A Novel Automated Screening Method For Combinatorial Small Molecules

Pingshi Yu,[*,†,‡] Alistair Sterling,[*,¶] and Jotun Hein[*,†]

†*Department of Statistics, University of Oxford*

‡*Department of Computer Science, University of Oxford*

¶*Department of Chemistry, University of Oxford*

E-mail: pingshiyu@gmail.com; alistair.sterling@chem.ox.ac.uk; hein@stats.ox.ac.uk

Abstract

A main challenge in the enumeration small molecule chemical spaces for drug design is to quickly and accurately differentiate between possible and impossible molecules. Current approaches for screening enumerated molecules (e.g. 2D heuristics, 3D force-fields) have not been able to achieve a balance between accuracy and speed. We have developed a new automated approach for fast and high-quality screening of small molecules, with the following steps: 1) for each molecules in the set, compute an ensemble of 2D descriptors as feature encoding, 2) on a random small subset, generate classification (feasible/infeasible) targets via a 3D-based approach, 3) form a classification dataset with the computed features and targets, and train a machine learning model for predicting the 3D approach's decisions, 4) use the trained model to screen the remainder of the enumerated set. Our approach is $\approx 8\times$ ($7.96\times$ to $8.84\times$) faster than screening via 3D simulations without significantly sacrificing accuracy; whilst compared to 2D-based pruning rules, this approach is more accurate, with better coverage of known feasible molecules. Once the topological features and 3D conformer

evaluation methods are established, the process can be fully automated, without any additional chemistry expertise.

Contents

Introduction

Isomer and chemical space enumeration is an inter-disciplinary endeavour for combinatorial mathematicians, computer scientists and chemists alike. The representation of molecular structures as mathematical graphs opens up an opportunity for application of methods in graph theory to chemistry. One one hand, the enumeration of isomers (or, in the language of graph theory, enumeration of sets of graphs with the same 'degree sequence'[1,2]) is a tool of central importance to structural elucidation.[3] On the other hand, the enumeration of small molecule chemical spaces have been an exciting approach in drug design and discovery.[4–8]

Graph enumeration methods will generate a list of chemical graphs that satisfies certain basic mathematical/chemical properties or restrictions (e.g. tree-like, chemical formula, molecular weight etc). However, not all of the mathematically possible graphs necessarily have a chemically feasible conformer (3D arrangement of the nodes) - hence, many of these graphs are unlikely to correspond to any reasonable molecules. It is thus desirable to obtain a list of graphs which are also 'chemically realistic', in that they have reasonable conformers.

Currently, there are two main paradigms for deciding the physical realism of chemical graphs:

1. Heuristics (2D) based: rules applied directly on the molecular graph. These usually involve chemical knowledge on what substructures will result in realistic molecules - for example rejecting all graphs that are non-planar,[4,9] or contains a fragment from a list of undesirable substructures.[10]

2. Force-field/simulation (3D) based: 1) for a molecular graph, first generate an explicit

3D conformer (e.g. via distance geometry,[11] or via fragment based constructions[12,13]), then 2) subjecting it to a physical simulation, finding a potential energy minima, and testing for its stability. When the conformer seems like a reasonable structure, the graph is accepted - and is otherwise rejected.

An example of the heuristics approach can be found as a part of MOLGEN,[10] which contains a *badlist* (list of undesirable substructures). MOLGEN provides the option to prevent graphs containing any substructures in the badlist from being enumerated. Faulon[14] has explored the force-field/simulation based approach to test molecular isomers in a series of studies on stochastic molecule generators, where the calculated potential energies is used to assess the quality of sampled isomers. The GDB chemical spaces[4] uses both force-fields and heuristics approaches during enumeration: each smaller graph (molecules with $\leq$ 11 heavy atoms) first has an estimated conformer computed, followed by a forcefield simulation thereafter to optimize the coordinates. The graph is rejected if the model fails a feasibility test (e.g. atomic volume around 1-4 carbons). For larger graphs, Reymond resorts back to increasingly aggressive graph-based rules to reject molecules, in order to tame the combinatorial explosion that comes with the increase in atom count.

