Electronic Supplementary Information

Predicting homopolymer and copolymer solubility through machine learning

Christopher D. Stubbs[a], Yeonjoon Kim[b], Ethan C. Quinn[a], Raúl Pérez-Soto[a], Eugene Y.-X. Chen[a*], and Seonah Kim[a*]

[a] *Department of Chemistry, Colorado State University, CO 80523, United States*

[b] *Department of Chemistry, Pukyong National University, Busan 48513, Republic of Korea*

[*] Corresponding Authors.

1. Computational

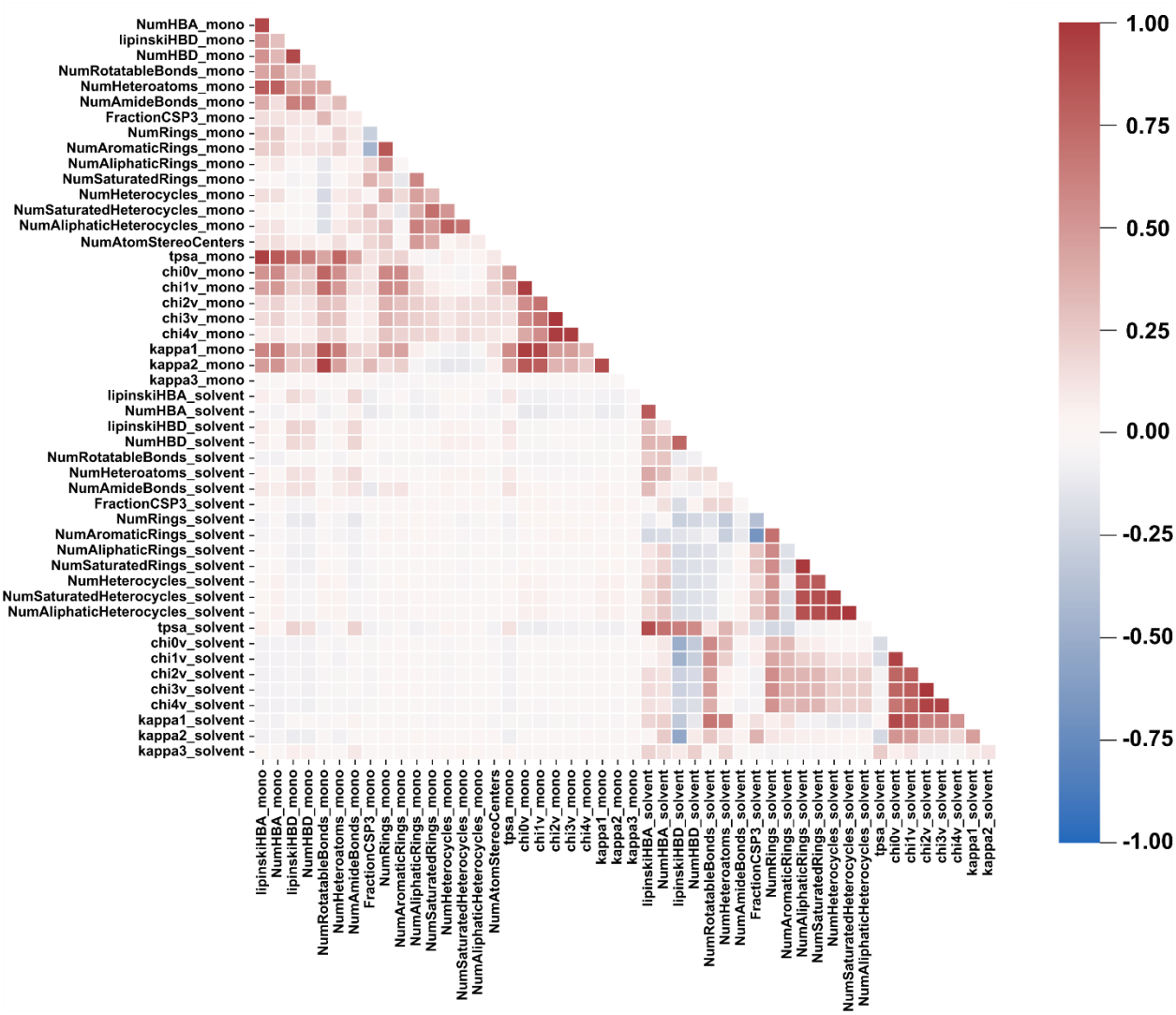

Figure S1: *Pairwise Pearson correlations for the 49 RDKit descriptors used in classical models of both homopolymer and copolymer solubility.*

Descriptors Used

Listed below are the RDKit descriptors used for descriptor-based RDKit models. For each datapoint, both monomer and solvent descriptors were calculated to yield 25 monomer + 24 solvent = 49 RDKit descriptors per datapoint. For solvents, the NumAtomStereoCenters descriptor was excluded from the solvent descriptor set as this descriptor had a value of 0 for almost every solvent in the dataset, meaning it would provide little predictive value.

