Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2022

Supplementary Information

CoPriNet: Graph Neural Networks provide accurate and rapid compound price prediction for molecule prioritisation.

Ruben Sanchez-Garcia, Dávid Havasi, Gergely Takács, Matthew C. Robinson, Alpha Lee, Frank von Delft, Charlotte M. Deane

Contents

1. Retrosynthesis-based score for PC and NPNP datasets

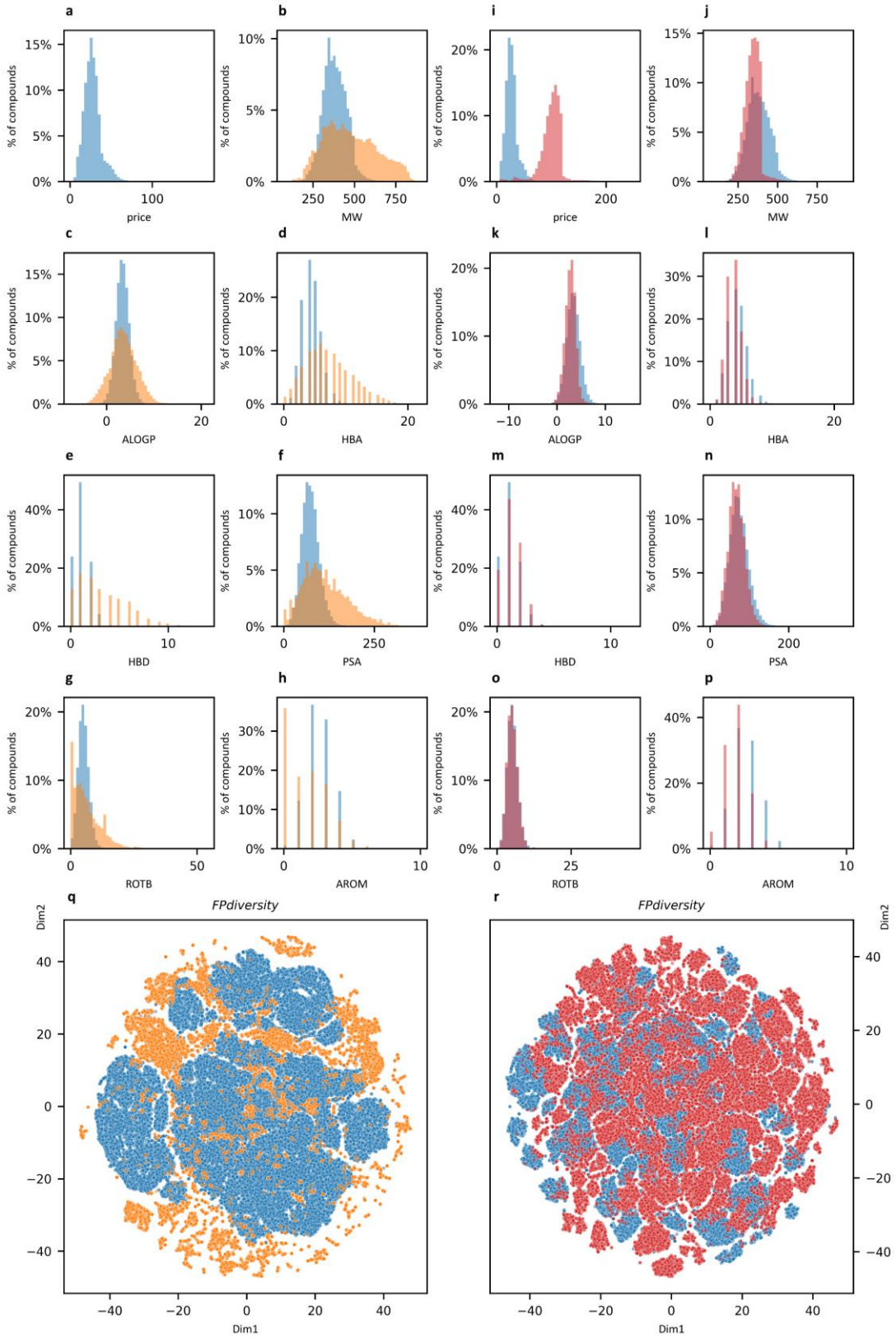
Supplementary Table 1. Statistics for different scores computed on the PC and NPNP datasets using as ground truth the class of the compound (PC vs NPNP) or the retrosynthesis-based score (ManifoldSA).

	PC vs NPNP	ManifoldSA		
	MCC	PCC	SRCC	MCC
SAScore	0.86	0.80	0.79	0.73
SCScore	0.18	0.13	0.12	0.23
SYBA	0.75	0.63	0.65	0.55
RAscore	0.78	0.81	0.77	0.72
IsolationForest (Supplementary Section 10)	0.77	0.66	0.74	0.63
CoPriNet	0.82	0.63	0.66	0.58

Note: PC: purchasable compounds; NPNP: Non-purchasable natural products; PCC: Pearson correlation coefficient; SRCC: Spearman's rank correlation coefficient; MCC: Matthew's correlation coefficient

2. Comparison of testing datasets

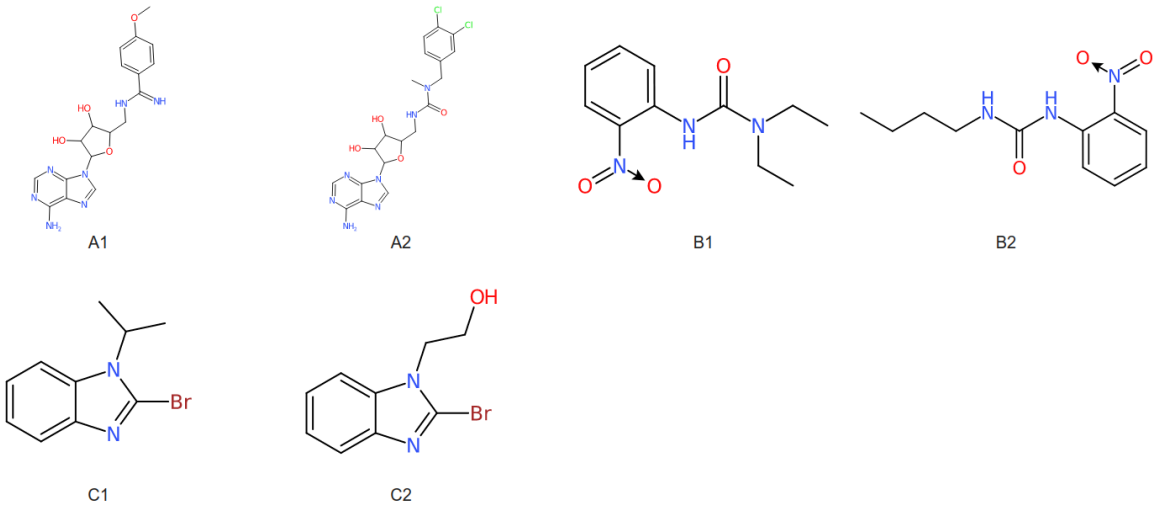
We characterized the differences between the set of PC vs NPNP compounds and PC vs Virtual Compounds by computing the distributions of several QED properties and also by comparing the regions of the chemical space that each dataset spans using Morgan fingerprints and dimensionality reduction techniques. From these calculations, it seems clear that PC compounds tend to be smaller than NPNP but with less aromatic rings and larger polar surface area values. The 2D representation of the chemical space also show non-overlapping regions, showing that the datasets are different from a fingerprint perspective. The same can be inferred from the 2D representation of the chemical space of the Virtual Compounds vs PC dataset. On the other hand, the Virtual Compounds vs PC QED descriptor distributions are similar except for the size. This result is perhaps no surprising as the main goal of commercial catalogues is to recapitulate molecules with good drug-like properties that are biased towards fixed values of these descriptors. However, Virtual Compounds exhibit one important difference with respect PC compounds: they tend to be far more expensive. This makes Virtual Compounds ideal to study the behaviour of CoPriNet when trying to predict prices that are systematically shifted with respect to the values of the training set.



Supplementary Figure 1. CoPriNet test compound dataset (PC, blue) is substantially different from the test set of Non-Purchasable Natural Products (NPNP, orange) and the Virtual compounds (red). a-p) Histograms of QED descriptors for PC and NPNP compounds (a-h) and PC and Virtual compounds (i-p). Bottom, PCA + t-SNE 2D projection of the Morgan fingerprints (radius 2) of the PC and NPNP compounds (q) and PC and Virtual compounds (r).