One limitation to the badlist, and other 2D based approaches, is that they may, at times, reject genuine molecules.[10] For example, as seen in GDB17 - around 15-20% of PubChem[15]/ChEMBL[16] molecules do not pass GDB's enumeration rules. The advantage is that they are very fast to run - hence these were used in GDB when the size of chemical space grew quickly. MOLGEN can even prune during enumeration by stopping the orderly generation[5,10] process when a badlist substructure appears. Force-field/simulation (3D) based methods are less likely to reject genuine molecules - as calculations are based on first principles.[13,17,18] In the majority of cases, when topology of real molecules are used, we would expect common conformer generators ([13]) to find reasonable conformers. The conformer generation and the subsequent physical simulation will, of course, be computationally much more demanding as compared to a purely 2D based computation. Thus, 2D based approaches

are fast, but suffer from low accuracy; whilst 3D based approaches are more accurate, but suffer from slow speed. This becomes problematic when screening large chemical spaces.

We propose a method which is a synthesis of the two paradigms, with advantages of both: still primarily based on 2D representation of molecules, the speed benefit is preserved; whilst largely maintaining the accuracy of simulation-based methods. For each molecule, an encoding with a collection of 2D descriptors is computed, each descriptor capturing a different quantitative aspect of the molecule's topology (similar to QSAR studies). The descriptors set should be chosen such that they contain enough information about the molecule's topology to determine whether it leads to 'realistic' conformers. A small set is sampled from the candidate structures, forming the training set - for each of these molecules, targets of 'realism' is generated by force-field/simulation based methods. In this study, ETKDG[19] and UFF[17] (as implemented in RDKit[20]) are used to generate and evaluate the realism of generated conformers. Only the molecules in the training set require explicit generation of conformers. Once trained, the model can be used on the remaining structures, where only 2D representations needs to be calculated for realism prediction. Due to the vast number of combinatorially possible molecules, the computational cost of generating conformers for the sampled training set is small relative to the prohibitively expensive 3D simulations for the remainder of the molecules, which is avoided.

Methods

Combinatorial Enumeration of Molecules

Combinatorial chemical spaces were generated with our extended version of PMG[21] (parallelized OMG[22]), to enumerate all $CHNOPS$ molecules under a certain total molecular weight w. OMG/PMG enumerates all isomers of a certain molecular formulae by orderly generation. To enumerate chemical spaces of molecular weight $(MW) \leq w$: 1) all molecular formulae with total MW under w is enumerated, and 2) for each formulae, PMG is called to

