Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Material for Solvent Selection for Polymers Enabled by Generalized Chemical Fingerprinting and Machine Learning

Joseph Kern, Shruti Venkatram, Manali Banerjee, Blair Brettmann, and Rampi Ramprasad*

School of Materials Science and Engineering, Georgia Institute of Technology

771 Ferst Drive NW, Atlanta, Georgia - 30332, USA.

E-mail: rampi.ramprasad@mse.gatech.edu

Dataset

Originally, 6,282 polymers and 58 solvents were collected. A subset of 3,373 polymers and 51 solvents is shown in Table S1 (with the solvent names and polymer counts), and a similar subset of 2,909 polymers and 7 solvents is shown in Table S2.

Table S1: Solvents in the training data with their counts of total, soluble, and insoluble polymer pairs.

Solvent	Total	Soluble	Insoluble
chloroform	2806	2138	668
THF	2375	1872	503
methanol	2187	159	2028
DMF	2152	1803	349

DMSO	1794	1451	343
nMP	1622	1507	115
DMAc	1522	1371	151
3-methylphenol	1289	1099	190
propan-2-one	1211	459	752
toluene	967	621	346
benzene	933	602	331
oxidane	898	107	791
pyridine	799	572	227
sulfuric acid	787	774	13
ethanol	785	228	557
DCM	734	580	154
ethyl acetate	446	255	191
acetonitrile	404	222	182
1,4-dioxane	371	284	87
formic acid	360	258	102
1,1,1,2-tetrachloroethane	354	285	69
chlorobenzene	342	221	121
tetrachloromethane	329	178	151
hexane	325	33	292
heptane	271	136	135
1,2-dichloroethane	259	219	40
2,2,2-trifluoroacetic acid	254	231	23
butan-1-ol	235	125	110
cyclohexanone	221	162	59
propan-1-ol	221	114	107
propane-1,2,3-triol	216	114	102

2-methylpropan-2-ol	210	113	97
acetic acid	195	152	43
nitrobenzene	142	80	62
cyclohexane	140	40	100
phenol	140	138	2
chloromethane	132	94	38
trichloro(fluoro)methane	132	94	38
ethoxyethane	124	9	115
butan-2-one	118	65	53
1,4-xylene	105	61	44
phenyl hypochlorite	75	39	36
chlorane	66	63	3
2,2,2-trifluoroethanol	53	46	7
formamide	46	43	3
1,2-dichlorobenzene	39	35	4
N-[bis(dimethylamino)phosphoryl]-N-methylmethanamine	33	31	2
1,2,3,4,4a,5,6,7,8,8a-decahydronaphthalene	28	15	13
oxolan-2-one	26	22	4
carbon disulfide	25	14	11
1,2,3,4-tetrahydronaphthalene	24	16	8

Table S2: Solvents in the held out data with their counts of total, soluble, and insoluble polymer pairs.

solvent	Total	Soluble	Insoluble
2,2-dichloroacetic acid	135	135	0

methanesulfonic acid	61	61	0
nitric acid	32	32	0
4-chlorophenol	32	32	0
1,2,3,4,5,6-hexafluorobenzene	15	15	0
1,1,2-trichloro-1,2,2-trifluoroethane	12	12	0
1,2,4-trichlorobenzene	12	12	0

This subset was chosen based on the number of soluble and insoluble combinations. If a solvent did not have an instance of being soluble and an instance of being insoluble it was removed. The same was done for the polymers. Figure S1 shows a graphic representation of the polymers. Each polymer has a certain number of combinations with various solvents (the colorbar) and a certain percentage of those combinations are soluble pairs (y-axis). Those encapsulated by a red square are the ones removed from the training data set.