3. Synthetic similarity in structurally similar compounds

It is well known that structurally similar compounds may have quite different synthetic accessibilities (see Supplementary Figure 2 and Supplementary Table 2), causing complexity-based measurements such as SAscore to underperform. However, in many benchmarks, such as the one conducted in this work, SAscore exhibits a decent performance, thus indicating that this problem is not so frequent. With the aim of shedding light to this question, we counted the number of compound pairs contained in our test dataset, that while being highly similar, exhibit totally different retrosynthesis scores (one regarded as synthesizable and the other as non-synthesizable). In our testing set, only 14 molecule pairs out of the 257 pairs that exhibited Tanimoto similarity > 70%, exhibited this duality. Although this number might suggest that the high similarity-different synthesizability scenario is infrequent, we acknowledge that our dataset, a random subset of a commercial catalogue, may not be representative of many use cases. Thus, we performed an additional experiment in which we compute the same statistics for a dataset obtained computing similarity searches for 100 randomly sampled catalogue compounds. This scenario represents the typical use case in which we are interested in finding analogues of a compound of interest. In this case, and similarly to the previous experiment, we only found two examples for which some of the analogues were much more difficult to synthesize than the query molecule. Based on these two results, it seems plausible that this problem, while overly concerning, is infrequent for catalogue compounds, which could explain the perceived reliable performance of methods like SAscore and their potential risks when used with *de novo* generated molecules.



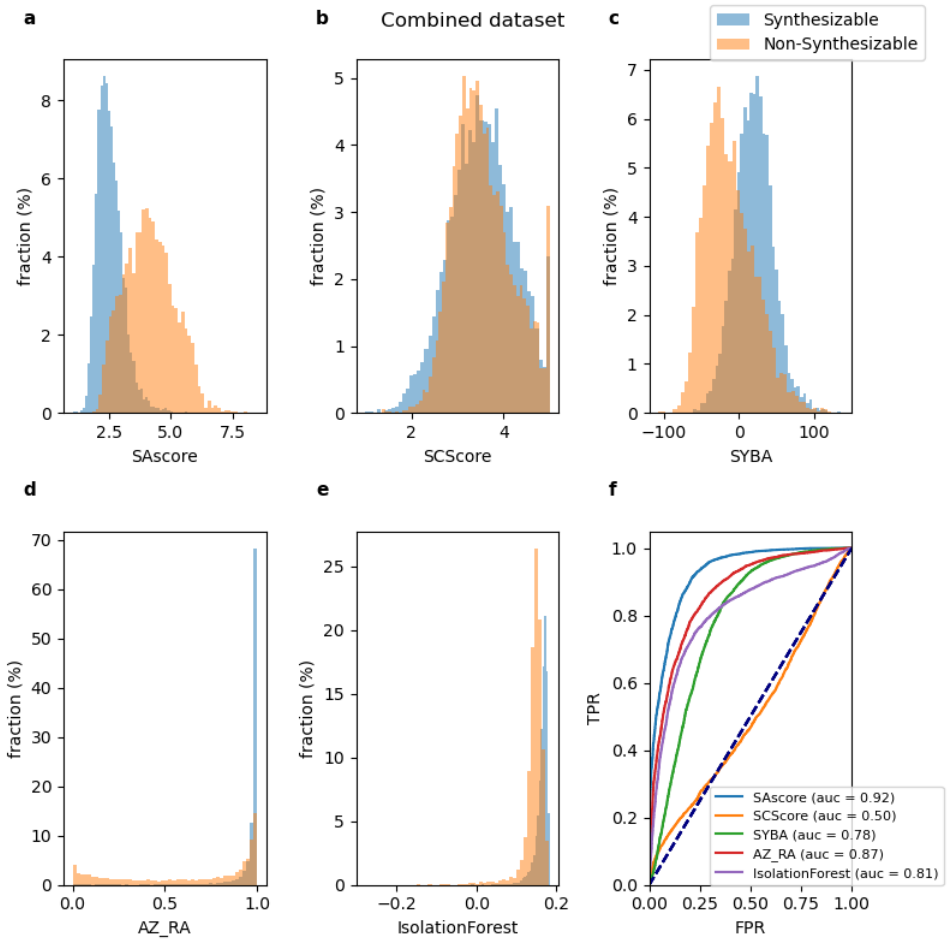
Supplementary Figure 2. Three examples of pairs of highly similar molecules with opposite synthesizability.

Supplementary Table 2. Different SA measurements for highly similar molecules with opposite synthesizability.

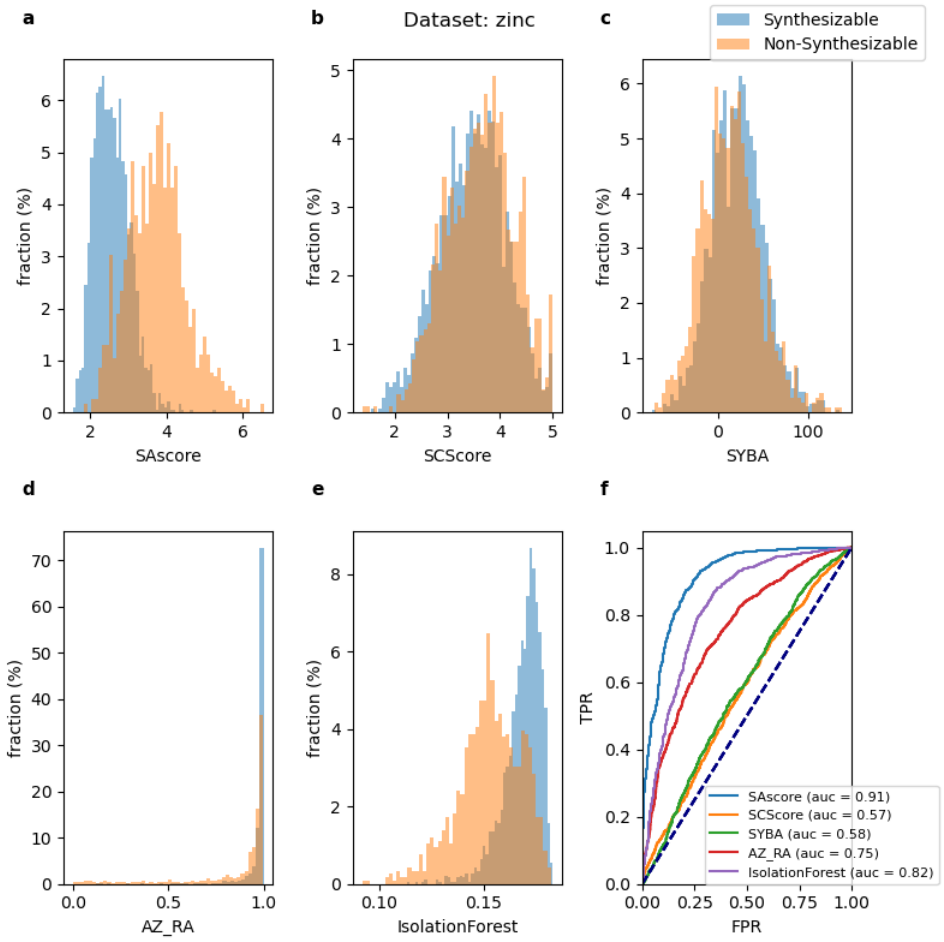
Molecule	ManifoldSA	IsolationForest	SYBA	SAscore	RAScore	SCScore	CoPriNet
A1	0.19	0.14	-7.91	3.83	0.82	3.63	4.42
A2	0.75	0.14	6.10	3.74	0.89	4.52	4.36
B1	0.00	0.16	31.61	1.91	1.00	1.88	2.84
B2	0.92	0.16	37.52	1.75	1.00	2.02	2.98
C1	0.00	0.15	1.27	2.21	0.99	2.48	3.13
C2	0.92	0.15	6.33	2.20	0.99	2.47	2.88

