reaction due to the implicit addition of zinc salts. However, it is also clear that a Kumada coupling with this functional group setting would not be first choice and was thus criticized by the raters.

3.9 Failure mode analysis and the effect of temperature

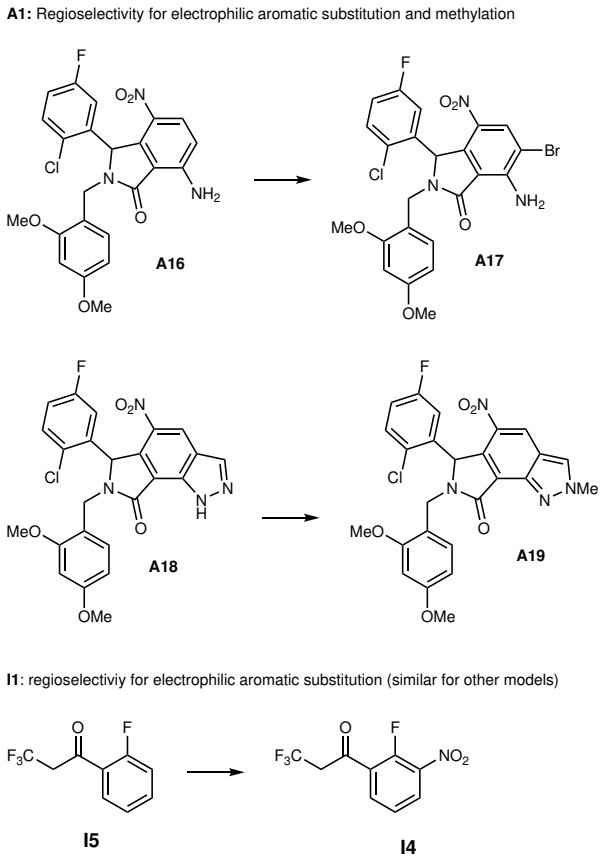


Supplementary Figure 10: Examples of rejected steps of R-SMILES 2 at $T = 0.09$.

We also explored different temperatures T for RetroChimera and R-SMILES 2 for the 1784 targets in the $\text{Pistachio}_{\text{Hard}}$ dataset (Supplementary Table 1). Consistent with previous experiments, tuning T increased the solve rate by approximately 1% for both models. We observed, however, that this increase in solve rate is anti-correlated with rater agreement of steps and routes accepted (Extended Data Table 3). RetroChimera with $T = 0.5$ showed a modest decline in the percentage of accepted steps but could not identify a route for target **A1**. Steps predicted by R-SMILES 2 at $T = 0.09$ were accepted less often (89.2% vs 85.0%) compared to those with standard temperature settings. As a consequence, only 4 instead of previously 5 routes were overall accepted. We found that for both models, optimizing T to increase number of targets solved increased the number of hallucinations and selectivity conflicts. Examples include cases of reactivity conflicts such as a Minisci reaction adjacent to a cyclopropyl moiety for the synthesis of **G1**, which is unprecedented due to the potential of ring-opening (cyclopropyl radical clock, Supplementary Figure 10). Other examples include chemically implausible and spontaneous appearances and migrations of entire substructures listed in the same figure. This finding corroborates findings of the preceding paragraph that highlight the importance of human expert ratings over solve rates alone.

From this analysis, we find that the most common failure modes for R-SMILES 2, especially at $T = 0.09$, are hallucinations and selectivity issues as well as reactivity conflicts. We also observed selectivity and reactivity conflicts for NeuralLoc. As shown in Supplementary Figures 2, 8, and 9, these conflicts mainly

arise when functional groups involved are relatively far away from each other, for instance in the cases of Friedel-Crafts acylations with mismatched electronics or selective Boc deprotections. All models introduced in this work, however, still mainly outperform NeuralSym, which gives most cases of hallucinations, selectivity and reactivity conflicts, and mechanistic plausibility.