Descriptors for the Mordred model will be made available online at the GitHub repository for this work. Fingerprint-type descriptors are already described in the Main Text (see *Methods*) and can be reproduced using their generation algorithm (Morgan/RDKit), bit length (32,768), and radius (3).

- lipinskiHBA
- NumHBA
- lipinskiHBD
- NumHBD
- NumRotatableBonds
- NumHeteroatoms
- NumAmideBonds
- FractionCSP3
- NumRings
- NumAromaticRings
- NumAliphaticRings
- NumSaturatedRings
- NumHeterocycles
- NumSaturatedHeterocycles
- NumAliphaticHeterocycles
- NumAtomStereoCenters (*monomers only*)
- tpsa
- chi0v
- chi1v
- chi2v
- chi3v
- chi4v
- kappa1
- kappa2
- kappa3

Table S1: *List of Mordred descriptor categories and their dimension, alongside a brief description.*

Name	Code Function	Dimension	Description
ABCIndex	ABC	2D	atom-bond connectivity index
AcidBase	nAcid	2D	acidic group count
AdjacencyMatrix	SpAbs_A	2D	SpAbs of adjacency matrix
Aromatic	nAromAtom	2D	aromatic atoms count
AtomCount	nAtom	2D	number of all atoms
Autocorrelation	ATS0dv	2D	moreau-broto autocorrelation of lag 0 weighted by valence electrons
BCUT	BCUTc-1h	2D	first highest eigenvalue of Burden matrix weighted by gasteiger charge
BalabanJ	BalabanJ	2D	Balaban's J index
BaryszMatrix	SpAbs_DzZ	2D	graph energy from Barysz matrix weighted by atomic number
BertzCT	BertzCT	2D	Bertz CT
BondCount	nBonds	2D	number of all bonds in non-kekulized structure
CPSA	PNSA1	3D	partial negative surface area (version 1)

CarbonTypes	C1SP1	2D	SP carbon bound to 1 other carbon
Chi	Xch-3d	2D	3-ordered Chi chain weighted by sigma electrons
Constitutional	SZ	2D	sum of constitutional weighted by atomic number
DetourMatrix	SpAbs_Dt	2D	graph energy from detour matrix
DistanceMatrix	SpAbs_D	2D	graph energy from distance matrix
EState	NsLi	2D	number of sLi
EccentricConnectivityIndex	ECIndex	2D	eccentric connectivity index
ExtendedTopochemicalAtom	ETA_alpha	2D	ETA core count
FragmentComplexity	fragCpx	2D	fragment complexity
Framework	fMF	2D	molecular framework ratio
GeometricalIndex	GeomDiameter	3D	geometric diameter
GravitationalIndex	GRAV	3D	heavy atom gravitational index
HydrogenBond	nHBAcc	2D	number of hydrogen bond acceptor
InformationContent	IC0	2D	0-ordered neighborhood information content
KappaShapeIndex	Kier1	2D	kappa shape index 1
Lipinski	Lipinski	2D	Lipinski rule of five
LogS	FilterItLogS	2D	Filter-it LogS
McGowanVolume	VMcGowan	2D	McGowan volume
MoRSE	Mor01	3D	3D-MoRSE (distance = 1)
MoeType	LabuteASA	2D	Labute's Approximate Surface Area
MolecularDistanceEdge	MDEC-11	2D	molecular distance edge between primary C and primary C
MolecularId	MID	2D	molecular ID
MomentOfInertia	MOMI-X	3D	moment of inertia (axis = X)
PBF	PBF	3D	PBF
PathCount	MPC2	2D	2-ordered path count
Polarizability	apol	2D	atomic polarizability
RingCount	nRing	2D	ring count
RotatableBond	nRot	2D	rotatable bonds count
SLogP	SLogP	2D	Wildman-Crippen LogP
TopoPSA	TopoPSA(NO)	2D	topological polar surface area (use only nitrogen and oxygen)
TopologicalCharge	GGI1	2D	1-ordered raw topological charge
TopologicalIndex	Diameter	2D	topological diameter
VdwVolumeABC	Vabc	2D	ABC Van der Waals volume
VertexAdjacencyInformation	VAdjMat	2D	vertex adjacency information
WalkCount	MWC01	2D	walk count (leg-1)
Weight	MW	2D	exact molecular weight
WienerIndex	WPath	2D	Wiener index
ZagrebIndex	Zagreb1	2D	Zagreb index (version 1)

Database Solvent Distribution

In **Tables S2-S4** below, we describe the most common solvents for each database alongside detailed statistics on the average number of datapoints per solvent. All values were calculated using solvent SMILES rather than name, as we found that this was a more consistent and reliable counting method due to solvent name collision (more than one name mapping to one SMILES).