4. SA measurements for Gao and Coley datasets

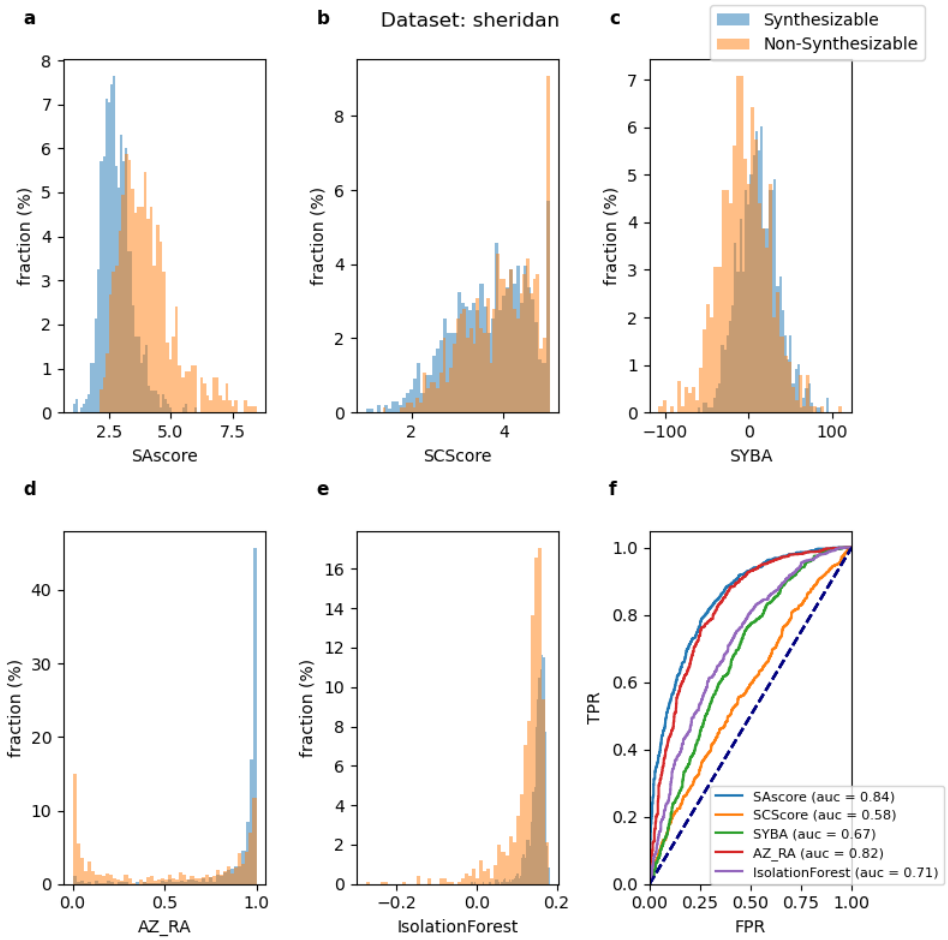
In this section we employed some of the datasets compiled by Gao and Coley (Gao and Coley 2020) and we complemented them by computing additional SA scores that were not studied in the original publication. Overall, the SAscore better reproduced retrosynthesis predictions, but performance varies depending on the dataset.



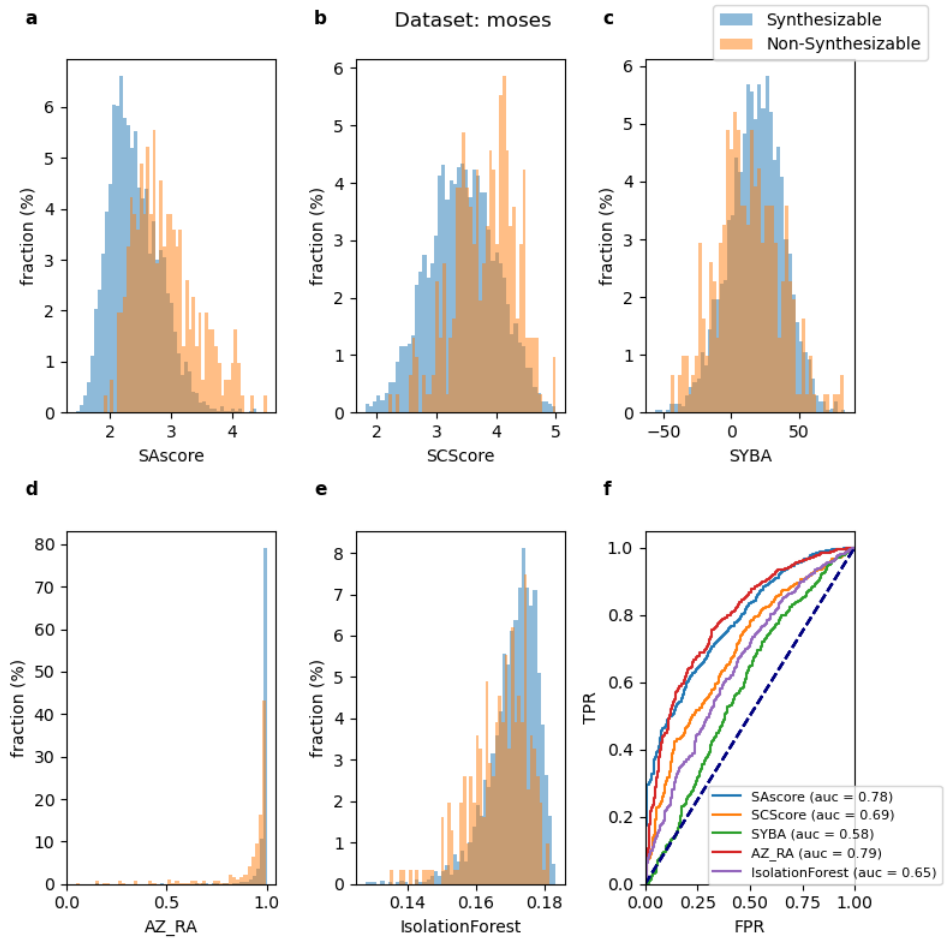
Supplementary Figure 3. Scores distributions for all datasets considered in Gao & Coley (Gao and Coley 2020) classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.



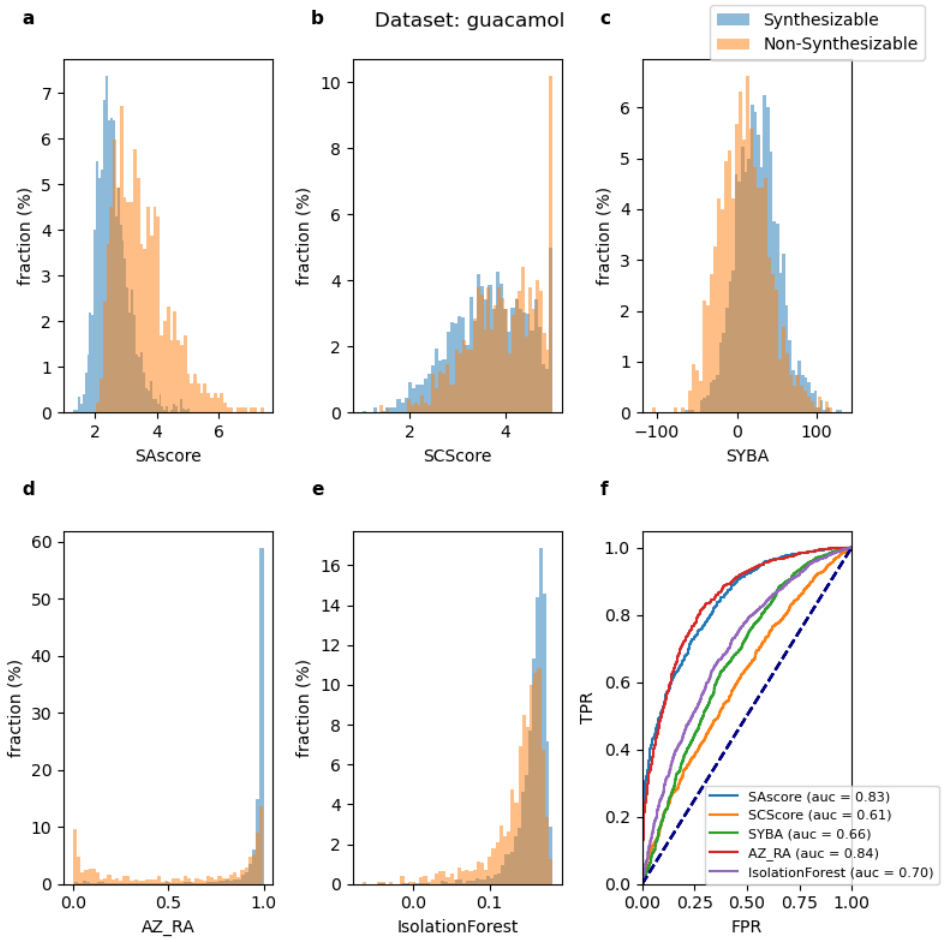
Supplementary Figure 4. Scores distributions for the ZINC dataset obtained from Gao & Coley (Gao and Coley 2020)classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.



