Supporting Information

An exploration of dataset bias in single-step retrosynthesis prediction

Sara Tanovic, Ewa Wieczorek, Fernanda Duarte

Contents

S1 Data preprocessing

The USPTO database was downloaded from https://figshare.com/s/5e57a3399c52701cbc15.[1] The 2023Q1 version of the commercial Pistachio database[2] was used, and reactions were deduplicated and mapped with RXNMapper.[3]

The preprocessing pipeline includes the following steps:

- Remove multi-product reactions by either filtering out small side products (<6 atoms) or removing reactions with large side products.
- Remove reactions with purely inorganic products.
- Remove reactions where the product is present as a reactant.
- Remove "stereoalchemy" by removing any stereochemistry tokens if present in the product but not in the reactants.
- Remove reagents (precursor species which do not contribute to the atom mapping).
- Remove reactions with >4 reactants.
- Canonicalise reaction SMILES.
- Remove reactions with over 512 tokens.
- Remove duplicate reactions.
- Extract templates using LocalTemplate[4] and filter out templates with under 6 occurrences.

The pipeline was applied to the USPTO and Pistachio databases, yielding 1,103,781 and 3,720,288 reactions, respectively.

The Pistachio database was further filtered by patent number to exclude all US patents and avoid overlap with the USPTO. Templates with fewer than 20 occurrences were removed, and the database was divided into two sets: one containing templates also present in USPTO-retro, and the other containing new templates. A subset of 10,000 reactions was randomly sampled from both sets to generate Pistachio ID and Pistachio OOD, respectively.

The code used for cleaning and splitting both datasets can be found in the Github repository: `https://github.com/duartegroup/template-splits`.

S2 Splitting distributions

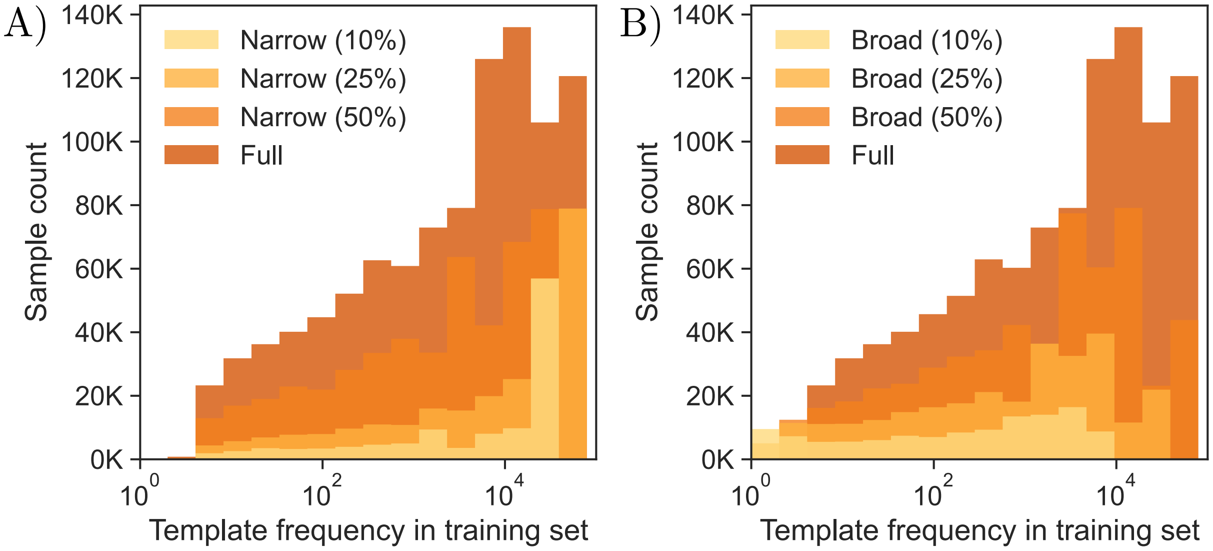

Figure S1: Histograms of sample counts in the training set by template frequency in the training set (on a log scale) across (A) the narrow split and (B) the broad split. Template frequency refers to the number of samples in the training set containing a specific template, while the sample count refers to the number of reactions with that template frequency.