Table S2: *Value counts for the ten most common solvents in the homopolymer database.*

Solvent	Count
chloroform	156
methanol	143
tetrahydrofuran	120
benzene	112
ethanol	95
acetone	85
N,N-dimethylformamide	80
water	74
toluene	69
diethyl ether	54

Table S3: *Value counts for the ten most common solvents in the copolymer database.*

Solvent	Count
chloroform	31
tetrahydrofuran	28
methanol	23
toluene	16
N,N-dimethylformamide	15
benzene	14
methyl ethyl ketone	13
acetone	13
methylene chloride	12
hexane	12

Table S4: *Database solvent statistics for the homopolymer and copolymer databases.*

Model	Num. Unique	Mean	Std. Dev.	Min. Count	25%	50%	75%	Max Count
Homopolymer	175	10.4	24.4	1	1	2	5	156
Copolymer	43	6.3	7.5	1	1	2	9	31

Definition of Binary Solubility

For our in-house experimental validation, we define the threshold between soluble and insoluble as >10% of the pre-dissolution mass present in filtrate, or greater than 25 mg dissolved in 10 mL of solvent. For these experiments, the temperature was 23 ºC and the filter pore size was 2.5 µm. For our homopolymer and copolymer database, we use the values reported in the literature and discard any datapoints with uncertain solubility (e.g. partial solubility, swelling noted, elevated temperature, etc.) We therefore believe that the solubility values in our database are reasonable and consistent despite being sourced from a wide variety of literature. Due to the broad scope of literature and chemical space covered, it is difficult to guarantee that the boundary between soluble and insoluble is the same for all data sources. However, to address this issue we validate our model through k-fold cross-validation – which should show significant differences in performance across folds if the boundary between soluble/insoluble is changing significantly across different portions of the database. We do not see large fluctuations in our accuracy and R2 values across these folds, and thus we indirectly guarantee the quality of the data and its consistency.

Copolymer Model Input

All copolymer models used a digital representation of copolymer stoichiometry and sequence to encode this information numerically. In **Figure S2** below, we describe the transformation of the stoichiometry/sequence into numeric form for an example alternating copolymer with a comonomer ratio of 0.6:0.4 (3:2).

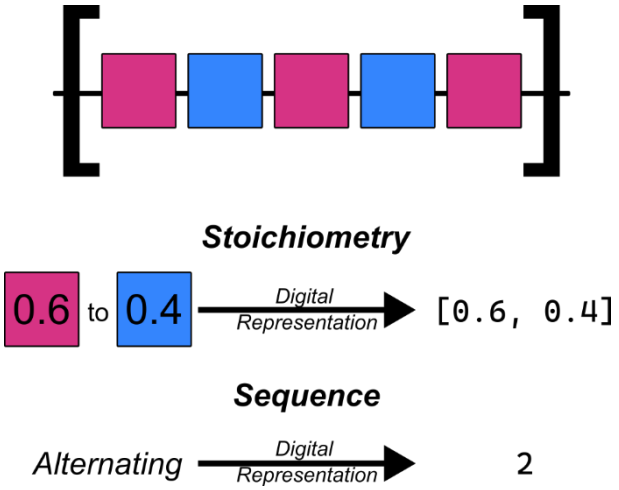

Figure S2: *Transformation of copolymer stoichiometry and sequence to digital representations used for model training.*

Hyperparameter Details

All classical homopolymer and copolymer models were trained using the hyperparameters listed below in **Table S5**. For computational efficiency, hyperparameter tuning was only performed on the most successful model (RDKit with Random Forest) and most diverse database (homopolymer), with the chosen hyperparameters in **bold with underline**.

Table S5: *Hyperparameters used for all classical models of polymer solubility.*

Architecture	SciKit-Learn Function	Hyperparameters Used
AdaBoost	AdaBoostClassifier	n_estimators=50 learning_rate=1.0 algorithm='SAMME.R'
Decision Tree	DecisionTreeClassifier	criterion='gini' splitter='best' max_depth=None min_samples_split=2 min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features=None random_state=None max_leaf_nodes=None min_impurity_decrease=0.0 class_weight=None ccp_alpha=0.0 monotonic_cst=None
Random Forest	RandomForestClassifier	n_estimators=[10,**100**] criterion='gini' max_depth=[**None**, 5, 4, 3, 2] min_samples_split=[**2**, 4] min_samples_leaf=1 min_weight_fraction_leaf=0.0 max_features='sqrt' max_leaf_nodes=None min_impurity_decrease=0.0 bootstrap=True oob_score=False n_jobs=None random_state=None verbose=0 warm_start=False class_weight=None ccp_alpha=0.0 max_samples=None monotonic_cst=None
Naive Bayes	GaussianNB	priors=None var_smoothing=1e-09

Classical Model Performance

Table S6*: Full model statistics for the Classical Homopolymer models of polymer solubility, sorted by descending accuracy. All values are rounded to 4 decimals. Archi. = model architecture, Acc. = model accuracy, Bal. Acc. = class-balanced model accuracy, Prec. = precision. For average K-Fold accuracies, accuracy values were averaged across 5 folds.*