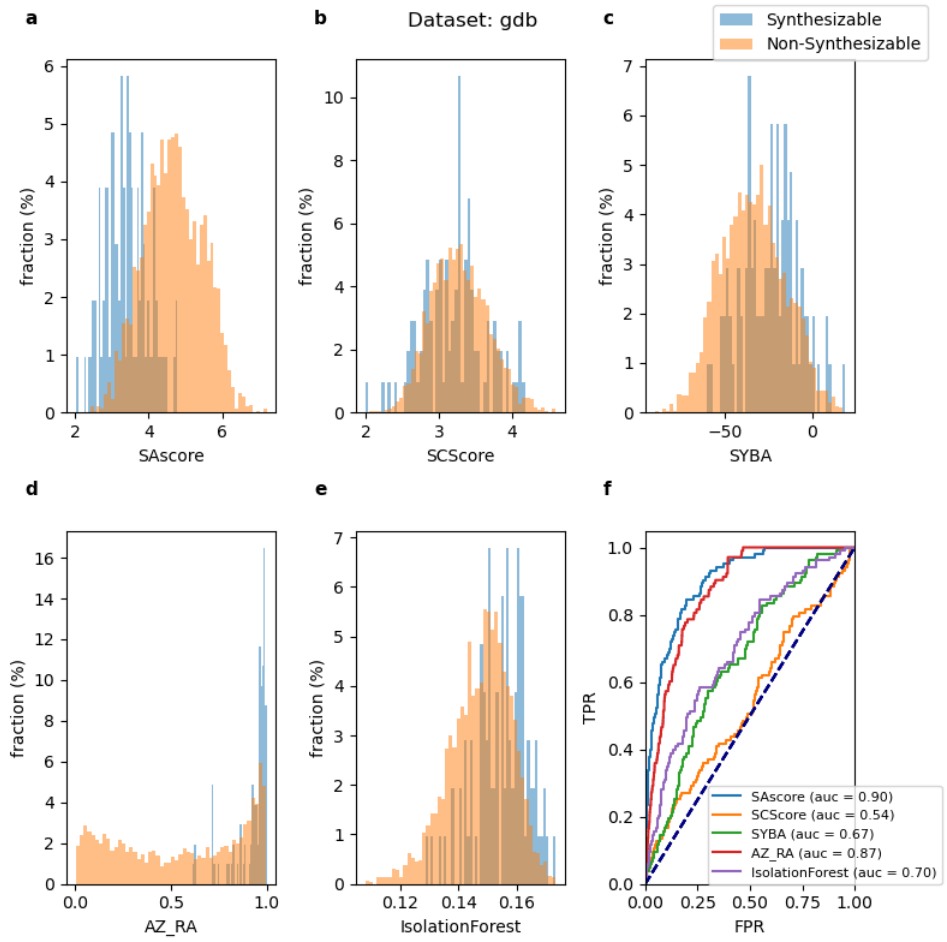
Supplementary Figure 5. Scores distributions for the Sheridan et al. (Sheridan et al. 2014) dataset obtained from Gao & Coley (Gao and Coley 2020) and classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.



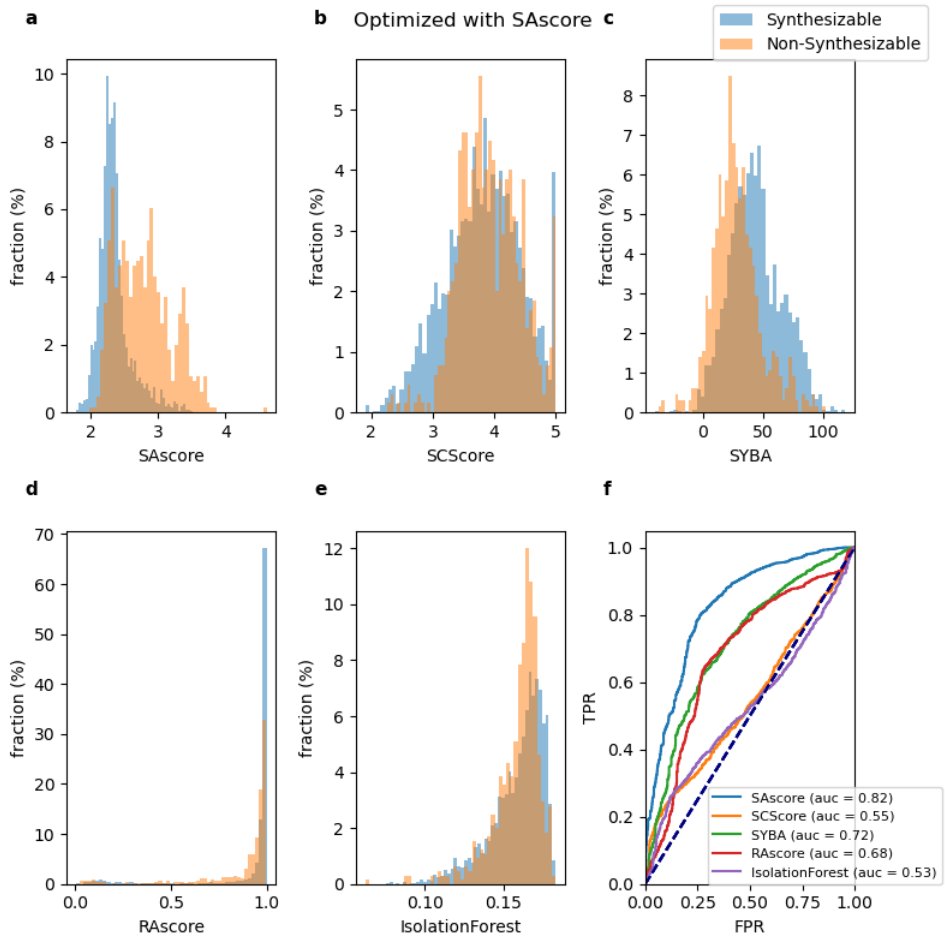
Supplementary Figure 6. Scores distributions for the MOSES dataset obtained from Gao & Coley (Gao and Coley 2020) and classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.



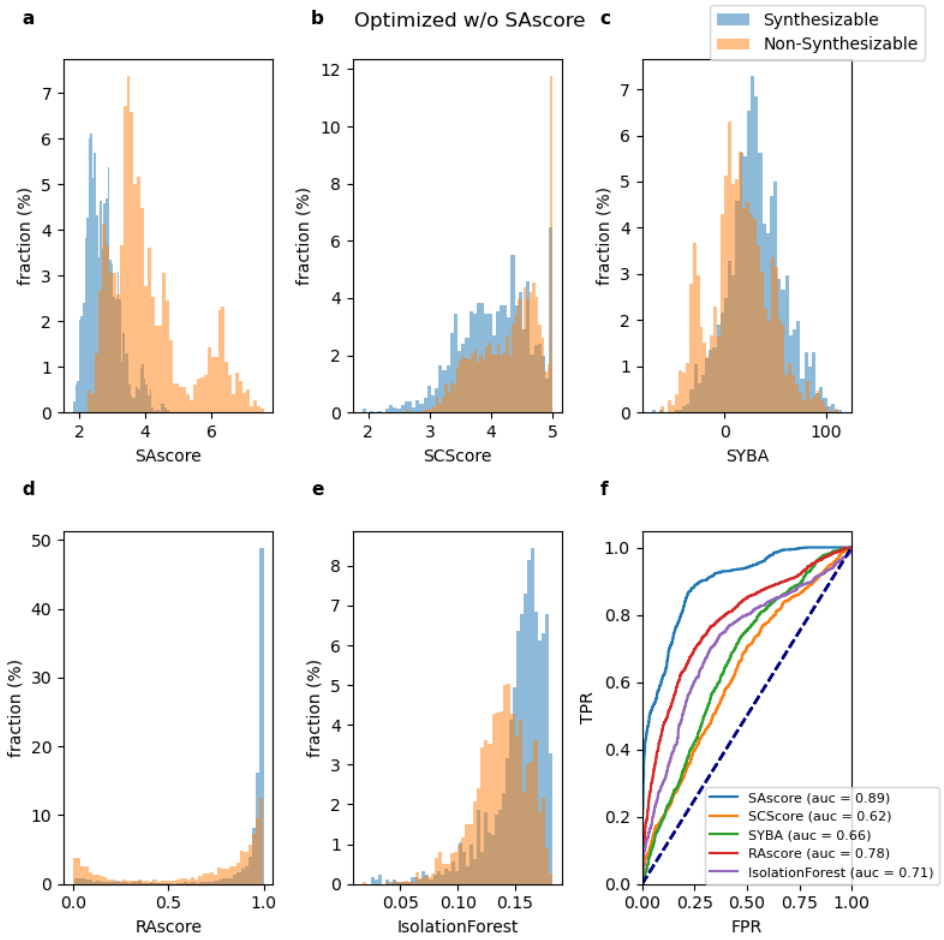
Supplementary Figure 7. Scores distributions for the guacamol dataset obtained from Gao & Coley (Gao and Coley 2020) and classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions



Supplementary Figure 8. Scores distributions for the GDB dataset obtained from Gao & Coley (Gao and Coley 2020) classified according to retrosynthesis predictions. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.