S3 Training hyperparameters

The configuration files used to train the models can be found in the Github repository: `https://github.com/duartegroup/template-splits`. Each model was trained using its respective repository.

The LocalRetro models were trained using default hyperparameters, and an early stopping patience of 5 epochs was implemented. The MEGAN models were trained with default hyperparameters. The RootAligned models trained on the narrow and full splits used the default hyperparameters; however, those trained on the broad split have a separate optimised set of hyperparameters. This optimisation was done to improve the suboptimal performance observed when training with the default hyperparameters.

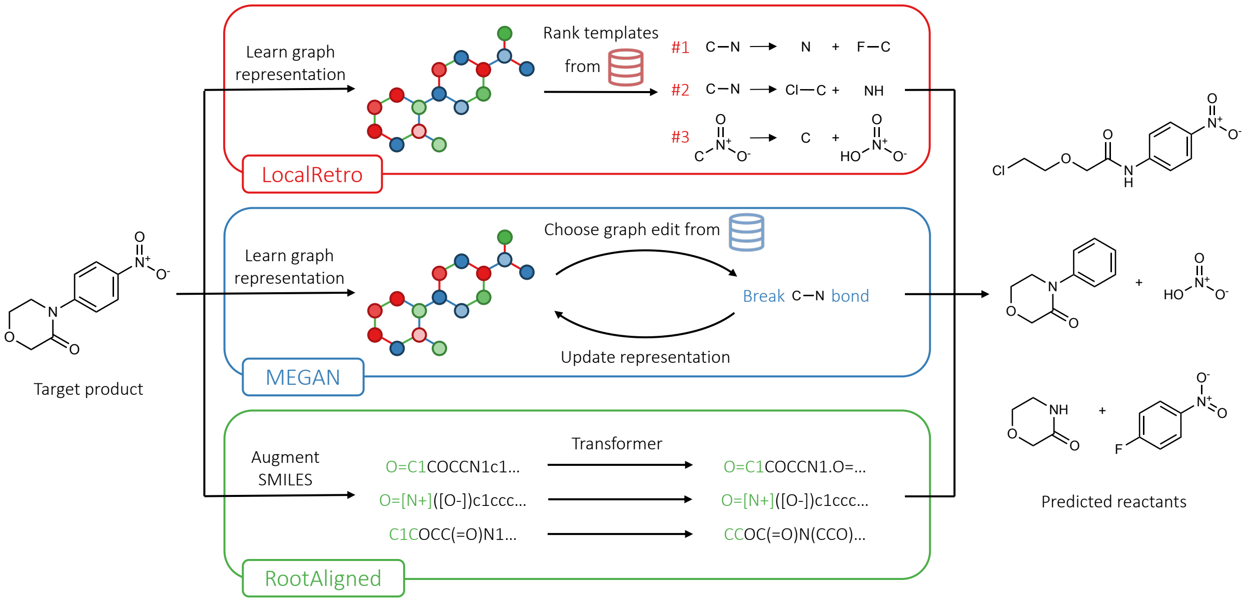

Figure S2: The three models used in this work: (i) LocalRetro[4] (red), a template-based algorithm that learns to choose the most suitable template from an extracted list of templates; (ii) MEGAN[5] (blue), a semi-template algorithm that formulates retrosynthesis a sequence of graph edits, and (iii) RootAligned[6] (green), a template-free algorithm that that treats retrosynthesis as a sequence-to-sequence translation task translating product SMILES strings into reactant SMILES.

S4 Evaluation metrics

The retrosynthesis platform *Syntheseus*[7] was used for evaluating all trained models. This involves automatically filtering out predictions which are invalid or duplicated, as shown in Figure S3. The evaluation metrics used include:

- **Top-k accuracy:** the proportion of test reactions with an exact ground truth match (i.e. the expected reactants from the test set) in the top-k predicted reactants.
- **Top-k round-trip accuracy:** the proportion of top-k predicted reactants which satisfy back-translation.[8] This is calculated by using a forward reaction model to predict the top-1 product of each set of predicted reactants, and this product is compared to the original target product.

A RootAligned forward prediction model is trained on the full USPTO-retro dataset for the purpose of back-translation and calculating top-k round-trip accuracy.