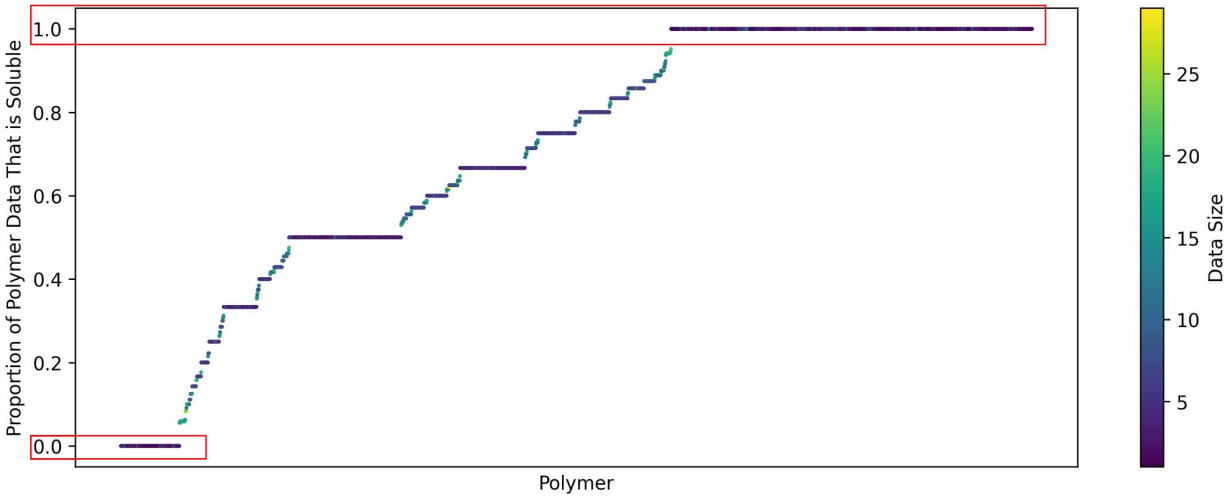

Figure S1: Proportion of polymer-solvent combinations that are soluble pairings per polymer. The color of the polymer data point represents the number of solvents the polymer is paired with. Those polymers encapsulated by the red square are the ones removed from the training data set.

The reason these polymers and solvents were removed was due to worry that the model would learn the specific polymer's or solvent's class as opposed to the interaction between the polymers and solvents. We believed this would skew the metrics used to assess model

performance, i.e., the model would always predict the correct class for these polymers or solvents, artificially inflating the recall or f1-score. If a split by solvent was used, the polymers might inflate the score. If a split by polymer was used, the solvents might inflate the score. Thus, both were removed.

The model might also fail to correctly predict the polymer's or solvent's solubility with a new combination. To test this, we trained a random forest machine learning model on the entire data set of 6,282 polymers and 58 solvents. We took 2,909 polymers that were only one class and used the model to predict their solubility in all 58 solvents. The color of each polymer corresponded to the average confidence of the random forest model in its prediction (which is the number of trees that predict the polymer is either soluble or insoluble over the total number of trees). These results are plotted in Figure S2.