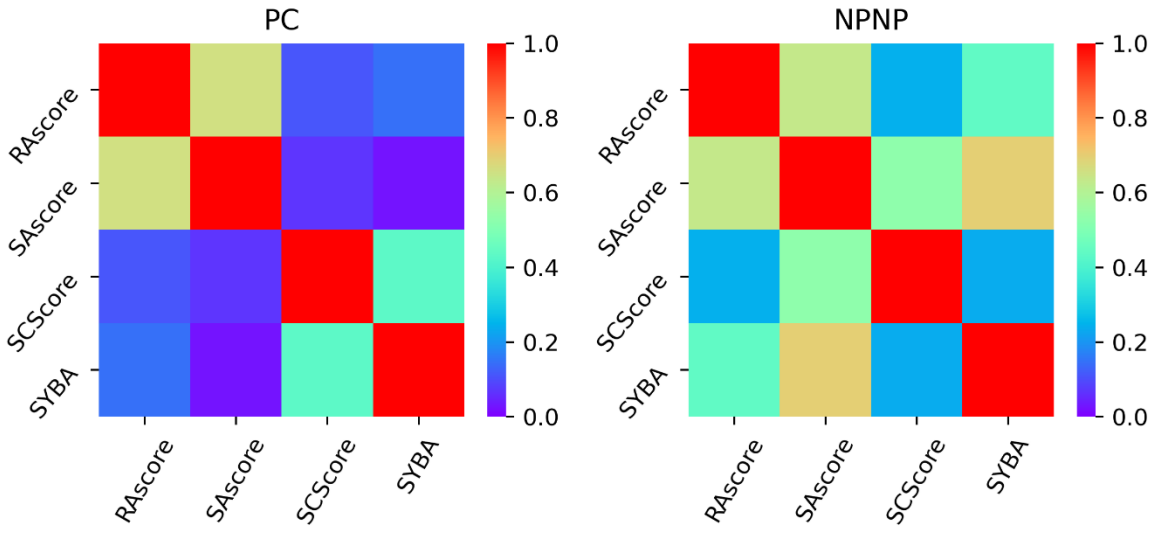


Supplementary Figure 9. Scores distributions for the goal_hard_cwa dataset obtained from Gao & Coley (Gao and Coley 2020) classified according to retrosynthesis predictions. The dataset contains de novo generated molecules optimized against multi-property objective functions and SAscore. **a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions.



Supplementary Figure 10. Scores distributions for the goal_hard_cwo dataset obtained from Gao & Coley (Gao and Coley 2020) classified according to retrosynthesis predictions. The dataset contains de novo generated and optimized molecules against different multi-property objective functions**. a-e**) Synthetic accessibility/feasibility scores computed with SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (AZ_RA) (Amol Thakkar et al. 2021), and proposed IsolationForest. **f**) ROC curves for the scores a-e using as ground truth retrosynthesis predictions

5. SA measurements correlation

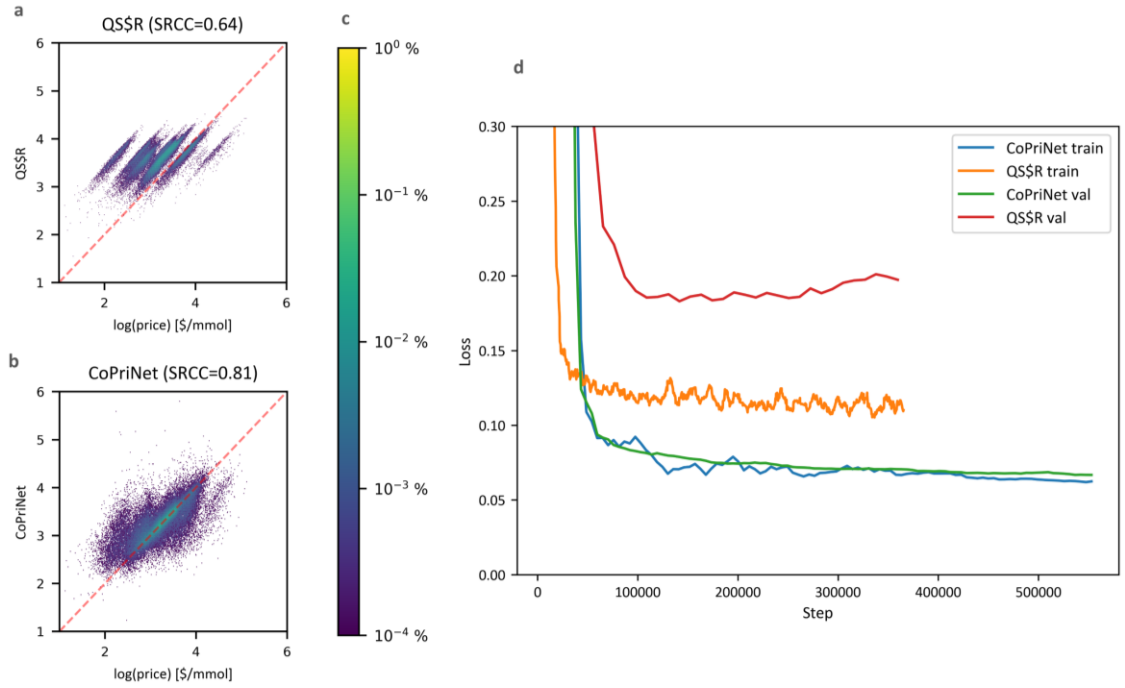


Supplementary Figure 11. Spearman's rank correlation coefficient (absolute value) for different SA measurements (SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020), RAscore (Amol Thakkar et al. 2021)) for the dataset of purchasable compounds (PC) and non-purchasable natural products (NPNP). The different scores exhibit quite different behaviour depending on the dataset leading to different most correlated scores.

6. Comparison with other approaches

Supplementary Table 3. Retrosynthesis-based price estimation compared to CoPriNet

	Original test set		Virtual test set	
	PCC	SRCC	PCC	SRCC
CoPriNet	0.77	0.80	0.27	0.56
Min building blocks price	0.73	0.75	0.52	0.68
Best building blocks price	0.85	0.87	0.62	0.75



Supplementary Figure 12. QS$R model performance when trained on the CoPriNet training set. a-b) Density heatmap for CoPriNet test set compound prices against QS$R model (a) and CoPriNet model (b) predictions when both are trained on the CoPriNet training set. Colour bar for (a-b) is shown in c). Compound prices are displayed as natural logarithm of catalogue prices. The absolute value of the Spearman's Correlation Coefficient is displayed in parenthesis (SRCC). d) Training curves for the CoPriNet model (blue and green) and the QS$R model (orange and red). Training loss (blue and orange) and validation loss (green and orange) are potted against the step number.

7. CoPriNet generalizability to virtual compounds.

CoPriNet predictions and SA measurements were computed for the testing dataset of virtual compounds obtained from the Mcule catalogue. The catalogue prices for these compounds are estimations provided by the vendors and exhibit quite a different distribution (see Supplementary Figure 13 a) compared to the distribution of prices in CoPriNet training/validation/testing sets, with most of the compounds showing larger prices. This is not surprising since virtual compounds need to be synthesized and success is not guaranteed. Additionally, the fact that different vendors have different price estimation protocols makes the problem even harder.