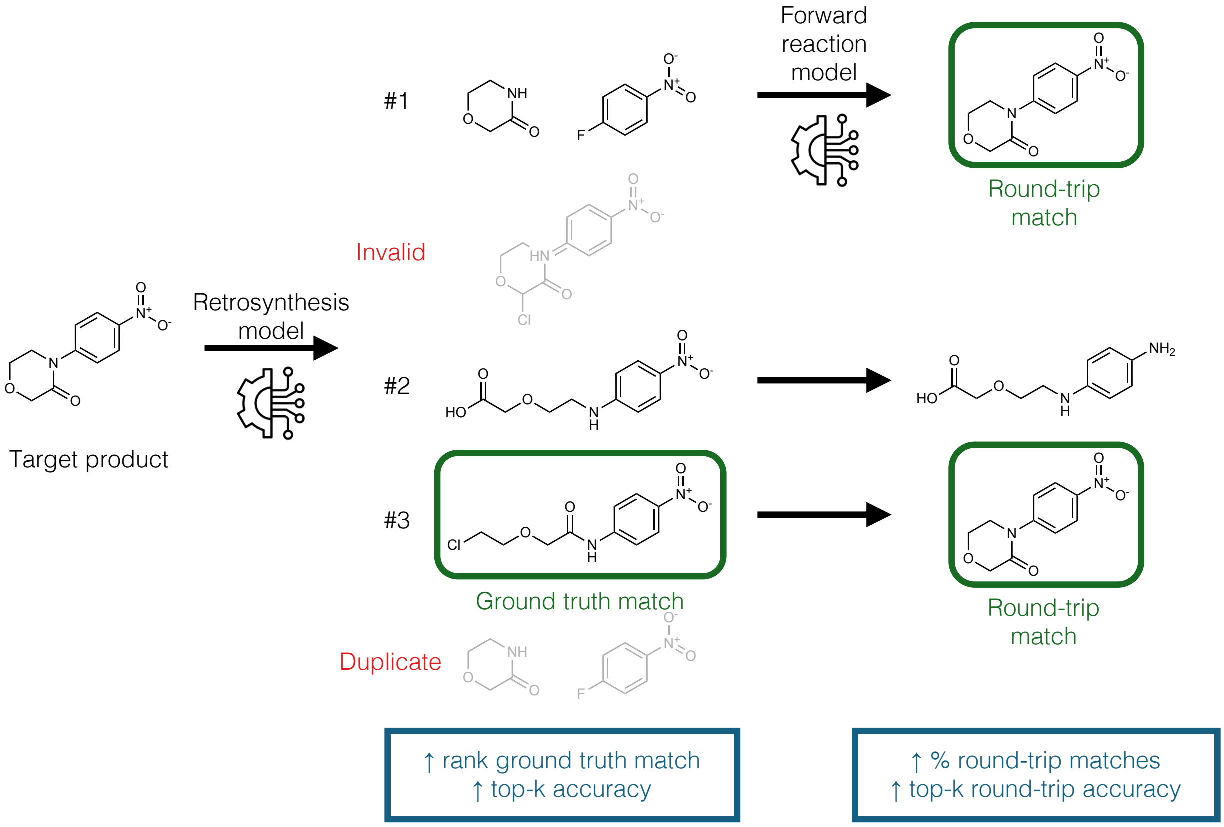

Figure S3: Visualisation of the retrosynthesis evaluation pipeline. Given a target product, the retrosynthesis model produces $n = 50$ sets of reactants, which are then filtered to contain only valid and unique predictions. The rank of the ground truth match determines the top-k accuracy. A forward reaction model is used to predict the product of each set of predicted reactants and this is compared to the target product to determine the top-k round-trip accuracy.

S5 Narrow split results

Table S1: Top-k accuracy and round-trip (RT) accuracy of models trained and tested on the narrow split.