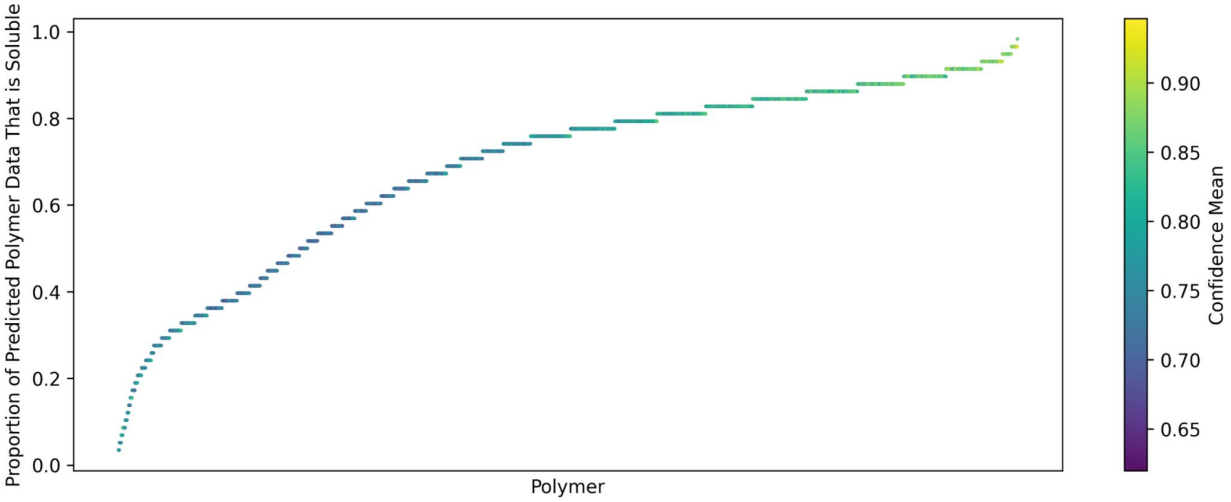

Figure S2: Proportions of polymers predicted as soluble with 58 solvents. The color of each data point represents the average model confidence in the prediction for all solvents. The model was trained on the full data set of 6,282 polymers and 58 solvents. The 2,909 polymers that contained only one class are plotted on the x-axis.

For 1,155 of the polymers, the model predicted it was soluble in at least 46 of the 58 of solvents with a average confidence 0.8. This seems unlikely though, as a visual screening of solvent functional groups indicated only 43 of the solvents are polar, and we would expect solubility to be limited based on the "like-dissolves-like" rule of thumb. The fact some

polymers are expected to dissolve in all solvents is especially suspect. This was why we decided to remove the solvents and polymers with only one class.

Class Imbalance

The f1 score is highly affected by class imbalance because precision is affected by class imbalance. To show this, we generated 10,000 data points labelled as either class 0 or class 1. We generated a fake prediction for the data points that was accurate for 75% of the data for each class. We did this 98 times, varying the proportion of class 0 data from 1% to 99% of data. The resutls are plotted in Figure S3.