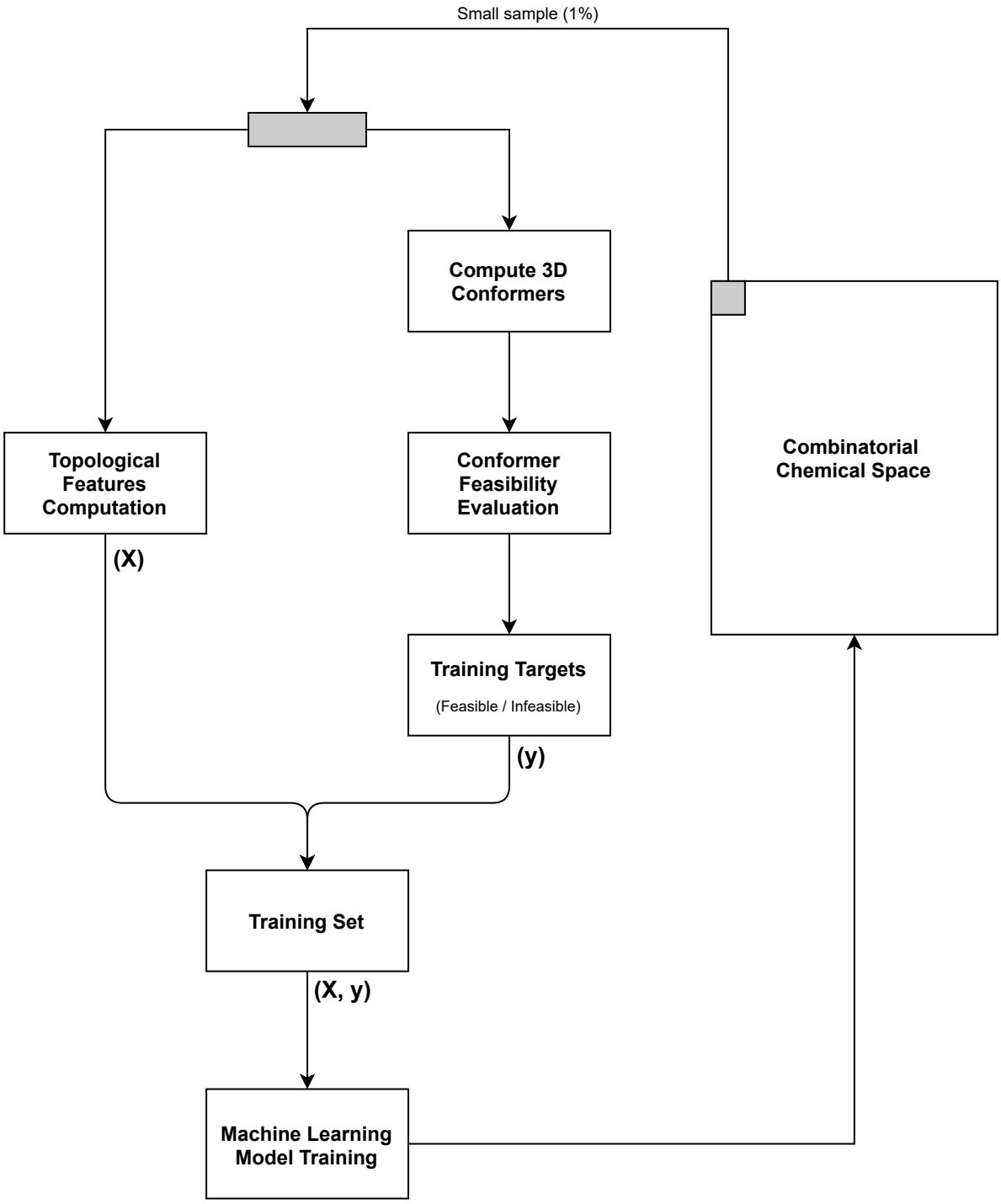

Figure 1: Overall steps of the proposed approach. For a set of combinatorially generated molecules: 1) compute a 2D-based feature encoding using descriptors, topological indices and fingerprints, 2) extract a small ($\approx 1\%$) uniform set of molecules to be included in the training set, 3) for this subset, evaluate, using 3D based methods, a binary (0/1) target corresponding to the predicted physical feasibility of each molecule, 4) using the 2D feature encoding, together with the computed 3D feasibility targets, train a machine learning model to predict the feasibility from the 2D features, and finally 5) use the learned model to screen the larger chemical space.

generate all the corresponding isomers. Unlike in the original OMG/PMG, pentavalent nitrogens were not included in our enumeration. In this fashion, we enumerate a combinatorial space of small molecules with $MW \leq 125$ ($33,846,411$ molecules).

Topological Features

Altogether, 570 2D features were computed in parallel. 206 of these were topological indices/descriptors from RDKit;[20] 167 were binary features extracted from the MACCS fingerprint,[23] indicating the presence/abasence of certain SMARTS patterns; 192 were the 2D Broto-Moreau (Autocorrelation) descriptors;[24] the Wiener index[25] which we implemented; 4 were the *crowding indices* which we developed (see below), evaluated at $d = 1, 2, 3, 4$.

Crowding Index

Given a graph $G = (V, E)$, and $v \in V$, define the distance d neighbourhood of v as:

$$N_G(v, d) = |\{u \in V : d(u, v) \leq d\}| \tag{1}$$

where for $x, y \in V$, $d(x, y)$ is the topological distance between nodes x, y. The crowding index $\phi_G(d)$, then, is the size of the largest distance d-neighbourhood within the graph G:

$$\phi_G(d) = \max_{v \in V} N(v, d) \tag{2}$$

Thus, the crowding index $\phi_G(d)$ is the size of the 'most crowded' region of the graph G within distance d of any vertex. Similar to the choice made in much of chemical graph theory,[26] we consider *hydrogen-depleted* graphs. Left unconstrained, combinatorially possible graphs (for example, n-furcating trees) can have their crowding indices grow exponentially with increasing d. However, physically realisable molecules are constrained by space, which would grow in order of $O(d^3)$. For example: for $d = 3$, the largest possible value that $\phi_G(3)$ can

take (assuming maximum degree of 4), is 53; however, in uniform random samples of $100,000$ PubChem and ChEMBL molecules, no molecules had $\phi(3)$ exceeding 33. A high crowding index would then act as a indication of sterically highly strained topologies, and is evidence against their physical feasibility.