		Top-k accuracy (%)					Top-k RT accuracy (%)	
Model	Training set (%)	k=1	5	10	20	50	1	5
LocalRetro	10	77.6	93.1	94.7	95.6	95.9	92.4	63.8
	25	74.2	91.6	93.4	94.6	95.1	93.2	70.3
	50	60.2	83.4	87.6	90.1	91.3	90.5	73.2
	90	50.5	77.4	83.1	86.7	88.5	90.8	78.0
MEGAN	10	76.9	91.8	93.3	94.6	95.1	91.8	58.6
	25	71.7	90.2	92.9	94.4	95.3	92.6	68.2
	50	53.2	78.0	83.6	87.1	89.6	83.6	61.2
	90	42.1	71.2	78.8	83.8	87.5	89.9	77.8
RootAligned	10	79.8	94.1	95.4	96.1	96.2	93.7	54.3
	25	76.2	92.8	94.5	95.4	95.5	94.5	63.1
	50	60.9	85.5	89.4	91.2	91.4	92.3	70.4
	90	49.1	78.8	85.0	87.6	87.8	92.1	75.2

S6 Broad split results

Table S2: Top-k accuracy and round-trip (RT) accuracy of models trained and tested on the broad split.

		Top-k accuracy (%)					Top-k RT accuracy (%)	
Model	Training set (%)	k=1	5	10	20	50	1	5
LocalRetro	10	42.1	71.8	79.6	84.8	87.4	88.3	75.4
	25	45.3	74.0	81.0	85.5	87.9	90.0	78.0
	50	48.3	75.9	82.1	86.3	88.3	90.5	78.2
	90	50.5	77.4	83.1	86.7	88.5	90.8	78.0
MEGAN	10	40.5	69.5	77.3	82.6	86.3	88.5	75.0
	25	40.2	69.2	77.0	82.5	86.3	88.4	75.0
	50	41.3	70.7	78.3	83.5	87.2	89.9	77.8
	90	42.1	71.2	78.8	83.8	87.5	89.9	77.8
RootAligned	10	29.8	58.9	70.5	77.7	79.1	77.7	62.3
	25	23.5	49.4	61.5	70.2	71.9	77.5	63.7
	50	18.3	43.1	54.8	62.6	63.9	67.0	59.1
	90	49.1	78.8	85.0	87.6	87.8	92.1	75.2

S7 Tanimoto similarity between training and test sets

Tanimoto similarity was used to calculate the similarity between product molecules present in the USPTO-retro training and test sets. Morgan fingerprints with default RDKit parameters were calculated for all product molecules using RDKit, and Tanimoto similarity was calculated between all pairs of fingerprints from the training and test sets. Molecules were deemed to be similar if the Tanimoto similarity score was over 0.4, and the total count of similar molecules in the training set for a given test set product molecule was collected. The counts were then divided by the total number of reactions in the training set to get the percentage similarity (Figure S4).