Model Label	Descriptor	Archi.	Acc. (test)	Acc. (Avg., K-Fold)	Bal. Acc. (test)	Prec.	Recall	F1 Score	Jaccard Ind.	ROC-AUC
atom_bd_RF	RDKit	Random Forest	0.8462	0.8232	0.8017	0.8418	0.9431	0.8896	0.8011	0.8017
atom_mordred_RF	RDKit + Mordred	Random Forest	0.8396	0.843	0.7921	0.8343	0.9431	0.8854	0.7944	0.7921
mordred_RF	Mordred	Random Forest	0.8308	0.8503	0.7839	0.8304	0.9331	0.8787	0.7837	0.7839
atom_morganfp_RF	RDKit + Morgan FP	Random Forest	0.8242	0.818	0.7681	0.8156	0.9465	0.8762	0.7796	0.7681

atom_rdfp_AB	RDKit + RDKit FP	AdaBoost	0.8088	0.785	0.7656	0.8232	0.903	0.8612	0.7563	0.7656
atom_rdfp_RF	RDKit + RDKit FP	Random Forest	0.7934	0.8085	0.7432	0.806	0.903	0.8517	0.7418	0.7432
mordred_DT	Mordred	Decision Tree	0.7868	0.774	0.7581	0.8301	0.8495	0.8397	0.7236	0.7581
atom_bd_DT	RDKit	Decision Tree	0.778	0.7557	0.7437	0.8173	0.8528	0.8347	0.7163	0.7437
atom_mordred_DT	RDKit + Mordred	Decision Tree	0.778	0.7784	0.7483	0.8235	0.8428	0.8331	0.7139	0.7483
atom_rdfp_DT	RDKit + RDKit FP	Decision Tree	0.7714	0.7909	0.7249	0.7982	0.8729	0.8339	0.7151	0.7249
mfp_DT	Morgan FP	Decision Tree	0.7692	0.7689	0.7370	0.8149	0.8395	0.827	0.7051	0.737
atom_bd_AB	RDKit	AdaBoost	0.7692	0.7857	0.7186	0.7922	0.8796	0.8336	0.7147	0.7186
rdfp_AB	RDKit FP	AdaBoost	0.7648	0.7549	0.6984	0.7727	0.9097	0.8356	0.7177	0.6984
atom_morganfp_AB	RDKit + Morgan FP	AdaBoost	0.7626	0.7931	0.7136	0.7903	0.8696	0.828	0.7065	0.7136
atom_morganfp_DT	RDKit + Morgan FP	Decision Tree	0.7626	0.7924	0.7213	0.7994	0.8528	0.8252	0.7025	0.7213
mordred_AB	Mordred	AdaBoost	0.7582	0.7857	0.7057	0.7838	0.8729	0.8259	0.7035	0.7057
atom_mordred_AB	RDKit + Mordred	AdaBoost	0.7582	0.7887	0.7057	0.7838	0.8729	0.8259	0.7035	0.7057
mfp_RF	Morgan FP	Random Forest	0.7473	0.7608	0.6881	0.7706	0.8763	0.82	0.695	0.6881
mfp_AB	Morgan FP	AdaBoost	0.7473	0.7799	0.6835	0.7659	0.8863	0.8217	0.6974	0.6835
rdfp_DT	RDKit FP	Decision Tree	0.7429	0.7498	0.7139	0.8033	0.806	0.8047	0.6732	0.7139
atom_mordred_NB	RDKit + Mordred	Naive Bayes	0.7297	0.7094	0.6747	0.7651	0.8495	0.8051	0.6737	0.6747
mordred_NB	Mordred	Naive Bayes	0.7297	0.7124	0.6763	0.7667	0.8462	0.8045	0.6729	0.6763
rdfp_RF	RDKit FP	Random Forest	0.7209	0.763	0.6773	0.7722	0.8161	0.7935	0.6577	0.6773
atom_bd_NB	RDKit	Naive Bayes	0.6593	0.6046	0.7009	0.8673	0.5686	0.6869	0.5231	0.7009
mfp_NB	Morgan FP	Naive Bayes	0.5231	0.5004	0.5927	0.7929	0.3712	0.5057	0.3384	0.5927
atom_morganfp_NB	RDKit + Morgan FP	Naive Bayes	0.5231	0.5092	0.5819	0.7662	0.3946	0.521	0.3522	0.5819
rdfp_NB	RDKit FP	Naive Bayes	0.5033	0.4703	0.5853	0.8017	0.3244	0.4619	0.3003	0.5853
atom_rdfp_NB	RDKit + RDKit FP	Naive Bayes	0.5011	0.4762	0.5713	0.7647	0.3478	0.4782	0.3142	0.5713