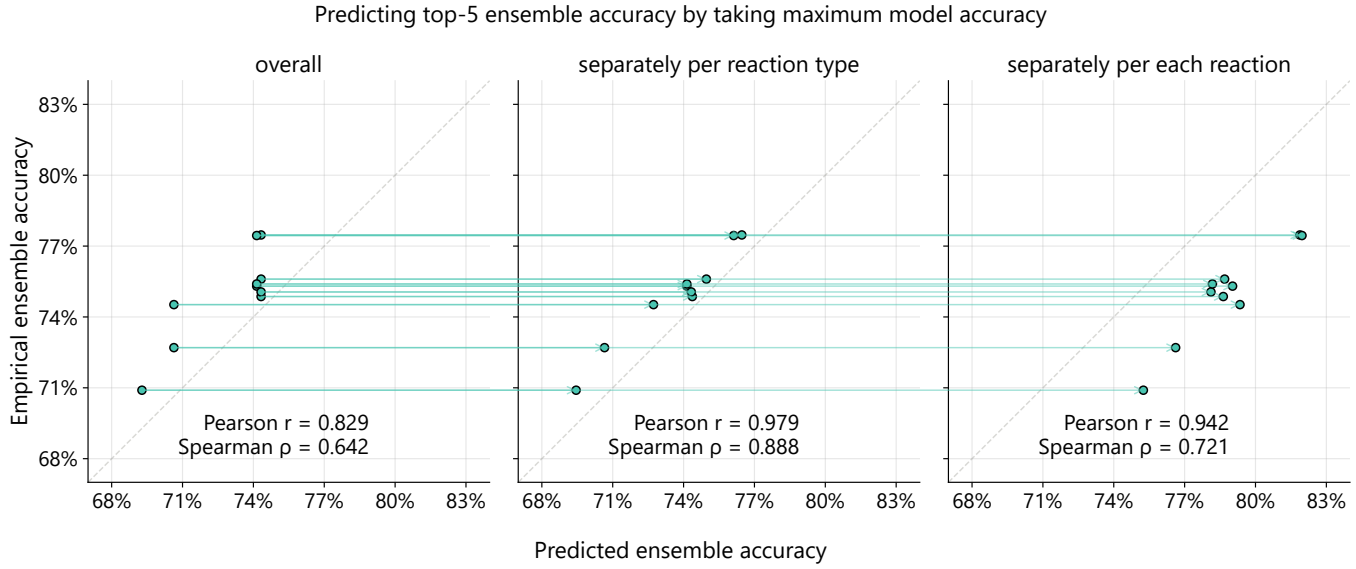


Supplementary Figure 11: Steps predicted by RetroChimera and rejected by raters.

RetroChimera's failures were mostly caused by subtle issues regarding chemo- and regioselectivity (Supplementary Figure 11). In almost all cases, raters were overall uncertain on the reaction outcome and product distribution and decisions were not unanimous, even after consulting reaction databases such as Reaxys and Scifinder. For instance, the product distribution of methylation of **A18** (N1 vs N2 methylation) is difficult to predict, and likely dependent on the specific substrate and conditions even if reports exist that are consistent with RetroChimera's predictions.[8] Similarly, raters and literature data of related systems agree that nitration of **I5** should give the *para*-substituted product over desired *ortho*-product **I4**. However, the exact ratio is unclear and the reaction may still be viable as shown in other reports.[9]