Conformer Generation

3D conformers were generated from 2D topologies with RDKit's[20] implementation of the ETKDG[19] approach for solving distance geometry (DG) instances. In addition to the standard bond lengths information used in distance geometry algorithms, torsional angles preferences from Cambridge Structures Database (CSD) were also used to improve the quality of conformers generated. Conformers generated by ETKDG were shown to be competitive with those generated by a standard distance geometry run, followed with a force-field minimization step. The outcome of a conformer generation step is dependent on the random choice of the initial atom positions. In particular, a run may fail to generate a valid conformer due to incompatibility of the atom's 3D arrangements with constraints. To differentiate between random failures and failures due to a genuinely 'bad' topology, each molecule was given three attempts to generate a valid conformer, and was only marked as 'failure' should all attempts fail. The conformer generation process was parallelized, and were performed for samples of molecules from both combinatorial (generated by PMG) and real/realistic (PubChem,[15] ChEMBL,[16] GDB17[4]) chemical spaces. The result of the generation process is tabulated in 1. Three attempts appeared sufficient for the vast majority ($\geq 98.88\%$) of the real/realistic molecules to generate a conformer.

Conformer Quality Evaluation

A molecular topology that does not lead to a valid 3D structure satisfying basic chemical constraints would, of course, serve as evidence for its 'bad' quality. However, even when a conformer was successfully generated, it may still be highly unstable (e.g. rapidly undergoing

Table 1: Percentage of molecules failing bounds smoothing during DG (distance geometry) runs, on random samples from real (PubChem, ChEMBL), realistic (GDB17), and from combinatorial ($MW \leq 125$) chemical spaces.

Chemical Space	Sample (n)	Failed DG	Failure %
PubChem	42,992	125	0.29
ChEMBL	46,964	525	1.12
GDB17	49,906	298	0.60
$MW \leq 125$	56,189	18,239	32.46

thermal decomposition). For example, due to strained geometries, steric constraints, or other reasons (e.g. unstable functional groups). Given a conformation, we seek general metrics from which we can determine its 'quality' - that is, metrics corresponding to the conformer's feasibility. As conformers of novel/unknown molecules are included, where a 'target' conformation is not known, methods to evaluate conformations based on deviation from an 'optimal' conformation (such as RMSD[27]) cannot be used. To this end, we applied several general quality measures for conformations: those based directly on the 3D geometry (success/failure of ETKDG; bond lengths and angles based metrics), and force-fields based (potential energy computed from molecular force-fields) - described in detail below.

In chemical space screening, we would like to differentiate between the combinatorially generated molecules (where a large proportion is expected to be infeasible), and realistic molecules. For this purpose, we propose several metrics, each aiming to capture aspects of the molecule's 'soundness'. In order for the metrics to be useful, there should be significantly differences in distribution (under the metric) between the conformers of real molecules (RDB: PubChem, ChEMBL) and conformers of combinatorially generated molecules (CDB: $MW \leq 125$). A candidate molecule can then be screened based on its deviation from realistic molecules' distribution on the metrics.

Metric 1: Success of Distance Geometry Runs

The failure of distance geometry during conformation generation serves as evidence for the infeasibility of the topology. In RDKit's implementation of distance geometry,[19] failures