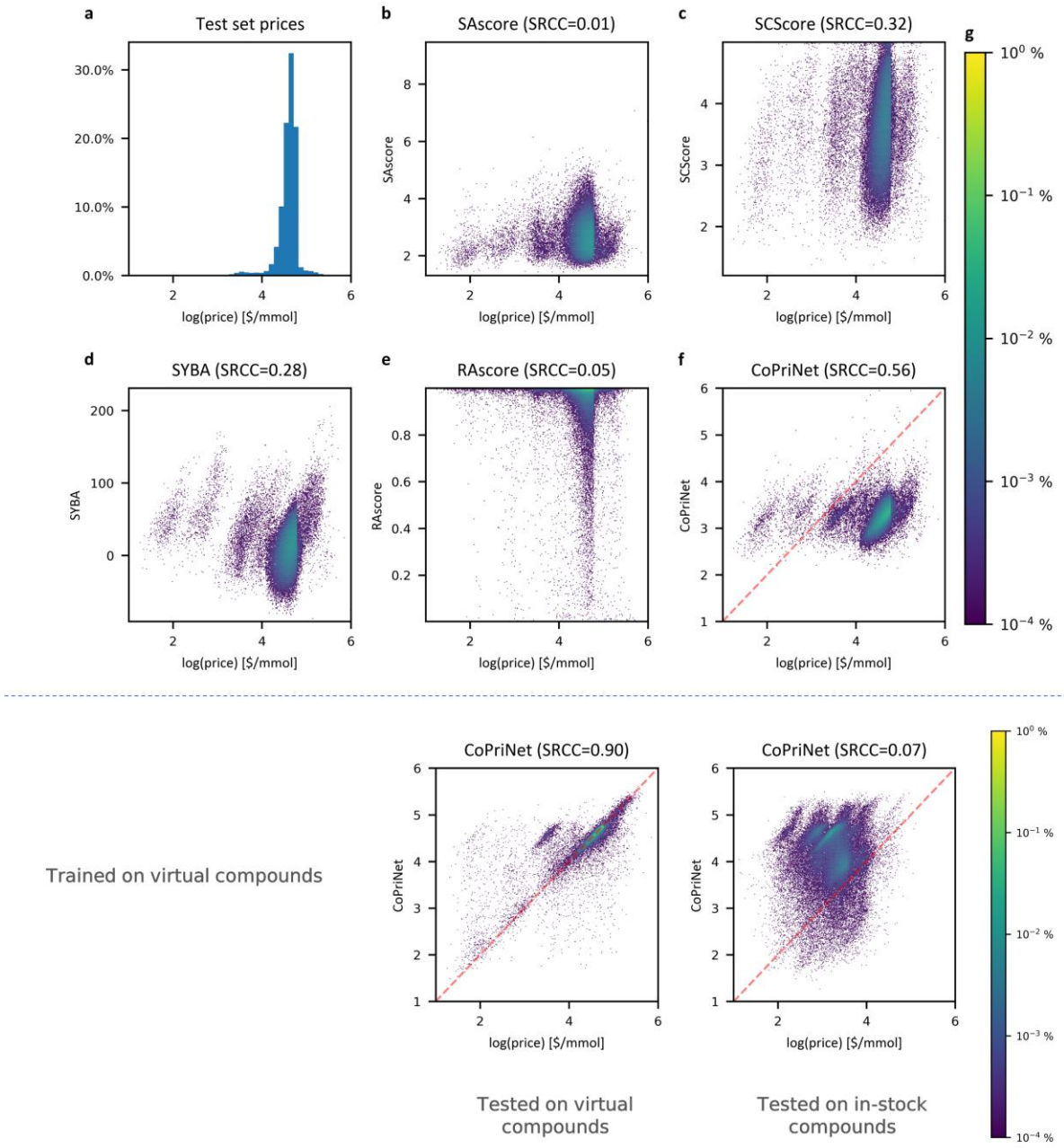
From direct inspection of Supplementary Figure 13, it is obvious that for this dataset the relationship between price and SA is far weaker as there is a massive drop in correlation for all scores. Nevertheless, CoPriNet correlation stills remains well above the other methods. Looking at Supplementary Figure 13 f, it can be observed that CoPriNet predictions systematically underestimate prices, which is also not surprising as the price distribution of the training set is shifted towards lower prices. Although this systematic bias in the predictions for this set of

compounds prevents CoPriNet from obtaining accurate price estimations, its impact in compound ranking by price is far less severe, as the smaller drop in SRCC demonstrates. Indeed, as Supplementary Figure 14 shows, much of this lack of linear correlation is caused by the differences in compound prices depending on the vendors. However, within a particular vendor, CoPriNet predictions also exhibit good linear correlation, suggesting generalizability beyond the training dataset.

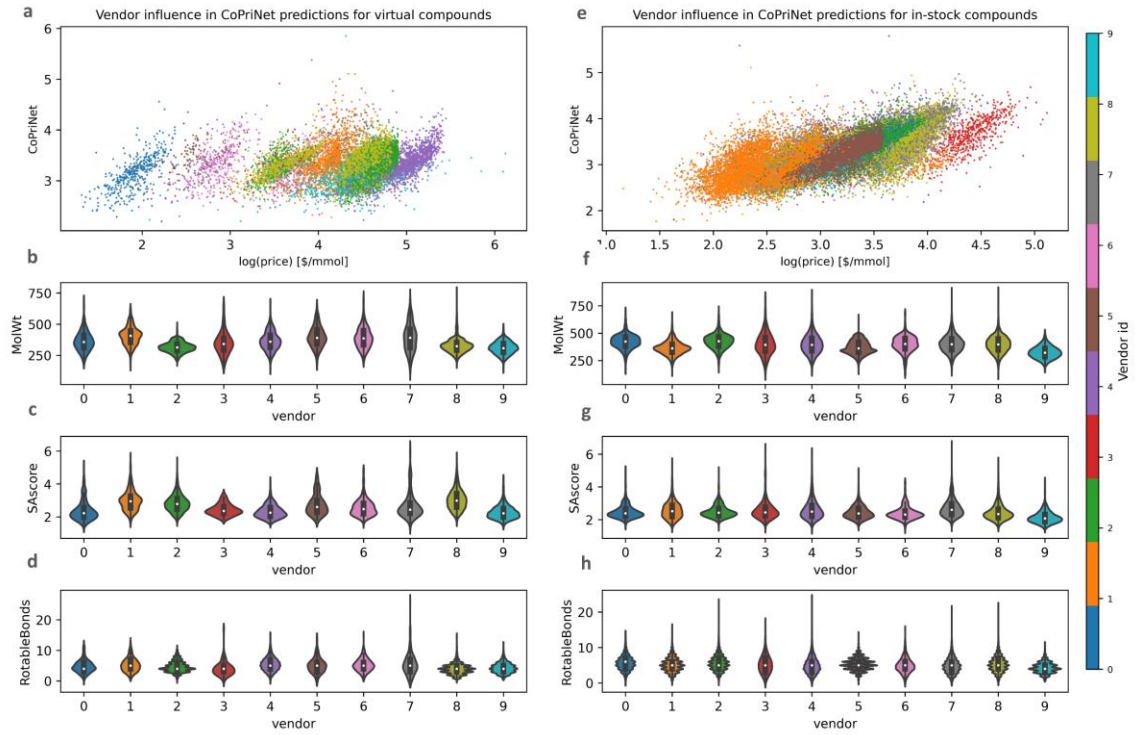
It is possible that the reason behind CoPriNet under-pricing virtual compounds are the extra fees that are applied to compounds never synthesized before. We have, therefore, also trained a model using a training set of virtual compounds only. This model is able to achieve excellent performance for the test dataset of virtual compounds (see Supplementary Figure 13 bottom), with a SRCC of 0.9. However, when evaluated on the PC dataset, the model tends to overprice all compounds and the overall performance is far worse (SRCC=0.07), which suggest poor generalization. Given these results, an alternative approach aimed to generalize better for the two scenarios could be combining a model trained with in-stock compounds and a model trained with virtual compounds, or a model trained on a mixture of both virtual and in-stock compounds. The challenge of this approach would be to define the weight for each model or dataset and how to properly evaluate the performance, since the proportion of each compound type is not known in advance. One way to circumvent this problem for a particular user could be to recalibrate such weights given its historical data.

Supplementary Table 4. Absolute value of the Pearson's correlation coefficient (PCC) and the Spearman's rank correlation coefficient (SRCC) for CoPriNet and SA measurements against estimated price for the testing dataset of virtual compounds obtained from the Mcule catalogue.

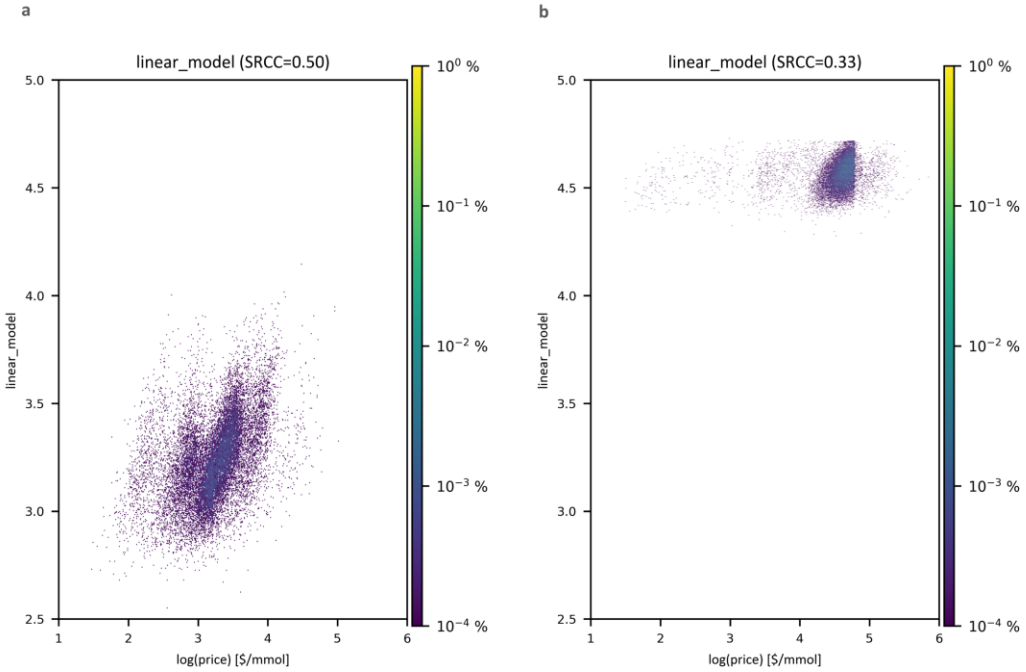
	Original test set		Virtual test set	
	PCC	SRCC	PCC	SRCC
CoPriNet	0.77	0.80	0.27	0.56
SAscore	0.16	0.16	0.03	0.00
RAscore	0.16	0.16	0.03	0.05
SCScore	0.32	0.32	0.19	0.32
SYBA	0.35	0.41	0.02	0.28