4 Ensemble accuracy prediction

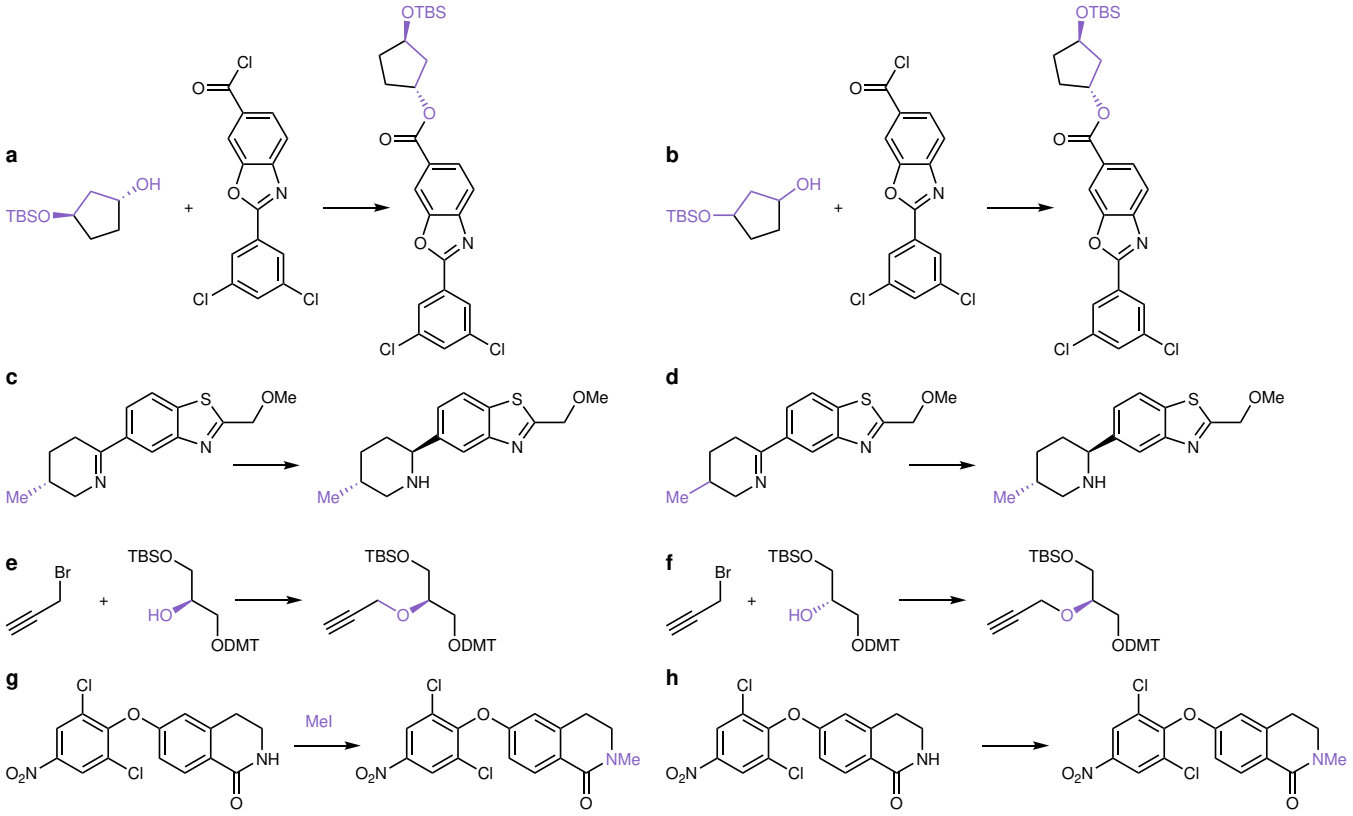
We assess the degree to which ensemble accuracy can be predicted from the accuracies of individual models (Supplementary Figure 12). Computing maximum accuracy over the ensembled submodels per each reaction type, and weighting the results according to frequency, is highly correlated to actual ensemble accuracy (Pearson correlation of 0.979). In contrast, maximum accuracy either at the entire dataset level, or individual reaction level, are less predictive, with the former usually underestimating ensemble performance, and the latter overestimating it. This suggests that accuracy split by reaction type provides a reliable "fingerprint" of the model strengths, and that this level of granularity is helpful in understanding how models interact when ensembled.



Supplementary Figure 12: Reliability of predicting top-5 accuracy for ensembles of pairs of Pistachio-based models given the accuracies of the two submodels. Prediction is made by considering the best of the two models (highest top-5 accuracy), either globally (left), separately for each reaction type (middle), or separately for each individual test reaction (right). Datapoints correspond to the 10 pairs of models from Extended Data Figure 2c; arrows connect points stemming from the same model pair. For each method, we report the Pearson and Spearman correlations between the predicted and actual ensemble accuracy. Gray diagonal line denotes perfect calibration ($y = x$).