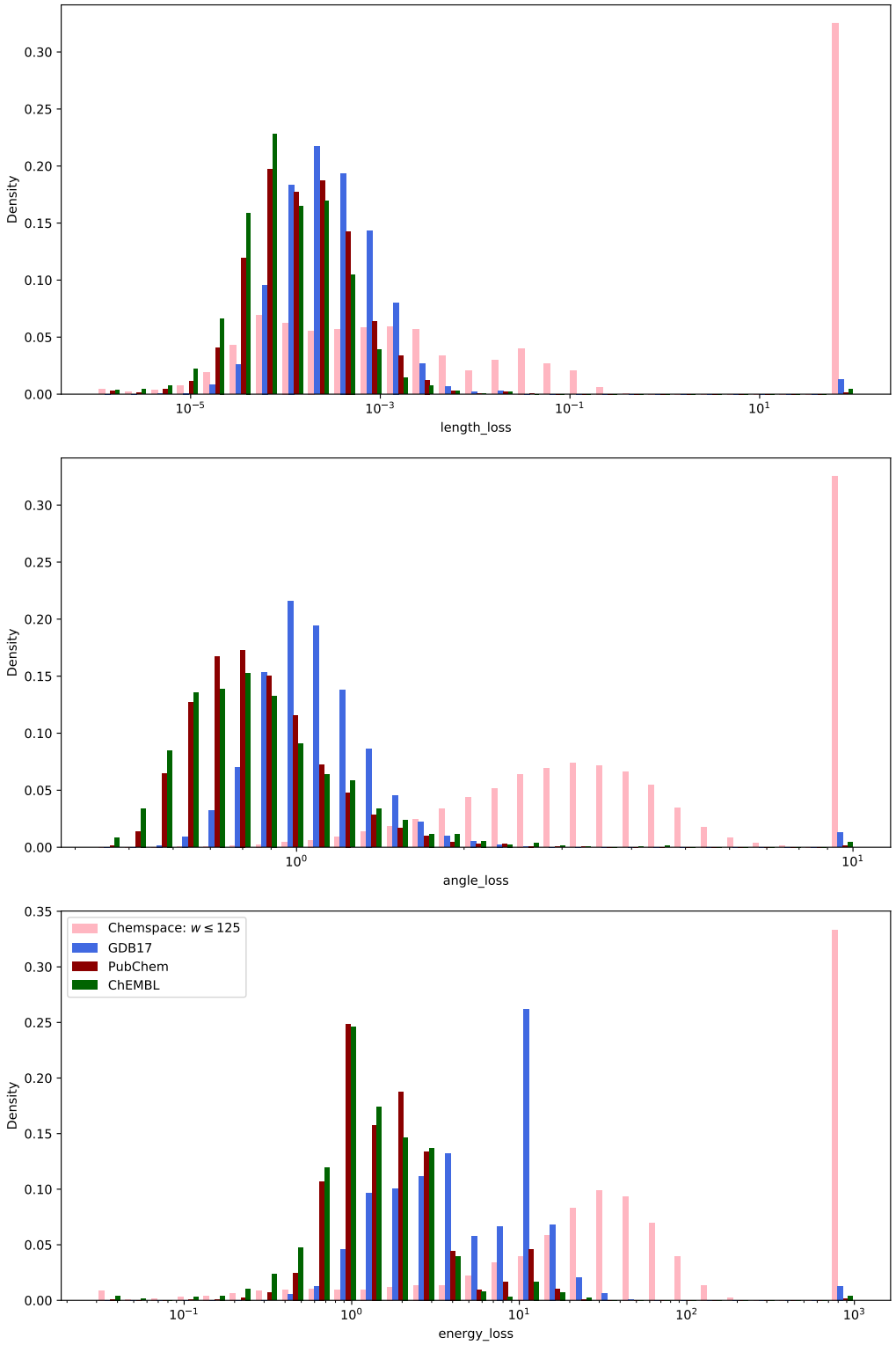

Figure 2: Distributions of the metrics (top: metric 2, middle: metric 3, bottom: metric 4) used, over realistic (PubChem, ChEMBL, GDB17) and combinatorially generated ($MW \leq 125$) molecules. The rightmost bin denotes molecules which has failed ETKDG conformer generation, i.e. metric 1.

are due to violation during bounds smoothing.[28] Structures that fails this criterion will have bond lengths that deviates significantly from the ideal values when embedding into 3D space. We define $L_{ETKDG}(G) = 0 \iff G$ was successful in conformer generation by ETKDG, and $L_{ETKDG}(G) = 1$ otherwise.

As seen in 1 and 2, significant proportions (32.46% in $MW \leq 125$) of CDB molecules failed DG runs, whilst the failure rate is much lower ($\leq 1.12\%$) in RDB molecules.

Metric 2: Average Deviation from Ideal Bond Lengths

The objective function of general distance geometry problems is a natural metric in evaluating the quality of conformers. From Liberti et al,[11] given a conformation of G, $x_u \in \mathbb{R}^3, u \in G$, and typical bond length ranges $[d^L_{uv}, d^U_{uv}], uv \in E(G)$, $L_{lengths}$ is defined as:

$$L_{length}(G, (x_u)_{u \in G}) = \frac{1}{|E|} \sum_{uv \in E} max(\frac{(d^L_{uv})^2 - \|x_u - x_v\|^2}{(d^L_{uv})^2}, 0)^2 + max(\frac{\|x_u - x_v\|^2 - (d^U_{uv})^2}{(d^U_{uv})^2}, 0)^2 \quad (3)$$

That is, L_{length} is the average deviation of each bond lengths from its ideal value.

2 (top) presents the distribution of molecules under L_{length}. The distributions of GDB17 and RDB molecules are largely the same, whilst around half (≈ 45.21) of CDB molecules have conformers with L_{length} values below the 99th percentile of realistic molecule's values (0.011).