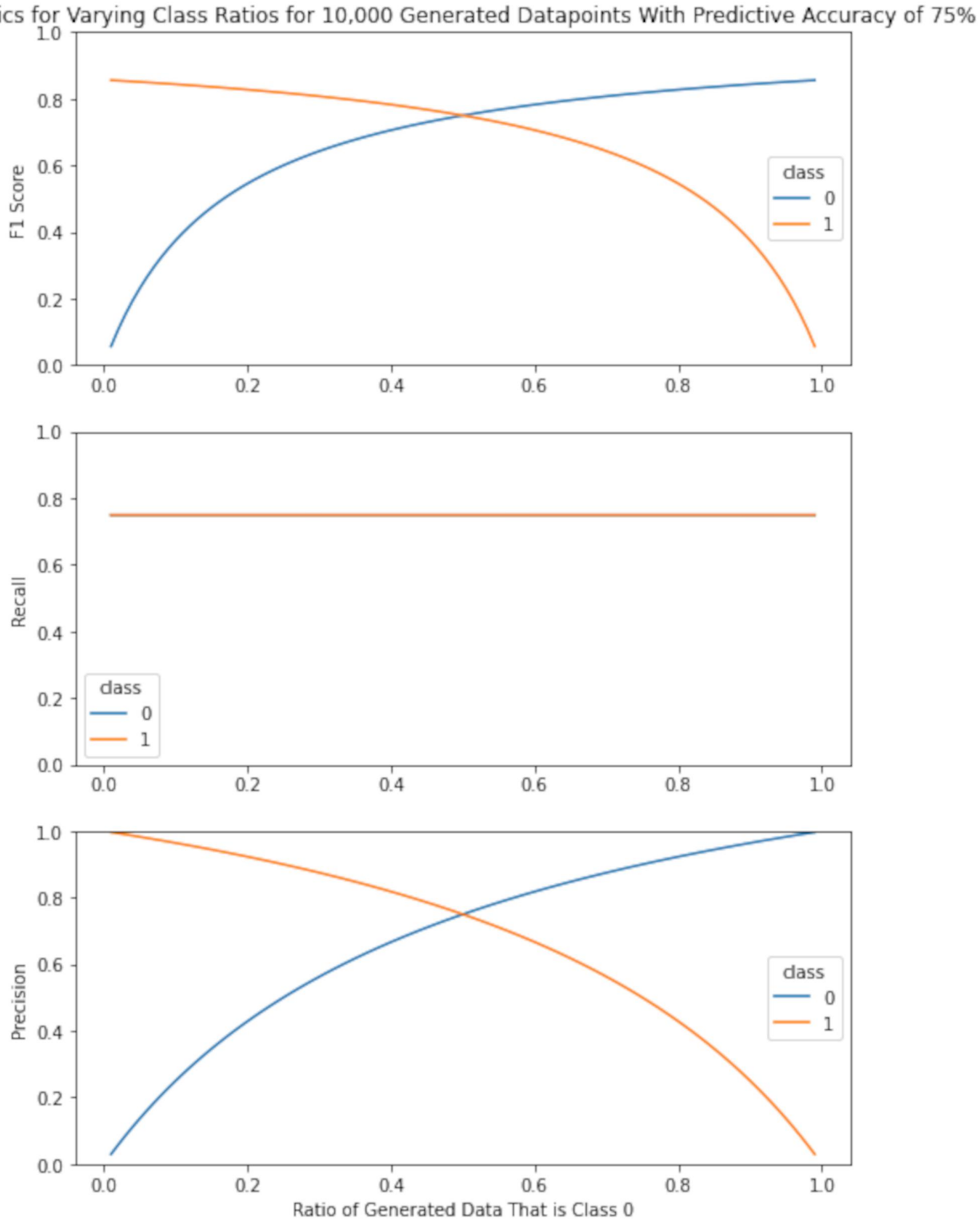

Figure S3: 10,000 data points were generated with varying proportion of class 0 and 1. 75% of the data was classified correctly for each class, while 25% was misclassified. The effect of class imbalance on f1 score (top), recall (middle), and precision (bottom) is plotted. The y-axis is the metric the imbalance is being simulated for, the x-axis is the proportion of simulated data that is class 0.

As we can see, recall is not biased by the proportion of class data, but precision and f1 score are. Since the precision is the true positive over the sum of true Positive and false Positive, when there is more negative data there is higher likelihood for false positives, skewing the results. Since recall only accounts for the correct and incorrect classifications of the positive data, it is not affected by the proportion of negative data. Based on our analysis of Figure S3, the bias does not affect f1 score too dramatically until one class makes up over 65 to 70% of data.

As stated in the manuscript, this is the case for the test data of two splits in the solvent splitting method (70 and 75% of the test data is soluble vs 41, 50 and 50% for the other three splits). However, an analysis of recall, seen in Figure S4, reveals there is still a large variation in the recall between splits when using a group split by solvent.