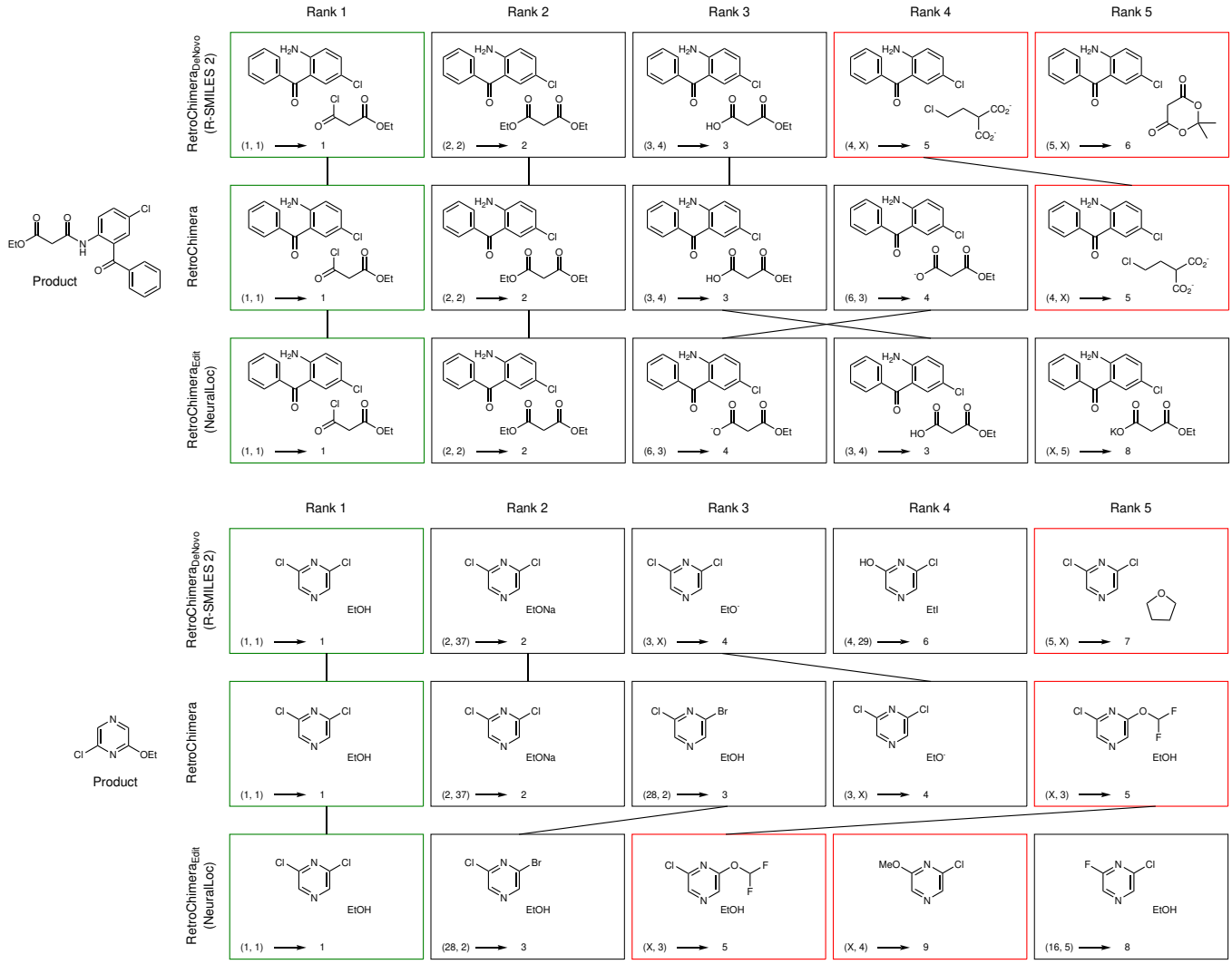
5 Denoising Behaviour

Denoising of potentially erroneous data In past work, AI retrosynthesis models have been criticized for their potential inability to deal with noise in the training data, which is expected to be present in large reaction datasets.[10] However, correctly trained probabilistic models, such as the ensemble component models in this work, can become robust towards errors in the training set, and effectively denoise the data. For this purpose, we qualitatively inspected random test reactions for given products from the Pistachio dataset that our model was unable to recover in its top 50 predictions, and asked expert chemists for an assessment (Supplementary Figure 13). We found that representative erroneous ground truth test reactions for example contain stereochemistry or other assignment errors, while our model returns the reactions that the chemist would expect, qualitatively suggesting the ability of the model to deal with partially noisy data.