Table S7: *Full model statistics for the Classical Copolymer models of polymer solubility, sorted by descending accuracy. All values are rounded to 4 decimals. Archi. = model architecture, Acc. = model accuracy, Bal. Acc. = class-balanced model accuracy, Prec. = precision. For average K-Fold accuracies, accuracy values were averaged across 5 folds.*

Model Label	Descriptor	Archi.	Acc. (test)	Acc. (avg., K-Fold)	Bal. Acc. (test)	Prec.	Recall	F1 Score	Jaccard Ind.	ROC-AUC
atom_mordred_RF	RDKit + Mordred	Random Forest	0.9412	0.931	0.9412	0.9796	0.9412	0.96	0.9231	0.9412
atom_morganfp_RF	RDKit + Morgan FP	Random Forest	0.9265	0.8713	0.8725	0.9259	0.9804	0.9524	0.9091	0.8725
atom_bd_RF	RDKit	Random Forest	0.9265	0.9212	0.8922	0.9423	0.9608	0.9515	0.9074	0.8922
mordred_RF	Mordred	Random Forest	0.9118	0.931	0.8824	0.9412	0.9412	0.9412	0.8889	0.8824
mfp_RF	Morgan FP	Random Forest	0.8971	0.8073	0.8137	0.8929	0.9804	0.9346	0.8772	0.8137
atom_rdfp_RF	RDKit + RDKit FP	Random Forest	0.8676	0.8172	0.7745	0.875	0.9608	0.9159	0.8448	0.7745
rdfp_RF	RDKit FP	Random Forest	0.7941	0.768	0.7059	0.8491	0.8824	0.8654	0.7627	0.7059

Table S8: *5-fold cross-validation accuracies and confusion matrices for Homopolymer Classical models. Each number in the column "K-Fold Accuracies" represents the accuracy for a given fold of the model training set during cross-validation. The diagonal elements of the confusion matrix represent true positives (top left) and true negatives (bottom right), whereas the off-diagonals elements represent false negatives (top right) and false positives (bottom left).*

Model Label	Descriptor	Archi.	K-Fold Accuracies	Confusion Matrix
at_NOPE_nr_atom_bd_RF	RDKit	Random Forest	[0.8242 0.8755 0.7875 0.8272 0.8015]	[[103 53] [17 282]]
at_NOPE_nr_atom_mordred_RF	RDKit + Mordred	Random Forest	[0.8645 0.8864 0.8095 0.8346 0.8199]	[[100 56] [17 282]]
at_NOPE_nr_mordred_RF	Mordred	Random Forest	[0.8681 0.8901 0.8205 0.8456 0.8272]	[[99 57] [20 279]]
at_NOPE_nr_atom_morganfp_RF	RDKit + Morgan FP	Random Forest	[0.8388 0.8535 0.7839 0.8088 0.8051]	[[92 64] [16 283]]
at_NOPE_nr_atom_rdfp_AB	RDKit + RDKit FP	AdaBoost	[0.8095 0.8059 0.7949 0.7574 0.7574]	[[98 58] [29 270]]
at_NOPE_nr_atom_rdfp_RF	RDKit + RDKit FP	Random Forest	[0.8498 0.8242 0.7839 0.7868 0.7978]	[[91 65] [29 270]]
at_NOPE_nr_mordred_DT	Mordred	Decision Tree	[0.7802 0.8095 0.7546 0.7904 0.7353]	[[104 52] [45 254]]
at_NOPE_nr_atom_bd_DT	RDKit	Decision Tree	[0.7473 0.7802 0.7399 0.7574 0.7537]	[[99 57] [44 255]]
at_NOPE_nr_atom_mordred_DT	RDKit + Mordred	Decision Tree	[0.7802 0.8022 0.7509 0.7941 0.7647]	[[102 54] [47 252]]
at_NOPE_nr_atom_rdfp_DT	RDKit + RDKit FP	Decision Tree	[0.7949 0.8498 0.7839 0.761 0.7647]	[[90 66] [38 261]]
at_NOPE_nr_mfp_DT	Morgan FP	Decision Tree	[0.7766 0.7729 0.7509 0.7794 0.7647]	[[99 57] [48 251]]