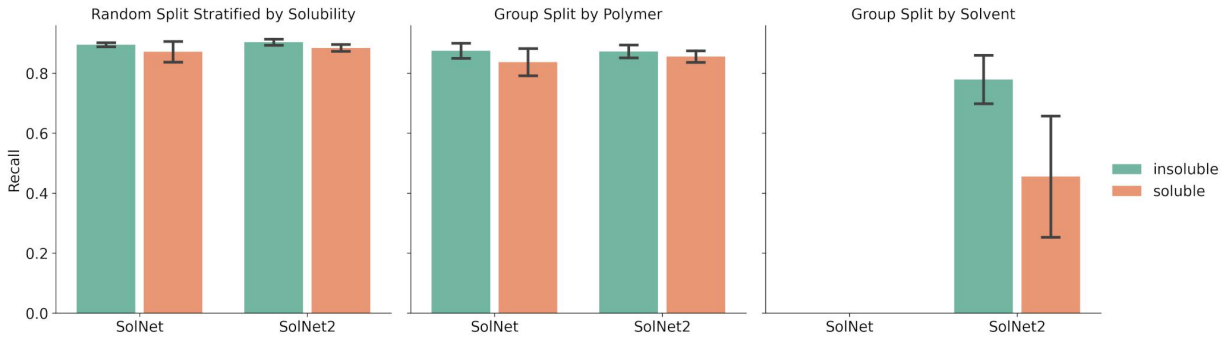

Figure S4: Average recall of SolNet infrastructure models for soluble and insoluble classification using either a one-hot encoding for solvents or a structural fingerprint. Five-fold cross validation splits were chosen using either a random split stratified by solubility (left), group split by polymer (middle), or group split by solvent (right). Error bars represent the standard deviation for the F1 score of those splits.

Tanimoto Similarity

$$\text{Tanimoto(x, y)} = \frac{\sum x_i y_i}{\sum x_i^2 + \sum y_i^2 - \sum x_i y_i} \tag{1}$$

where x and y are either both polymers or both solvents and i is their ith fingerprint. A Tanimoto score of 0 means x and y are completely different while a value of 1 indicates they are the same.

Learning Curve

To assess the random forest model's performance on completely unseen solvents and polymers, we did a learning curve analysis. First, we used the leave-one-out method to generate 51 splits of the data where one solvent and all of its associated polymers were held out as a test set. To start each learning curve, we trained a random forest model with one solvent and all of its polymers and tested the model on the held out solvent-polymer pairings. Then, we added another solvent and its polymers and ran the analysis again. This was done 50 times until each solvent and their polymers were added to the training model. Solvents were added based on the number of polymer pairings they had, so the solvents with the fewest pairings were added first. This was done 51 times so each solvent and its polymers were the test pairings at least once. For instance, chloroform and all 2,806 polymers would be removed for the test set, and then a model would be trained with 1,2,3,4-tetrahydronaphthalene and all of its polymers (excluding any that were in the chloroform's set) and tested on chloroform's set. After, a new model would be trained with both 1,2,3,4-tetrahydronaphthalene and carbon disulfide with all of their polymers and tested on the chloroform set. This would continue until all 50 solvents were added to the training data (in order of smallest to largest data set size), then it would be repeated with a new test solvent.

The results of this analysis are shown in Figure S5. Recall is used as the metric due to the large test set imbalance in classes for some solvents. Figure S5a shows the average $\pm$ standard deviation recall for predictions on all solvents (when used as the test set) as a function of the number of solvents in the training data set. Figure S5b shows the recall when methanol is the test solvent (with 1,915 test polymers) as a function of data set size. It also plots the proportion of test and training data that is soluble as a function of data set size. Figure S5c plots the same information for when chlorobenzene is the test solvent (with 226 test polymers). The discrepancy in data set size between S5b and S5c is because methanol has more test polymers. All polymers are removed from the training data, which results in a variable amount of data loss depending on how many solvents each polymer has been paired

with and how many polymers are in the test set.