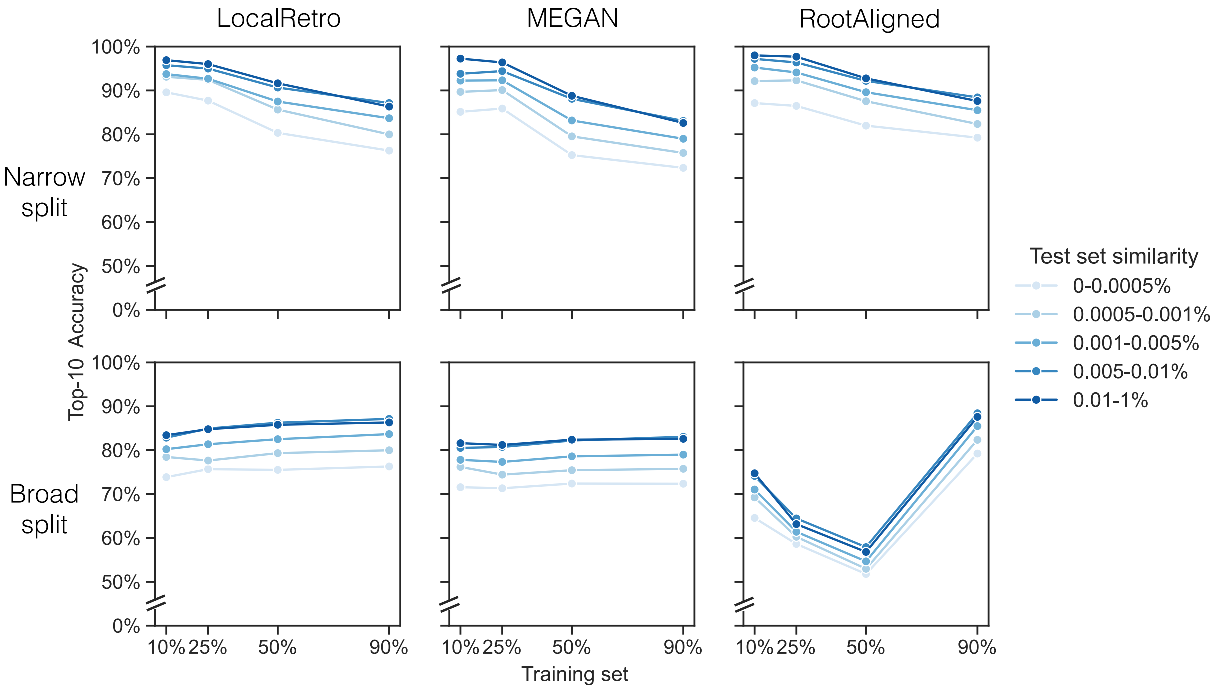

Figure S4: Top-10 accuracy of models trained and tested on splits of USPTO-retro, with test set reactions binned by their percentage similarity to the training set product molecules. Similarities are calculated as the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

S8 ID results

Tanimoto similarity was used to calculate the similarity between product molecules present in the full USPTO-retro training and the USPTO-retro test set and Pistachio ID test set, using the same method as discussed in Section S7. Figure S5 shows that Pistachio ID has a high proportion of products with 0% similarity to the training set, and is overall less similar to the training set than the USPTO-retro test set. Figure S6 shows the resulting top-10 accuracy of models tested on Pistachio ID, which suggests that the lowered accuracy of these models is due to the increased proportion of dissimilar test products.

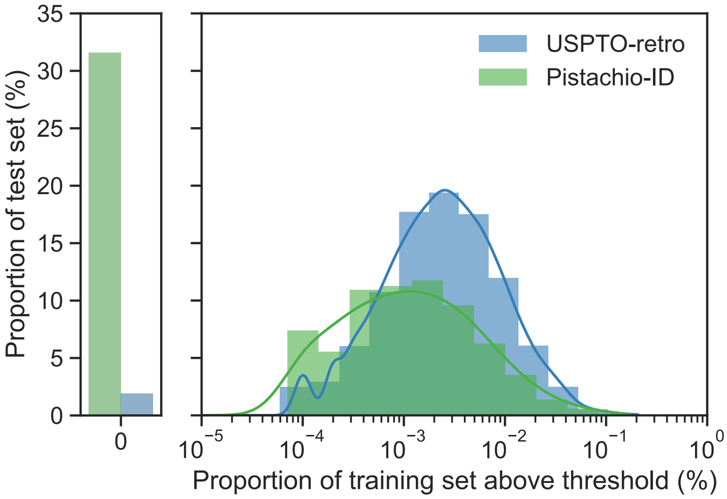

Figure S5: Similarity of the USPTO-retro (blue) and Pistachio ID (green) test sets to the USPTO-retro full training set. Similarities are calculated as the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