at_NOPE_nr_atom_bd_AB	RDKit	AdaBoost	[0.7802 0.8388 0.7619 0.7794 0.7684]	[[87 69] [36 263]]
at_NOPE_nr_rdfp_AB	RDKit FP	AdaBoost	[0.7729 0.7509 0.7619 0.7353 0.7537]	[[76 80] [27 272]]
at_NOPE_nr_atom_morganfp_AB	RDKit + Morgan FP	AdaBoost	[0.8095 0.8242 0.7729 0.7684 0.7904]	[[87 69] [39 260]]
at_NOPE_nr_atom_morganfp_DT	RDKit + Morgan FP	Decision Tree	[0.7949 0.8132 0.7656 0.8051 0.7831]	[[92 64] [44 255]]
at_NOPE_nr_mordred_AB	Mordred	AdaBoost	[0.7949 0.8168 0.7949 0.7426 0.7794]	[[84 72] [38 261]]
at_NOPE_nr_atom_mordred_AB	RDKit + Mordred	AdaBoost	[0.7985 0.8168 0.7912 0.7426 0.7941]	[[84 72] [38 261]]
at_NOPE_nr_mfp_RF	Morgan FP	Random Forest	[0.7546 0.7766 0.7289 0.7647 0.7794]	[[78 78] [37 262]]
at_NOPE_nr_mfp_AB	Morgan FP	AdaBoost	[0.7692 0.8022 0.7656 0.7684 0.7941]	[[75 81] [34 265]]
at_NOPE_nr_rdfp_DT	RDKit FP	Decision Tree	[0.7436 0.7399 0.7436 0.7684 0.7537]	[[97 59] [58 241]]
at_NOPE_nr_atom_mordred_NB	RDKit + Mordred	Naive Bayes	[0.7289 0.7436 0.685 0.7096 0.6801]	[[78 78] [45 254]]
at_NOPE_nr_mordred_NB	Mordred	Naive Bayes	[0.7289 0.7509 0.6886 0.7132 0.6801]	[[79 77] [46 253]]
at_NOPE_nr_rdfp_RF	RDKit FP	Random Forest	[0.7766 0.7692 0.7106 0.761 0.7978]	[[84 72] [55 244]]
at_NOPE_nr_atom_bd_NB	RDKit	Naive Bayes	[0.5128 0.619 0.6154 0.6324 0.6434]	[[130 26] [129 170]]
at_NOPE_nr_mfp_NB	Morgan FP	Naive Bayes	[0.5165 0.4982 0.4469 0.511 0.5294]	[[127 29] [188 111]]
at_NOPE_nr_atom_morganfp_NB	RDKit + Morgan FP	Naive Bayes	[0.5165 0.5128 0.4505 0.5147 0.5515]	[[120 36] [181 118]]
at_NOPE_nr_rdfp_NB	RDKit FP	Naive Bayes	[0.5092 0.4652 0.4322 0.4816 0.4632]	[[132 24] [202 97]]
at_NOPE_nr_atom_rdfp_NB	RDKit + RDKit FP	Naive Bayes	[0.5092 0.4762 0.4432 0.4816 0.4706]	[[124 32] [195 104]]

Table S9: *5-fold cross-validation accuracies and confusion matrices for Copolymer Classical models. Each number in the column "K-Fold Accuracies" represents the accuracy for a given fold of the model training set during cross-validation. The diagonal elements of the confusion matrix represent true positives (top left) and true negatives (bottom right), whereas the off-diagonals elements represent false negatives (top right) and false positives (bottom left).*

Model Label	Descriptor	Archi.	K-Fold Accuracies	Confusion Matrix
co_di_atom_mordred_RF	RDKit + Mordred	Random Forest	[0.878 0.9268 0.925 0.975 0.95]	[[16 1] [3 48]]
co_di_atom_morganfp_RF	RDKit + Morgan FP	Random Forest	[0.8293 0.9024 0.8 0.9 0.925]	[[13 4] [1 50]]
co_di_atom_bd_RF	RDKit	Random Forest	[0.8293 0.9268 0.925 0.975 0.95]	[[14 3] [2 49]]
co_di_mordred_RF	Mordred	Random Forest	[0.878 0.9268 0.925 0.975 0.95]	[[14 3] [3 48]]
co_di_mfp_RF	Morgan FP	Random Forest	[0.7317 0.8049 0.775 0.85 0.875]	[[11 6] [1 50]]

co_di_atom_rdfp_RF	RDKit + RDKit FP	Random Forest	[0.7561 0.8049 0.725 0.875 0.925]	[[10 7] [2 49]]
co_di_rdfp_RF	RDKit FP	Random Forest	[0.7073 0.6829 0.675 0.875 0.9]	[[9 8] [6 45]]

Small Molecule Random Forest Performance

In order to predict additive solubility, a random forest model was trained on a pre-existing database of small molecule solubility with a train/test split ratio of 75/25. This model was trained to predict $\Delta G_{solvation}$ for small molecules, but was used to predict binary solubility (soluble/insoluble) for additives by mapping negative $\Delta G_{solvation}$ values to soluble and positive $\Delta G_{solvation}$ values to insoluble. For the additives studied, this appeared successful as only a few cases resulted in misassignment of the soluble label to insoluble datapoints (see **Table 2** in the main text). This occurred because the model incorrectly predicted a negative $\Delta G_{solvation}$. As stated in the main text (Methods – Additive Removal) we believe that these prediction errors are not reflective of our model's predictive ability for soluble compounds, but rather are due to the structures of these molecules (azodicarbonamide, melamine, and decabromodiphenyl ether). To provide further evidence of our model's performance on positive $\Delta G_{solvation}$ values, in **Table S10** we provide performance metrics for the train/test splits and any positive $\Delta G_{solvation}$ values within those splits. Additionally, in **Figure S3** we show the kernel density estimate (KDE) of the predicted $\Delta G_{solvation}$ values versus the actual $\Delta G_{solvation}$ values for both train and test. We find that for positive $\Delta G_{solvation}$ values the small molecule RF model achieves a low MAE and high R^2 in both train and test, and that the predicted distribution of $\Delta G_{solvation}$ values is similar but not identical to the input data. For these reasons, we believe that our small molecule RF model is able to effectively predict $\Delta G_{solvation}$ values.