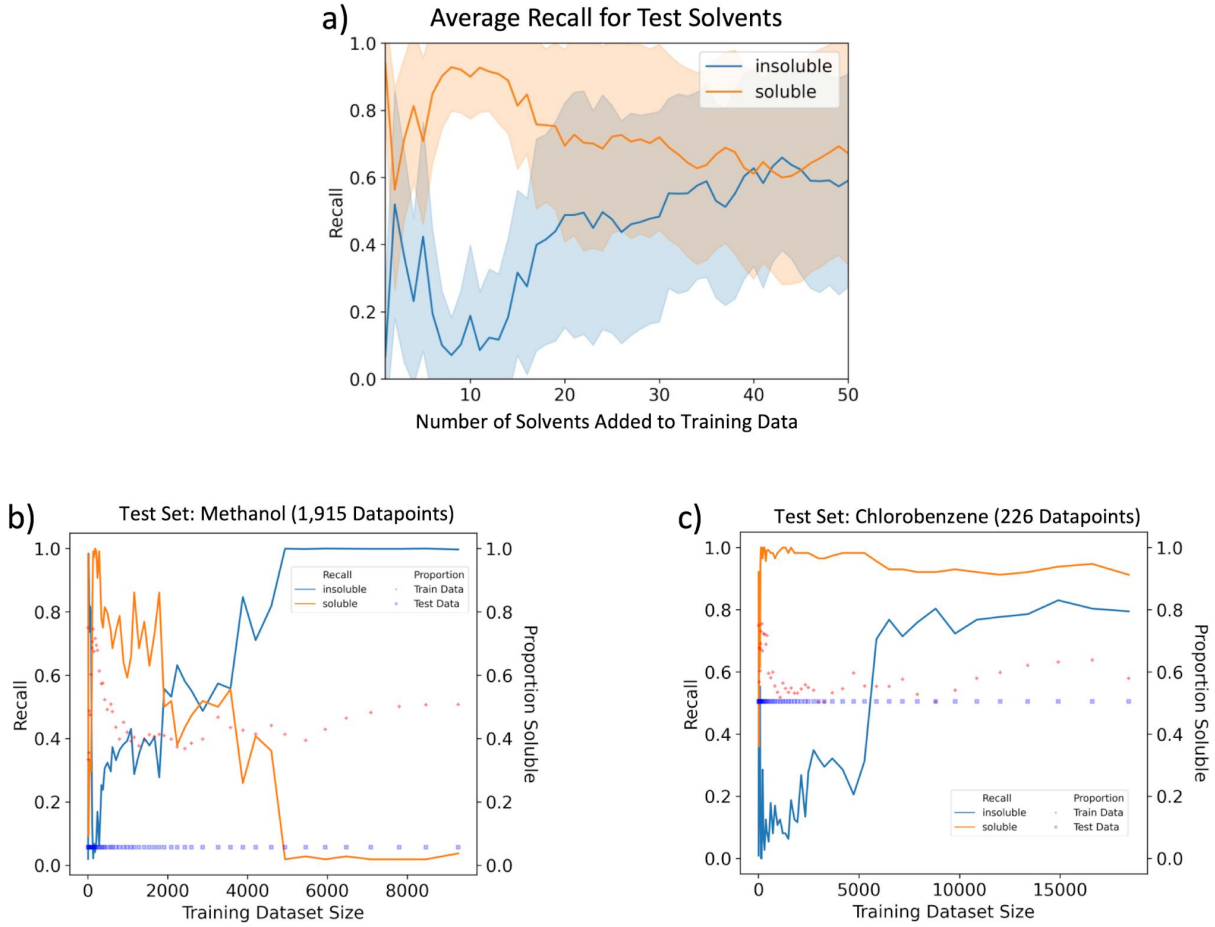

Figure S5: Random forest classifier learning curves for test solvents with all solvents and polymers held out of the training data. In (a), the average ± standard deviation recall for soluble and insoluble predictions are shown for all 51 test solvents as a function of the number of solvents in the training data. In (b) and (c), the soluble and insoluble recall for predictions on methanol (b) and chlorobenzene (c) is plotted as a function of training data set size for the prediction models. The proportion of training and test that is soluble is also plotted as a function of data set size, with each point corresponding to when another solvent was added.

In Figure S5a, there was a high recall for soluble combinations initially because the training data was overwhelmingly soluble. As more solvents were added, more insoluble data was in the model training data, and insoluble recall improved while soluble recall degraded as the model became less over-fit. On average, the model achieved a predictive recall of 0.67 ± 0.33 for soluble pairings and 0.59 ± 0.32 for insoluble pairings after all 50 training

solvents were added. This large variability can be explained by Figure S5b and Figure S5c in conjunction with Figure S6a and Figure S6b.