Metric 3: Aggregated Deviation from Realistic Bond Angles

The deviation of a conformer's bond angles from 'realistic' values were used as an estimate of angle strain. The distribution of angles on realistic molecule's conformations were used as a reference to evaluate new conformers. For each ETKDG conformation generated from a sample of $89,956$ $CHNOPS$ PubChem and ChEMBL molecules, $1,188,650$ angles (formed by 3 consecutively bonded atoms) were extracted. The angles were grouped by their con-

stituent atoms, the bonding type between the atoms, and the degree of the central atom - in total yielding 378 groups. For each group with ≥ 300 observations, a histogram of the cosine-angles was generated, splitting the $[-1, 1]$ output space into 40 equally sized bins, and a probability mass function was extracted from each histogram. 104 groups had ≥ 300 observations, which includes $1,179,256$, or 99.21% of all angles observed. The 104 extracted histograms are presented in 3.

Given a conformation $Conf(G)$ of G, the angles deviation metric is the average negative log-likelihood for each of its angles ϕ formed, based on the extracted probability distribution p_{X_ϕ} for ϕ's group, X_ϕ.

$$L_{angle}(Conf(G)) = -\frac{1}{N_a}\sum_{\phi} \min(-10, \log p_{X_\phi}(\phi)) \tag{4}$$

where $N_a = \sum_{v \in G} \binom{d(v)}{2}$ is the total number of angles in G. For angles from groups without sufficient observations in the realistic molecules sample, a flat probability of 0.79%, which is the overall likelihood of observing an angle outside of the 104 included groups, was used. To avoid infinity terms when $p_{X_\phi}(\phi) = 0$, each log term was capped at -10.

As seen in 2 (middle), significant differences in distribution can be observed between the realistic and enumerated molecules under this metric. Only a small proportion (≈ 4.84) of the CDB molecules had conformers within 99th percentile (1.53) of RDB molecule's values.

Metric 4: Potential Energy From Universal Force-Field (UFF)[17]

Forcefields define a potential energy surface for arrangements of bonded atoms, and serves as an approximation for the physical soundness of conformers.[14,29] As forcefields typically involve electrostatic, van der Waals, and geometrical (bond, angles, torsional) terms, the previous metrics (bond length / angle based) can be viewed as special cases, correlating to a subset of the terms in the forcefield.

An implementation of UFF, as a part of RDKit, was used for potential energy computa-

tions. An initial conformer was first generated by ETKDG, followed by an optimization step via UFF. The optimized conformer was then evaluated for its total potential energy. From this, $L_{energy}(Conf(G))$ was defined as the mean potential energy over the atoms of G.

L_{energy} over different databases are presented in 2 (bottom). Notably, GDB17, although largely similar to RDB under L_{angle} and L_{length}, had a marked spike of high density at $L_{energy} \approx 10$, which was not observed in RDB molecules.

Extracting Classification Targets For Screening

The screening of combinatorial molecules for physical feasibility can be viewed as classification problem - with classes of 'feasible' and 'infeasible', where only 'feasible' molecules should be kept. A candidate CDB molecule with calculated metric values lying significantly outside those seen in RDB molecules is 'unrealistic' in some aspects. Therefore, it is unlikely that the topology will lead to realistic conformers, and can be marked as 'infeasible'. In 2, we see that RDB (known examples of 'feasible') and CDB (a large proportion of which is expected to be 'infeasible') molecules do indeed take significantly different values under all of the metrics (1-4).

A topology G with $L_{ETKDG}(G) = 1$ (metric 1) will not have values defined for the remainder of the metrics $L_m, m \in \{length, angle, energy\}$, and is marked as 'infeasible'. Otherwise, for metric L_m, and a conformer $Conf(G)$ of G, we define the target y_m as:

$$y_m = 1 \iff L_{ETKDG}(G) = 1 \text{ or } L_m(Conf(G)) \geq r_m \tag{5}$$

where r_m is the value at the 99-th percentile of RDB moleucles under L_m ($r_{length} = 0.01105$, $r_{angle} = 1.53004$, $r_{energy} = 10.55727$). r_m was chosen as a cutoff for CDB molecules as it is unlikely that molecules beyond this value will have realistic conformers. At the same time, a significant (45.20% based on r_{length}, 95.16% based on r_{angle}, 81.72% based on r_{energy}) proportion of CDB molecules can be ruled out.

Generation of Classification Datasets

With both targets and features defined, a classification dataset for screening combinatorial chemical spaces is generated from a small sample of the molecules. For this study, a sample of 1% ($n = 337,924$) was taken from CDB ($MW \leq 125$ molecules). For each molecule in this sample, the topological features and conformer feasibility targets y_m, (where $m \in \{length, angle, energy\}$ were computed. Of the 570 computed topological features, 74 were dropped due to the presence of NaN values in its output; 45 were dropped as all molecules considered mapped to a constant value; 61 were dropped due to high Pearson correlation ($|r| \geq 0.95$) with another feature in order to reduce collinearity: leaving 390 features remaining in the dataset. The processed features, together with the feasibility targets y_{length}, y_{angle} and y_{energy}, led to the generation of three classification datasets.