Supplementary Figure 13. **Top**. Synthetic accessibility scores correlate poorly with virtual compound price while CoPriNet prediction exhibits better correlation. **a**) Histogram of the compound prices of the virtual compounds test set; **b**-**e**) Density heatmaps for CoPriNet virtual compounds test set compound prices against four different SA scores: SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020) and RAscore (Amol Thakkar et al. 2021); **f**) Density heatmap for CoPriNet virtual compounds test set compound prices against CoPriNet predictions. Compound prices are displayed as natural logarithm of catalogue prices. The absolute value of the Spearman's Rank Correlation Coefficient is displayed in parenthesis (SRCC). **g**) Colour bar showing percentages of the total test size per bucket. **Bottom**. CoPriNet model trained on virtual compounds can accurately predict virtual compound prices (left), but it fails to predict prices for in-stock compounds (PC, right).



Supplementary Figure 14. Different vendors use different pricing strategies for virtual compounds, which makes price prediction challenging, but has little impact for compound ranking within a given catalogue. Compound price, in $/mmol, against CoPriNet predictions for the compounds included in the test set of virtual compounds (a, same data as in Supplementary Figure 13) and the set of in-stock compounds (e, PC), coloured by vendor. For visualization reasons, only the top-10 most frequent vendors have been displayed. Subplots b-d display the per-vendor distributions of compound molecular weight, SAscore and number of rotatable bonds for the virtual compounds set and f-h for the set of in-stock compounds.

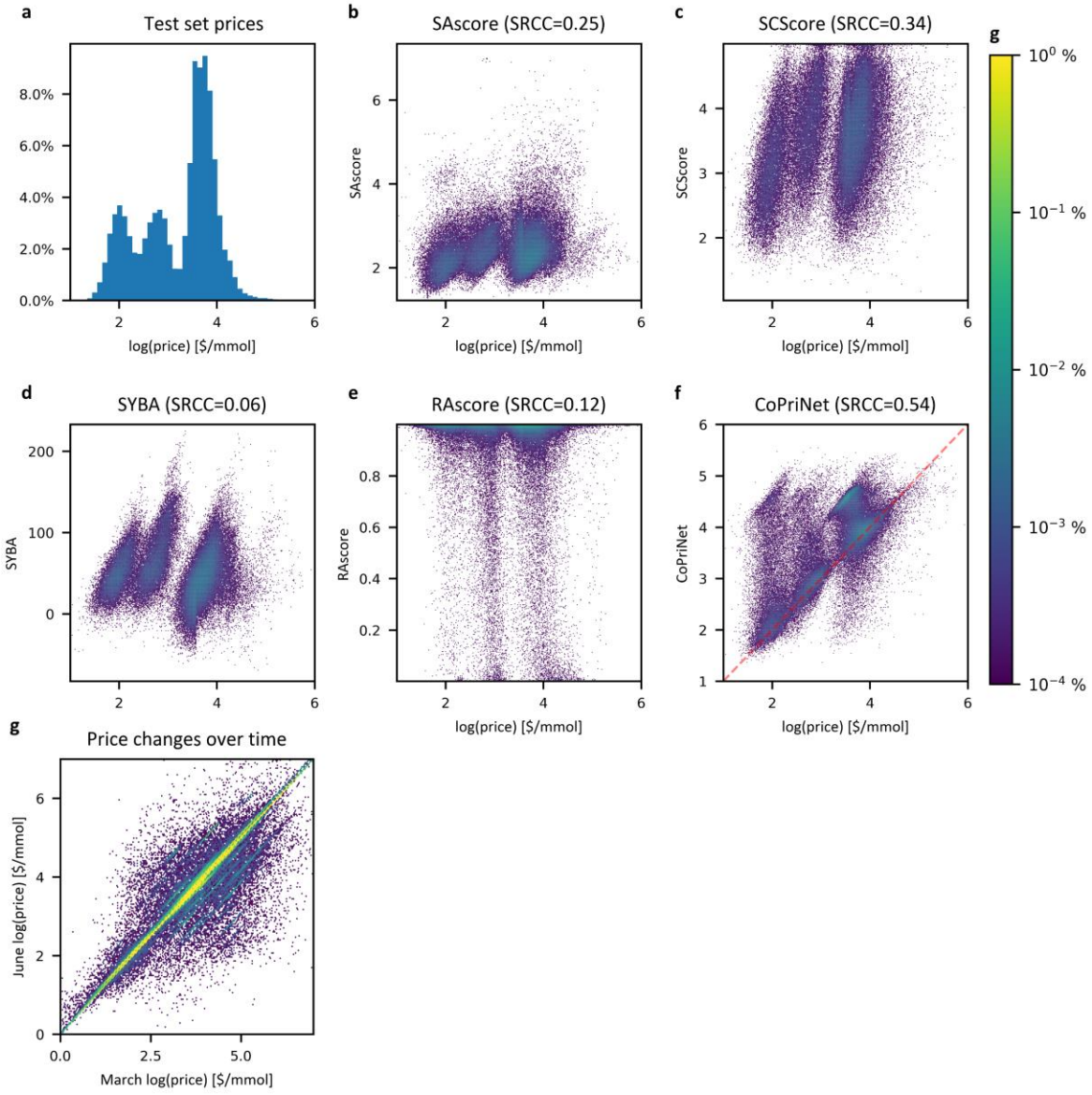


Supplementary Figure 15. Linear model combining SA scores do not predict compound prices accurately. a) linear regression model trained on purchasable in-stock compounds and evaluated on the CoPriNet test set (PC). b) linear regression model trained on virtual compounds and evaluated on the CoPriNet virtual compounds test set. Linear models combine four SA scores: SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020) and RAscore (Amol Thakkar et al. 2021).

8. CoPriNet generalizability over time.

CoPriNet predictions and SA measurements were computed for the testing dataset comprised of in-stock compounds that were present on the June 2021 Mcule database release but that were not

present in the March 2021 release, the one used for training. Supplementary Figure 16 shows that, as in all other experiments, CoPriNet scores (f) correlate much better with compound price than any other SA score (b-e). Although it is true that the correlation is not as strong as in the default testing set (PCC of 0.65 vs 0.77 and SRCC 0f 0.54 vs 0.81), it is important to highlight that prices are not static values and tend to change over time, as shown in Supplementary Figure 16 (g) for the compounds that were present in both releases. As a consequence, drops in performance are expected based on the evolution of compound prices.



Supplementary Figure 16. CoPriNet predictions are relatively robust over time, as shown when evaluated on the temporal test set that is comprised of molecules that were added to the June 2021 Mcule release and were not present in the March 2021 release. a) Histogram of the temporal test compound prices (compounds that were added to the new version of the catalogue); b-e) Density heatmaps for the temporal test set compound prices against four different SA scores: SAscore (Ertl and Schuffenhauer 2009), SCScore (Coley et al. 2018), SYBA (Voršilák et al. 2020) and RAscore (Amol Thakkar et al. 2021); f) Density heatmap for the temporal test set compound prices against CoPriNet predictions. Compound prices are displayed as natural logarithm of catalogue prices. The absolute value of the Spearman's Rank Correlation Coefficient is displayed in parenthesis (SRCC). h) Colour bar for subplots b-g displaying the percentage of the temporal test dataset in each bucket.

9. GNN hyperparameters

The hyperparameters of our method were selected by random search using the validation error as optimization goal. The model showed robustness against hyperparameter choices. The selected hyperparameters and the candidate values were:

- Number of layers: 10 [4 ,6, 10]
- PNA:
    - Node channels: 75 [25, 50, 75, 100]
    - Edge channels 50 [25, 50, 75, 100]
    - Towers: 5 [1, 5]
- Dense layers: 2 [1,2,3]
    - Units: 50 [25, 50, 75, 100]
    - Dropout 0 [0, 0.5]
    - Set2set: Steps: 6 [4, 6]

10. Purchasability as a ground truth for SA

While the correlation between SA scores and price is weak, main text Figure 3 shows that more expensive compounds tend to be less synthetically accessible. This suggests that the purchasability of the compounds, and more specifically their prices, could be reasonable criteria for compiling datasets for machine learning-based SA estimators. Such ground truth could be regarded as a refined version of the presence in a commercial catalogue that is exploited in the SYBA method (Voršilák *et al.*, 2020). However, with the aim of avoiding the usage of artificially generated negative compounds, instead of a binary classifier, we have employed a simple anomaly detection method, the isolation forest (Liu *et al.*, 2008). Anomaly detection methods allow to identify instances that differ from the training examples, thus only requiring a dataset of positive compound. This allows a better understanding on the ground truth data impact on the results.