In Figure S6, the number of test polymer pairings with methanol (Figure S6a) or chlorobenzene (Figure S6b) that were either predicted as soluble correctly (top left), predicted as soluble incorrectly (top right), predicted as insoluble correctly (bottom left), or predicted as insoluble incorrectly (bottom right), were plotted as a function of the proportion of similar training polymers paired with water (Figure S6a) or benzene (Figure S6b) that were of the same true class as the test pairing. To compare polymer similarity, the Tanimoto similarity score (a mathematical measure of fingerprint similarity) was used (Equation S1). Training polymers were considered similar to test polymers if this value was greater than 0.75. If no similar training polymers existed, the test polymer count was plotted as blue at the 0 to 0.1 position on the x-axis. Else, it was plotted as orange at the appropriate x-axis position.

For some test solvents, as additional solvents were added to the training data, the performance for one class improved, but the performance of the other class got worse. This occurred for 20 test solvents, including methanol (shown as an example in Figure S5b). Initially, soluble recall was very high due to the soluble heavy training data. As more data was added, insoluble recall improved, while soluble recall degraded. Dramatic shifts in performance were due to the addition of similar training solvents that had training polymers similar to the test polymers. In methanol's case, the solvent with the most dramatic effect (dropping soluble recall close to zero and raising insoluble recall to one) was water. Reviewing Figure 1 of the manuscript, we see that, water was relatively close to the alcohols (which include methanol) in our 2d projection, indicating the solvents are fairly similar. Reviewing Figure S6a, we see that a significant number of the training polymers paired with water were similar to test polymers paired with methanol, however, for insoluble pairings, they were the same class as the similar test pairing, whereas for soluble pairings, they were the opposite. This could explain why the insoluble recall jumped for methanol whereas the soluble recall dropped. The insoluble methanol-polymer pairings could be learned from the