Table S10: *Performance metrics for the small molecule random forest model used in additive solubility predictions.*

Metric	Mean Absolute Error (MAE) (kcal/mol)	Mean Squared Error (MSE) (kcal/mol)	R2
All Train	0.14	0.08	1.00
All Test	0.36	0.52	0.97
Positive ΔG (Train)	0.18	0.11	0.95
Positive ΔG (Test)	0.13	0.04	0.93

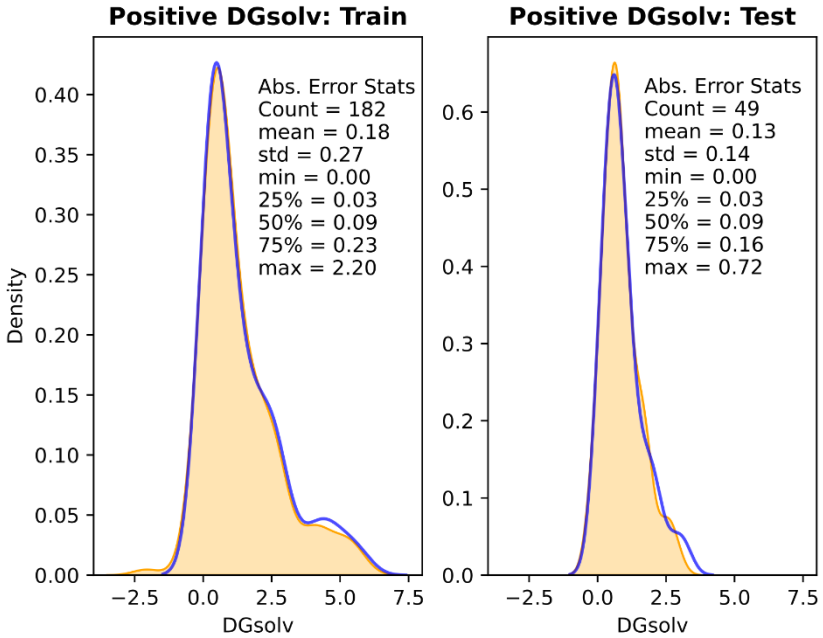

Figure S3: *Kernel density estimate (KDE) plot for the predicted (orange) vs actual (blue) $\Delta G_{solvation}$ values in the small molecule dataset used to train the small molecule RF model. Only positive $\Delta G_{solvation}$ were considered to demonstrate our model's ability to predict $\Delta G_{solvation} > 0$. All E values are given in kcal/mol.*

Fingerprint Dimensionality Studies

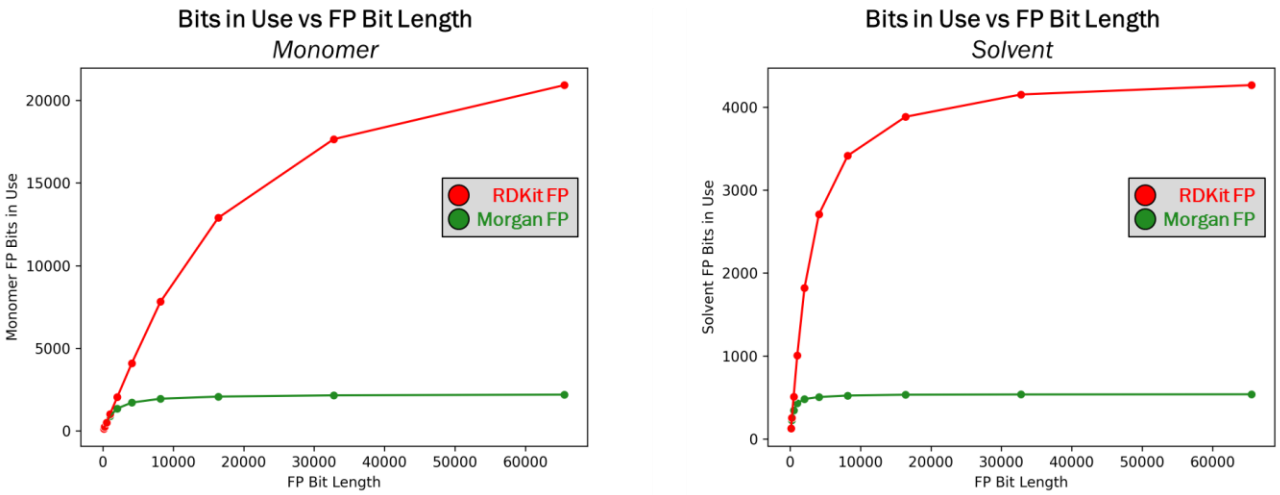

Figure S4: *Fingerprint (FP) bits in use as a function of FP bit length for monomer (left) and solvent (right). All values were calculated using our homopolymer database.*