Note that the comparatively expensive computation of the conformers (relative to computing topological indices) is only required for the molecules in the selected sample. Once a model has been developed, only topological features needs to be computed for prediction.

Model Development and Evaluation

An 80/10/10 training/validation/testing split is used for model development (training, validation) and evaluation (testing). A variety of classification models were used, including logistic regression (LR), multi-layer perceptron / fully-connected neural network (NN), decision tree (DT), random forest (RF), and histogram gradient boosting (HGB). HGB is a gradient boosting method utilising a similar technique used in LightGBM.[30] For each model, a grid search is performed for parameter selection, selecting for parameters with the best combined rankings among AUC-ROC, precision and recall on the validation set.

The classification models are evaluated by their accuracy, precision, recall, AUC-ROC, and the confusion matrix.[31] The classes 'infeasible' and 'feasible' are mapped to 0 and 1 respectively. In the context of molecule screening, precision corresponds to the probability that a molecule classified as 'feasible' is indeed 'feasible'; whilst recall corresponds to the

proportion of 'feasible' molecules correctly classified.

The i, j entry of the confusion matrix is the number of samples in class i, classified as j, where $i, j \in \{0, 1\}$. In addition to the classification accuracy metrics, low prediction complexity (time taken to generate each prediction) is also desirable - as the model will need to be used to generate predictions on billions of molecules.

All models and evaluation metrics were implemented using version 0.23.02 of the `sklearn`[32] library in Python 3.7. All computations were run on a 4 core, 8 thread Ryzen 5 2400G CPU clocked at 3.60GHz. For the parameters chosen by grid search, the training time, prediction time, accuracy, precision, recall and AUC-ROC values are cross validated using `sklearn`'s `KFold` utility to perform randomized 5-fold cross-validation, with `random_state` set to 42.

Results and Discussion

Training and Prediction Complexity

In this study, the time taken to screen $33,846,411$ molecules with $MW \leq 125$ using the proposed hybrid approach was between 29.55 to 35.99 hours on our setup, depending on the machine learning model used. This was the sum of:

1. Enumeration of chemical space via PMG (≈ 550 seconds)
2. ETKDG conformer generation and evaluation on a 1% sample via ETKDG and UFF ($\approx 10,300$ seconds)
3. Computation of 2D features on all molecules (extrapolated from feature computation on the 1% sample: $\approx 105,700$ seconds)
4. Training models on the 1% dataset (37 to $1,820$ seconds, depending on model, see 2)
5. Generating predictions on all remaining molecules (107 to 11210 seconds, depending on model, see 2)

where the majority of time is spent on steps 2 and 3. In comparison, the time required for screening via the 3D-based approach for the same chemical space is expected to take an excess of 286 hours (1,031,100 seconds) on the same setup. The proposed hybrid approach thus represents a speedup of $\geq 7.95\times$ over the 3D-based approach. The speed of our hybrid approach also has the potential for further improvement with a more tailored selection of features, according to situational needs.

[ref use cpu hours instead of raw hours?]

Model Performance

The models' performances on the training sets were presented in 2. The fastest model to train was LR, whilst the fastest to predict was DT. In the L_{length} and L_{energy} datasets, the best performances were from either the neural network or gradient boosting models. In L_{angle}, LR, being a model with relatively few parameters, somewhat surprisingly had the best AUC-ROC and recall scores; whilst GB_{500} had the best accuracy and precision. There were a notable lack of performance improvement between GB_{500} and GB_{3000}, despite the increased limit on the number of weak learners used. This was because the training of GB_{500} and GB_{3000} were automatically stopped early before the limit on weak learners were reached, due to saturating performance improvements on the validation set. As a result, the training time were shorter than on other datasets. The best models were able to correctly classify, respectively, $94.88\%, 98.07\%, 97.45\%$ of the molecules for L_{length}, L_{angle} and L_{energy} derived feasibility. Overall, the L_{length} dataset had best performances based on precision and recall, whilst the L_{energy} dataset had best performance based on the AUC-ROC score. Despite having the best overall accuracy, relative to the other datasets, the L_{angle} dataset had worse performance according to other measures.