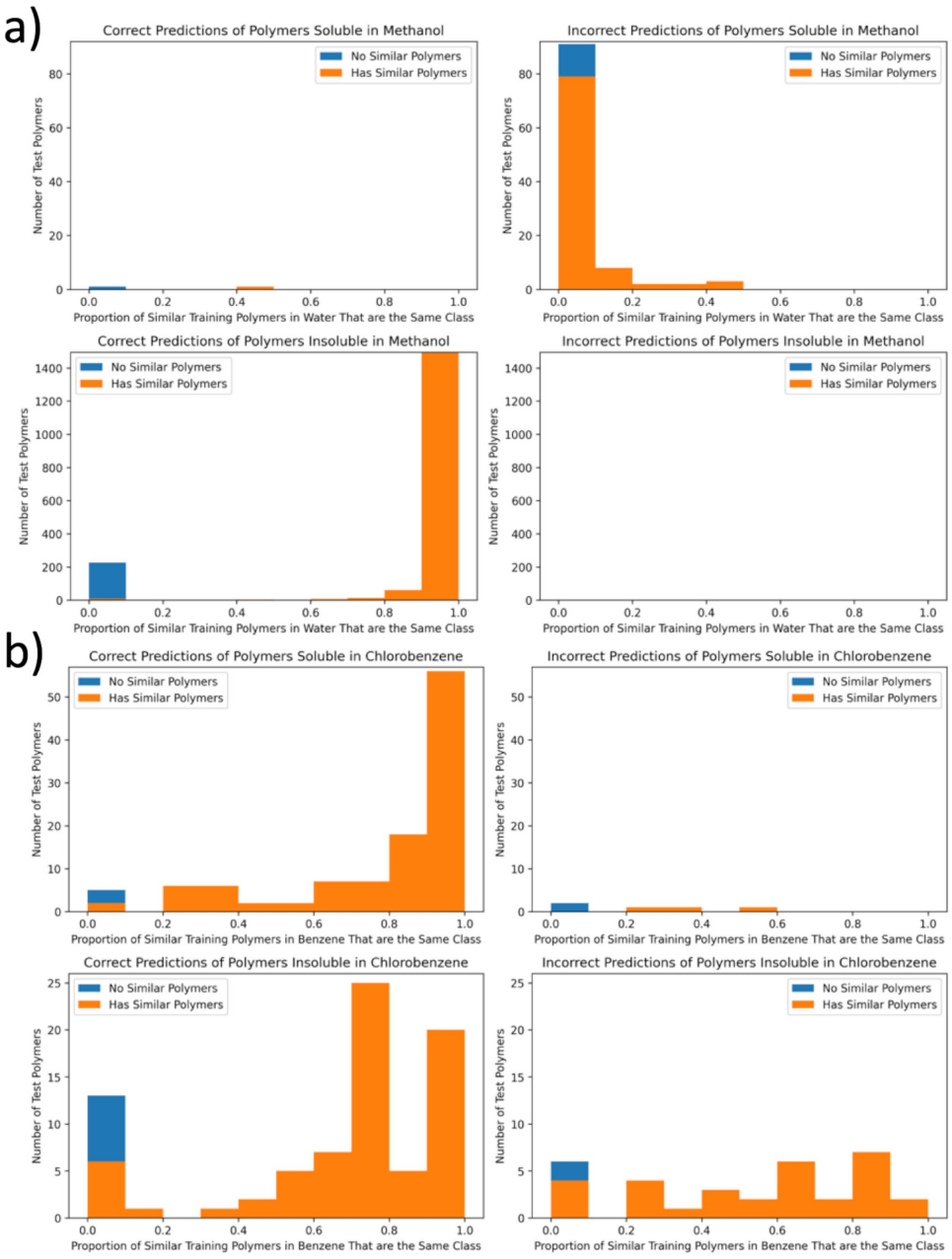

Figure S6: Number of test polymer pairings with methanol (a) or chlorobenzene (b) that were either predicted as soluble correctly (top left), soluble incorrectly (top right), insoluble correctly (bottom left), or insoluble incorrectly (bottom right), as a function of the proportion of similar training polymers paired with water (a) or benzene (b) that were the same true class as the similar test polymer-solvent pairing. Training polymers were similar to the test polymers if their Tanimoto similarity score was greater than 0.75. If a test polymer had no similar training polymers paired with water (a) or benzene (b) they were plotted in blue.

insoluble water-polymer pairings. In contrast, the model could not determine the soluble methanol-polymer pairs because the similar polymers paired with water were insoluble. This might indicate additional similar polymers that are soluble in water need to be added to the data set, or it could indicate the classification is mislabeled for these specific test polymers. Water didn't necessarily cause this issue for the other 19 test solvents where this occurred though, it may have been a different solvent. For instance, for Dimethylformamide it was Dimethylacetamide that lowered insoluble recall and raised soluble recall.

For chlorobenzene (Figure S5c), and 15 other solvents that achieve high recall for both soluble and insoluble pairings, something different occurs. After certain solvents are added, the insoluble recall improves, but the soluble recall also remains high. For chlorobenzene, the solvent that causes the most dramatic increase in insoluble recall was benzene. Looking at Figure 1, we see all of the aromatic compounds are extremely close to one another in the 2d projection, implying benzene and chlorobenzene are chemically similar solvents. Looking at Figure S6b, we find that benzene had many polymers similar to chlorobenzene's polymers, and these similar polymers have the same solubility when paired with benzene as the test polymers had when paired with chlorobenzene. As such, the model was able to learn the soluble and insoluble pairings of the test polymer-solvent combinations.

For the other solvents, both soluble and insoluble recall remained relatively constant, regardless of the number of solvents added. This could be because the new solvent-polymer combinations added are not similar to the test polymer-solvent combinations.