To verify the validity of the chosen fingerprint (FP) bit length, we examined the number of bits in use by the RDKit/Morgan FPs as their bit length was increased from 128 to 65,536 (**Figure S4**). Bit lengths 128, 256, 512, 1024, 2048, 4096, 8192, 16,384, 32,768, and 65,536 were all evaluated with radius 3 on the homopolymer database due to its greater size and chemical diversity. As can be seen in **Figure S4**, for both monomer and solvent the number of unique bits in the Morgan FP plateaus quickly – implying that the fingerprint has covered all of the chemical space that it can. However, this is not true for the RDKit FP which does not appear to plateau for the monomer case and only begins to plateau at 32,768 for the solvent. We chose the bit length of 32,768 in our manuscript to maximize the amount of chemical space described by the FPs, as this bit length is sufficient to describe most monomer moieties and almost all solvent moieties. Higher bit lengths than 32,768 were not used due to the large storage space and processing costs larger fingerprints require, alongside the marginal gains in chemical space coverage for solvent seen by increasing bit length. As part of the motivation for this study was a thorough investigation into descriptor performance for polymer property prediction, we did not wish to reduce the required bit length used by omitting the RDKit FP from our calculations. Lastly, we chose to solely consider our homopolymer database for bit length determination as it was more extensive than our copolymer database, and we chose to use the same bit length for our copolymer studies for consistency and easy of comparison.

SHAP Outlier Analysis

To evaluate the reliability of our SHAP analysis, we examined datapoints which appeared as red datapoints within blue clusters or as blue datapoints within red clusters. Specifically, we examined exceptions to the observed trend for three features: ‘chi1vsolvent’, ‘kappa3solvent’, and ‘tpsa_mono’. For each feature, we identified these exceptions by examining the 10 largest or smallest SHAP values and looking for datapoints within the 25/75% quartiles of the feature values. High feature values were within the 75-100% quartile range, while low feature values were within the 0-25% quartile range. By looking for high feature values (red) in clusters of low feature values (blue) and vice versa, we were able to identify exceptions to the observed trend. From this analysis, we found that these datapoints with anomalous SHAP values did not show clear trends in monomer/solvent chemical identity or in their solubility label, but our homopolymer RF model accurately predicted the solubility of all the datapoints examined. Given that there were no clear trends seen in the exceptions studied, we cannot conclusively identify the source of these clusters, but we speculate that this is due to a combination of model and SHAP algorithm inaccuracies. As the true SHAP values are typically only estimated, it is possible that the dataset provided to the SHAP algorithm for background sampling is insufficient to provide reasonable explanations for all datapoints. Despite this possibility, the majority of datapoints appear to obey similar trends in SHAP vs feature values and so we believe our analysis remains valid.

2. Experimental

Table S11: *Overview of single homopolymer solubility results. The best model for homopolymer solubility (RF with RDKit descriptors) was used for all predictions. [SW]Datapoint removed due to swelling. *Polypropylene datapoints, which were used to demonstrate challenges with tacticity in model prediction but were not included in accuracy calculations.*

Polymer	Solvent	% Original Mass in Filtrate	Soluble	Predicted Solubility	Agreement
PLA	Cyclohexane	0.0	No	Yes	No
PLA	DCM	84.4	Yes	Yes	Yes
PLA	THF	82.6	Yes	Yes	Yes
PLA	Toluene	-	No[SW]	-	-
PMMA	Cyclohexane	0.0	No	Yes	No
PMMA	DCM	85.0	Yes	Yes	Yes
PMMA	THF	69.7	Yes	Yes	Yes
PMMA	Toluene	-	No[SW]	-	-
PS	Cyclohexane	33.2	Yes	Yes	Yes
PS	DCM	82.7	Yes	Yes	Yes
PS	THF	75.9	Yes	Yes	Yes
PS	Toluene	91.1	Yes	Yes	Yes
PP	Toluene	8.0	No*	Yes*	No*
PP	Cyclohexane	-	No*[SW]	-	-
PP	DCM	8.0	No*	Yes*	No*
PP	THF	7.5	No*	Yes*	No*
PAA	Cyclohexane	0.0	No	No	Yes
PAA	DCM	1.2	No	No	Yes
PAA	THF	2.7	No	No	Yes
PAA	Toluene	0.4	No	No	Yes