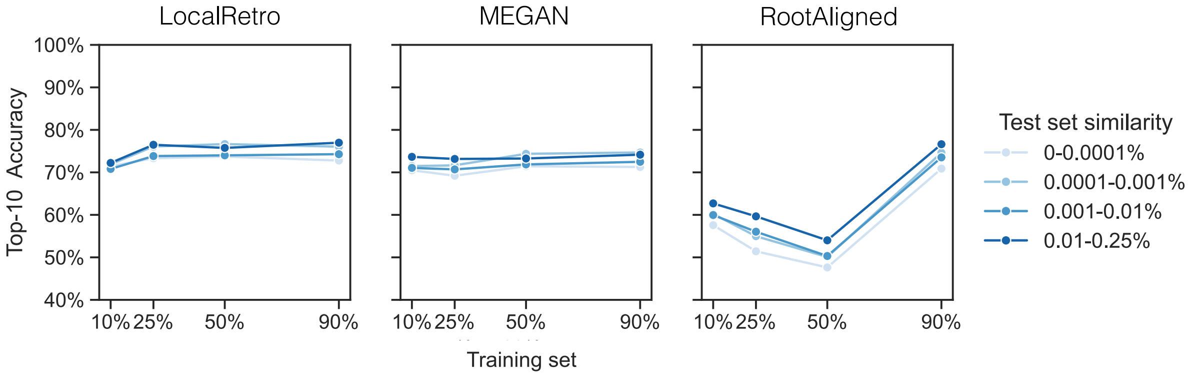

Figure S6: Top-10 accuracy of models trained on the broad splits of USPTO-retro and tested on Pistachio ID, with test set reactions binned by their percentage similarity to the training set product molecules. Similarities are calculated by the proportion of pairwise Tanimoto similarities over a threshold score of 0.4.

Table S3: Top-k accuracy of models trained on the broad split and tested on the Pistachio ID test set.

		Top-k accuracy (%)				
Model	Training set (%)	k=1	5	10	20	50
LocalRetro	10	35.6	62.9	71.4	77.0	80.4
	25	38.1	66.1	74.2	79.1	82.2
	50	38.7	67.0	74.7	79.9	82.9
	90	39.5	67.5	74.5	79.7	82.8
MEGAN	10	35.4	62.9	71.0	76.2	79.8
	25	35.5	62.7	70.4	75.9	79.9
	50	37.4	64.7	72.5	77.9	81.5
	90	37.6	65.1	72.9	78.1	82.2
RootAligned	10	21.6	47.6	58.9	66.2	67.7
	25	20.1	43.0	54.0	62.4	64.0
	50	17.2	39.1	49.5	56.9	58.1
	90	38.8	66.5	73.2	76.5	76.8

S9 OOD results

Table S4: Top-k accuracy of models trained on the narrow split and tested on the Pistachio OOD test set.

		Top-k accuracy (%)				
Model	Training set (%)	k=1	5	10	20	50
LocalRetro	10	0.08	0.20	0.20	0.22	0.23
	25	0.02	0.08	0.10	0.13	0.20
	50	0.05	0.10	0.12	0.18	0.20
	90	0.03	0.07	0.07	0.15	0.20
MEGAN	10	0.03	0.27	0.45	0.71	1.03
	25	0.03	0.22	0.48	0.93	1.40
	50	0.04	0.40	0.83	1.32	2.22
	90	0.02	0.59	1.01	1.53	2.28
RootAligned	10	0.05	0.22	0.51	0.79	0.99
	25	0.21	0.69	1.07	1.43	1.62
	50	0.10	1.14	1.84	2.51	2.71
	90	0.28	1.08	1.89	2.97	3.23

References

(1) Gil, V. S.; Bran, A. M.; Franke, M.; Schlama, R.; Luterbacher, J. S.; Schwaller, P. In *NeurIPS 2023 AI for Science Workshop*; arXiv: 2312.09004v1, 2023.

(2) Mayfield, J.; Lagerstedt, I.; Sayle, R. *Pistachio "Fantastic reactions and how to use them"*; tech. rep.; 2021.

(3) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. *Extraction of organic chemistry grammar from unsupervised learning of chemical reactions*; tech. rep., Publication Title: Sci. Adv Volume: 7 Issue: 7; 2021.

(4) Chen, S.; Jung, Y. *JACS Au* **2021**, *1*, 1612–1620.

(5) Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowski-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. *J. Chem. Inf. Model.* **2021**, *61*, arXiv: 2006.15426 Publisher: American Chemical Society, 3273–3284.

(6) Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; Song, M. *Chem. Sci.* **2022**, *13*, arXiv: 2203.11444 Publisher: Royal Society of Chemistry, 9023–9034.

(7) Maziarz, K.; Tripp, A.; Liu, G.; Stanley, M.; Xie, S.; Gainski, P.; Seidl, P.; Segler, M. *Faraday Discuss,* **2024**, arXiv: 2310.19796v1, –.

(8) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. *Chem. Sci.* **2020**, *11*, Publisher: Royal Society of Chemistry, 3316–3325.