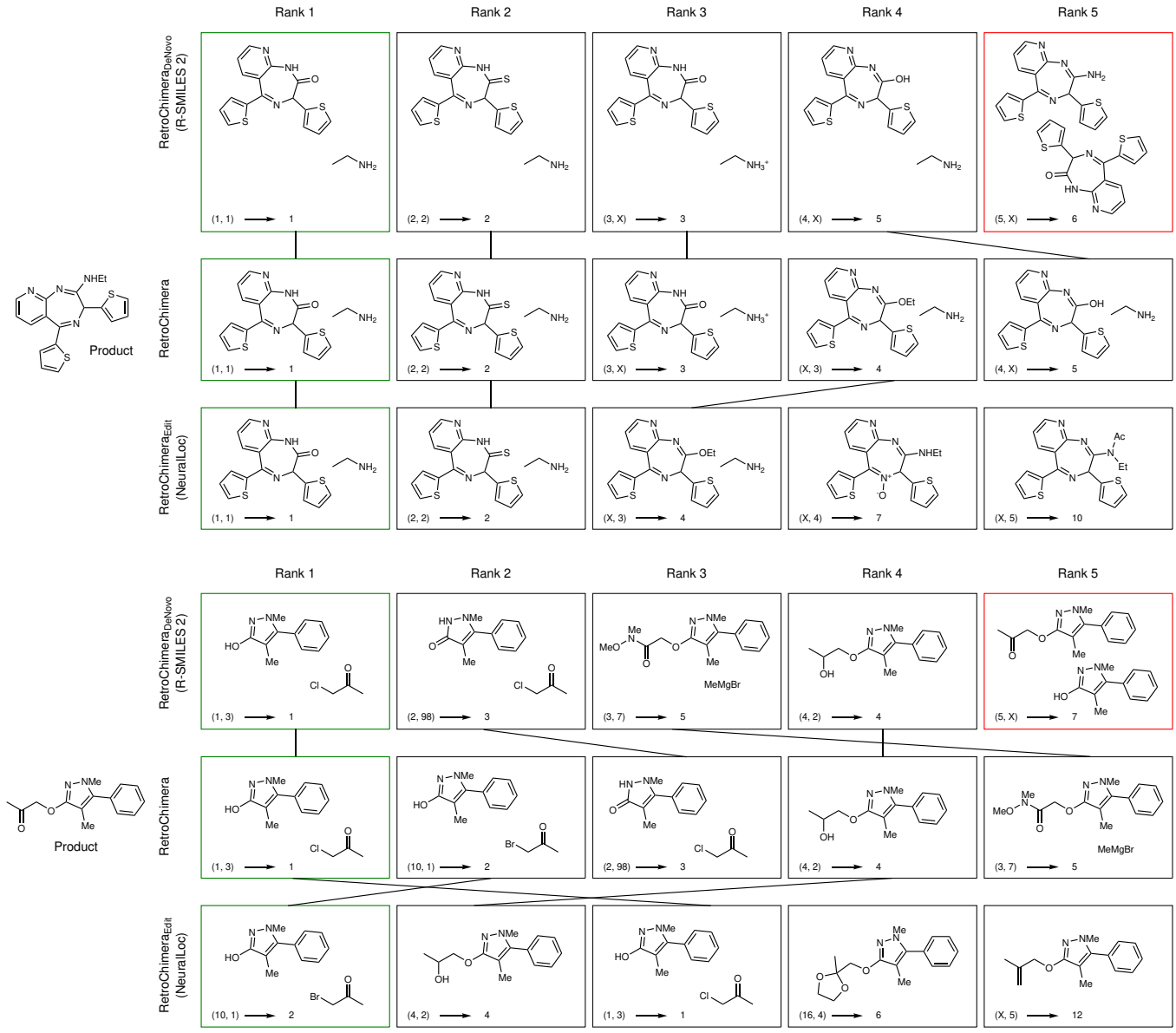


Supplementary Figure 13: Examples of the denoising behaviour of RetroChimera. Left column (**a**, **c**, **e**, **g**) are model predictions, in each case preferred by expert chemists. Ground truth examples (right column; **b**, **d**, **f**, **h**) are taken from test set. Examples **a**-**f** demonstrate denoising for stereochemistry assignments. Our model has learned to ignore likely incorrect assignments (right) and instead is aligned with expert expectations (left). Furthermore, the model exhibits the ability to infer missing reactants. In example **h** the test data does not specify the exact alkylating agent, whereas the model infers to use methyl iodide (Example **g**). `DMT` is dimethoxytrityl.

6 Worked examples for how the ensemble aggregates predictions



Supplementary Figure 14: Visualization of how predictions from R-SMILES 2 and NeuralLoc are combined. Molecule in row i and column j is the j-th reactant set predicted by the i-th model. `(A, B)` $\rightarrow$ `C` denotes that a prediction was rank `A` in the output of R-SMILES 2, rank `B` in the output of NeuralLoc, and rank `C` in the combined output (`X` signifies a prediction was not found in one of the lists). Segments connect molecules that are shared. Green box is ground-truth, red box highlights a hallucinated prediction which is chemically implausible. For certain inputs, the R-SMILES 2 model might predict bond-breaking reactions which are chemically implausible (ranks 4 and 5 in the top example; rank 5 in the bottom one). These cases are down-weighed during model ensembling as they are not predicted by NeuralLoc. In contrast, NeuralLoc can fail due to noise in the underlying data and incorrect template extraction (ranks 3 and 4 in the bottom example), which is in turn down-ranked by R-SMILES 2, highlighting the power of the ensembling approach.