We implemented an isolation forest model (Liu *et al.*, 2008) with 2000 trees, each trained on a random 10% subset of the data. The training set is constructed as a random subset of 1M inexpensive in-stock compounds (price < 250$/g) extracted from the Mcule dataset. This dataset is comparable in size to the SYBA training set. The price threshold selection has only a minor impact in the model performance (see Supplementary Table 5). Compounds are encoded using 209 descriptors computed with RDKit version 2020.09.1 (RDKit) (see Supplementary Material 6 for a complete list).

Our best performing isolation forest, trained on compounds with prices <$250/g, is able to compete with most of the studied SA scores. Thus, when evaluated on the SYBA testing set, the most similar approach, we measured a ROC AUC of 0.99, comparable to the one reported for SYBA. More interestingly, as shown in Supplementary Table 1, the isolation forest produces score distributions for the NPNP and PC datasets that are as separable as in some of the other methods. Indeed, for this experiment, the isolation forest score correlates better with the retrosynthesis-based score than the SCScore or SYBA. Similar results are obtained when evaluated in additional dataset, as displayed in Supplementary Material Section 3, which suggests that this simple proof-of-concept is able to compete with state-of-the-art approaches.

Supplementary Table 5. Correlation for the IsolationForest scores against the retrosynthesis-based score ManifoldSA on the validation set depending on the composition of the training set. It is important to note that less than 5% of the compounds included in the datasets have prices >500$/g, thus the difference between threshold 500 and infinite is minor.

	Validation set	
Price <	PCC	SRCC
50	0.733	0.618
250	0.738	0.623
500	0.737	0.615
Inf	0.734	0.616

Supplementary Table 6. Name of the descriptors used for the anomaly detection approach (IsolationForest).

MaxEStateIndex	HeavyAtomCount	PEOE_VSA4	fr_benzene
MinEStateIndex	NHOHCount	PEOE_VSA5	fr_benzodiazepine
MaxAbsEStateIndex	NOCount	PEOE_VSA6	fr_bicyclic
MinAbsEStateIndex	NumAliphaticCarbocycles	PEOE_VSA7	fr_diazo
qed	NumAliphaticHeterocycles	PEOE_VSA8	fr_dihydropyridine
MolWt	NumAliphaticRings	PEOE_VSA9	fr_epoxide
HeavyAtomMolWt	NumAromaticCarbocycles	SMR_VSA1	fr_ester
ExactMolWt	NumAromaticHeterocycles	SMR_VSA10	fr_ether
NumValenceElectrons	NumAromaticRings	SMR_VSA2	fr_furan
NumRadicalElectrons	NumHAcceptors	SMR_VSA3	fr_guanido
MaxPartialCharge	NumHDonors	SMR_VSA4	fr_halogen
MinPartialCharge	NumHeteroatoms	SMR_VSA5	fr_hdrzine
MaxAbsPartialCharge	NumRotatableBonds	SMR_VSA6	fr_hdrzone
MinAbsPartialCharge	NumSaturatedCarbocycles	SMR_VSA7	fr_imidazole
FpDensityMorgan1	NumSaturatedHeterocycles	SMR_VSA8	fr_imide
FpDensityMorgan2	NumSaturatedRings	SMR_VSA9	fr_isocyan
FpDensityMorgan3	RingCount	SlogP_VSA1	fr_isothiocyan
BCUT2D_MWHI	MolLogP	SlogP_VSA10	fr_ketone
BCUT2D_MWLOW	MolMR	SlogP_VSA11	fr_ketone_Topliss
BCUT2D_CHGHI	fr_Al_COO	SlogP_VSA12	fr_lactam
BCUT2D_CHGLO	fr_Al_OH	SlogP_VSA2	fr_lactone
BCUT2D_LOGPHI	fr_Al_OH_noTert	SlogP_VSA3	fr_methoxy
BCUT2D_LOGPLOW	fr_ArN	SlogP_VSA4	fr_morpholine
BCUT2D_MRHI	fr_Ar_COO	SlogP_VSA5	fr_nitrile
BCUT2D_MRLOW	fr_Ar_N	SlogP_VSA6	fr_nitro
BalabanJ	fr_Ar_NH	SlogP_VSA7	fr_nitro_arom
BertzCT	fr_Ar_OH	SlogP_VSA8	fr_nitro_arom_nonortho
Chi0	fr_COO	SlogP_VSA9	fr_nitroso
Chi0n	fr_COO2	TPSA	fr_oxazole
Chi0v	fr_C_O	EState_VSA1	fr_oxime
Chi1	fr_C_O_noCOO	EState_VSA10	fr_para_hydroxylation
Chi1n	fr_C_S	EState_VSA11	fr_phenol

Chi1v	fr_HOCCN	EState_VSA2	fr_phenol_noOrthoHbond
Chi2n	fr_Imine	EState_VSA3	fr_phos_acid
Chi2v	fr_NH0	EState_VSA4	fr_phos_ester
Chi3n	fr_NH1	EState_VSA5	fr_piperdine
Chi3v	fr_NH2	EState_VSA6	fr_piperzine
Chi4n	fr_N_O	EState_VSA7	fr_priamide
Chi4v	fr_Ndealkylation1	EState_VSA8	fr_prisulfonamd
HallKierAlpha	fr_Ndealkylation2	EState_VSA9	fr_pyridine
Ipc	fr_Nhpyrrole	VSA_EState1	fr_quatN
Kappa1	fr_SH	VSA_EState10	fr_sulfide
Kappa2	fr_aldehyde	VSA_EState2	fr_sulfonamd
Kappa3	fr_alkyl_carbamate	VSA_EState3	fr_sulfone
LabuteASA	fr_alkyl_halide	VSA_EState4	fr_term_acetylene
PEOE_VSA1	fr_allylic_oxid	VSA_EState5	fr_tetrazole
PEOE_VSA10	fr_amide	VSA_EState6	fr_thiazole
PEOE_VSA11	fr_amidine	VSA_EState7	fr_thiocyan
PEOE_VSA12	fr_aniline	VSA_EState8	fr_thiophene
PEOE_VSA13	fr_aryl_methyl	VSA_EState9	fr_unbrch_alkane
PEOE_VSA14	fr_azide	FractionCSP3	fr_urea
PEOE_VSA2	fr_azo	PEOE_VSA3	nStereo
fr_barbitur			

11. References

Amol Thakkar et al. 2021. “Retrosynthetic Accessibility Score (RAscore) – Rapid Machine Learned Synthesizability Classification from AI Driven Retrosynthetic Planning.” *Chemical Science* 12(9): 3339–49. https://pubs.rsc.org/en/content/articlehtml/2021/sc/d0sc05401a (September 1, 2021).

Coley, Connor W., Luke Rogers, William H. Green, and Klavs F. Jensen. 2018. “SCScore: Synthetic Complexity Learned from a Reaction Corpus.” *Journal of Chemical Information and Modeling* 58(2): 252–61. https://pubs.acs.org/doi/full/10.1021/acs.jcim.7b00622 (August 31, 2021).

Ertl, Peter, and Ansgar Schuffenhauer. 2009. “Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions.” *Journal of Cheminformatics 2009 1:1* 1(1): 1–11. https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-1-8 (August 18, 2021).

Gao, Wenhao, and Connor W. Coley. 2020. “The Synthesizability of Molecules Proposed by Generative Models.” *Journal of Chemical Information and Modeling* 60(12). https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00174 (September 1, 2021).

Sheridan, Robert P. et al. 2014. “Modeling a Crowdsourced Definition of Molecular Complexity.” *Journal of Chemical Information and Modeling* 54(6): 1604–16. https://pubs.acs.org/doi/abs/10.1021/ci5001778 (October 11, 2021).

Voršilák, Milan, Michal Kolář, Ivan Čmelo, and Daniel Svozil. 2020. “SYBA: Bayesian Estimation of Synthetic Accessibility of Organic Compounds.” *Journal of Cheminformatics 2020 12:1* 12(1): 1–13. https://jcheminf.biomedcentral.com/articles/10.1186/s13321-020-00439-2 (September 6, 2021).