Supplementary Figure 15: In certain cases, the R-SMILES 2 model appears to produce the same reactant twice, either as an exact copy or with minor variation.

7 Architectural, training and inference hyperparameters

Parameter	USPTO-50K	USPTO-FULL	Pistachio
Vocab size	72	235	346
Number of layers	6	6	8
Hidden dim	256	512	512
Feedforward dim	512	2048	2048
Number of heads	8	8	8
Number of KV heads	8	2	2
Batch size	128	128	512
Number of epochs	30	60	30
Learning rate scheduler	Noam	Noam	Noam
Learning rate	1.0	1.0	1.0
Warmup steps	8000	8000	8000
Dropout	0.3	0.1	0.1
Number of augmentations	20	5	10
Beam size	10	50	20
Total parameter count	17.4M	44.5M	66.7M

Supplementary Table 2: Architectural, training, and inference hyperparameters of the R-SMILES 2 model across the datasets investigated in this work.

	Parameter	USPTO-50K	USPTO-FULL	Pistachio
	d_{clf}	256	256	256
	d_{free}	0	32	32
	Number of templates	9735	228 127	146 256
GNN^{in}	Layer type	GPS + PNA	PNA	PNA
	Number of layers	3	5	5
	Hidden dim	64	768	1024
	Output dim (node-level)	256	128	128
	Output dim (graph-level)	512	1024	1024
	Aggregation heads	8	8	8
	Dropout (inter-layer)	0.1	0.0	0.05
	Dropout (post aggregation)	0.4	0.4	0.4
GNN^{tpl}	Layer type	GPS + PNA	PNA	PNA
	Number of layers	4	5	5
	Hidden dim	64	192	192
	Output dim (node-level)	256	128	128
	Output dim (graph-level)	512	512	512
	Aggregation heads	8	8	8
	Dropout (inter-layer)	0.1	0.0	0.0
	Dropout (post aggregation)	0.4	0.4	0.4
	Batch size	128	256	512
	Number of epochs	600	130	85
	Initial learning rate	10^{-3}	10^{-3}	10^{-3}
	Loss type	softmax	sigmoid	sigmoid
	r^{clf}	-	30	18
	r^{loc}	1	4	4
	r^{app}	100	10	10
	Total parameter count	1.9M	103M	165M

Supplementary Table 3: Architectural, training and inference hyperparameters of the NeuralLoc model across the datasets investigated in this work.

References

[1] Stansfield, I. *et al.* Heteroaromatic derivatives as NIK inhibitors (2018).

[2] Ma, C., Zhao, L., He, X., Jiang, Y. & Yu, B. Visible-light-induced direct 3-ethoxycarbonylmethylation of 2-aryl-2H-indazoles in water. *Org. Chem. Front.* **9**, 1445–1450 (2022).

[3] Salvatore, R. N., Yoon, C. H. & Jung, K. W. Synthesis of secondary amines. *Tetrahedron* **57**, 7785–7811 (2001).

[4] Chen, Y., Tian, S.-K. & Deng, L. A highly enantioselective catalytic desymmetrization of cyclic anhydrides with modified cinchona alkaloids. *J. Am. Chem. Soc.* **122**, 9542–9543 (2000).

[5] De Vargas, E. B., De Rossi, R. H. & Veglia, A. V. Amines as leaving groups in nucleophilic aromatic substitution reactions. *J. Org. Chem.* **51**, 1976–1981 (1986).

[6] Zuo, Z. *et al.* Merging photoredox with nickel catalysis: Coupling of α-carboxyl sp3-carbons with aryl halides. *Science* **345**, 437–440 (2014).

[7] Zuo, Z. *et al.* Enantioselective decarboxylative arylation of α-amino acids via the merger of photoredox and nickel catalysis. *J. Am. Chem. Soc.* **138**, 1832–1835 (2016).

[8] Baddam, S. R., Uday Kumar, N., Panasa Reddy, A. & Bandichhor, R. Regioselective methylation of indazoles using methyl 2, 2, 2-trichloromethylacetimidate. *Tet. Lett.* **54**, 1661–1663 (2013).

[9] Andreani, A. *et al.* Imidazo[2, 1-b]thiazole guanylhydrazones as rsk2 inhibitors. *Eur. J. Med. Chem.* **46**, 4311–4323 (2011).

[10] Mikulak-Klucznik, B. *et al.* Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).