Comparison of Metrics

A reason for the performance discrepancy between the results is due to unbalanced classes in the datasets. For many machine learning models, learning on unbalanced datasets is a well known challenge.[33,34] The L_{length} targets was the most balanced, with a roughly even split between 'feasible' and 'infeasible' examples; L_{energy}, and L_{angle} targets were respectively a $4:1$ and $19:1$ split between the two classes. The L_{angle} targets, being the most unbalanced, was also the one with comparatively worst results.

It appears that L_{angle}, despite of being a seemingly highly discriminating metric between CDB and RDB molecules, presented a challenge when attempting to derive a model for chemical space screening. At the same time, L_{angle} may overestimate the strain in non-RDB molecules, leading to targets that were more unbalanced. As the angles profiles were extracted from an empirical distribution of RDB (PubChem and ChEMBL) molecules, the calculated angle strain of molecules from other databases may be biased towards higher values. This is due to intrinsic differences between the molecules of different databases (see fig. 4 of Ruddigkeit et al[4]). This effect is observed in 2, where GDB17 molecules, despite selected for topological characteristics leading to low angle strain (e.g. small rings, aromaticity and bridgehead filters),[4] still had higher L_{angle} values compared to PubChem and ChEMBL molecules.

As L_{energy} corresponds to the overall potential energy of the molecule's conformers from a forcefield computation, where terms such as angle and bond potential energy were included as a component, the L_{energy} metric represents a more robust measure for the overall strain of the molecule. The other metrics, L_{angle} and L_{bond}, will correlate highly with a term in the forcefield equation. Overall, L_{energy} targets ruled out fewer CDB molecules than L_{angle} targets, which is counter intuitive. In addition to the aforementioned possible bias in the L_{angle} metric, it may require further investigation to see whether other factors also play a role.

Table 2: Performance of models on classification datasets of the $MW \leq 125$ molecules. Here LR = logistic regression, NN = neural network (multi-layer perceptron), DT = decision tree, RF = random forest, GB = gradient boosting (histogram-based). The subscript x on GB_x denotes the limit on the number of weak learners used. Training time is in terms of seconds. Prediction time is in terms of seconds per 10,000 molecules. Reference times are on a 3.60GHz AMD Ryzen 2400G CPU.

	LR	NN	DT	RF	GB_{500}	GB_{3000}
Accuracy	90.24%	94.01%	91.75%	91.73%	94.29%	**94.88%**
Precision	90.79%	94.69%	93.09%	92.99%	95.14%	**95.60%**
Recall	91.47%	94.36%	91.76%	91.82%	94.39%	**95.03%**
AUC-ROC	0.9011	0.9397	0.9175	0.9172	0.9428	**0.9487**
Training time	290.51	631.76	**49.07**	287.04	514.50	1820.42
Prediction time	**0.0378**	0.0987	0.0417	0.3745	1.4003	3.3173

(a) Performance on the L_{length} derived targets.

	LR	NN	DT	RF	GB_{500}	GB_{3000}
Accuracy	97.51%	98.08%	96.86%	97.27%	**98.16%**	98.10%
Precision	72.05%	83.41%	65.93%	71.12%	**85.40%**	84.63%
Recall	**79.36%**	75.48%	72.53%	73.48%	74.74%	74.18%
AUC-ROC	**0.8890**	0.8736	0.8531	0.8598	0.8705	0.8675
Training time	299.69	558.11	**37.28**	213.15	360.51	339.57
Prediction time	**0.0316**	0.1461	0.0348	0.2689	1.0013	0.8891

(b) Performance on the L_{angle} derived targets.

	LR	NN	DT	RF	GB_{500}	GB_{3000}
Accuracy	94.91%	96.96%	95.22%	94.97%	97.18%	**97.52%**
Precision	87.71%	91.92%	87.99%	84.34%	93.48%	**94.37%**
Recall	83.90%	91.39%	85.48%	89.02%	90.93%	**91.91%**
AUC-ROC	0.9063	0.9480	0.9144	0.9266	0.9476	**0.9534**
Training time	314.75	830.63	**37.48**	619.72	509.78	1606.85
Prediction time	**0.0365**	0.1425	0.0418	0.6545	1.4577	3.1794

(c) Performance on the L_{energy} derived targets.

Precision-Recall Tradeoff

One way to overcome the problem of unbalanced datasets is via under-sampling on the more populous class.[33,34] For the most unbalanced L_{angle} targets, we performed between $\frac{1}{10}$ to $\frac{1}{2}$ under-sampling. The results on the best models (gradient boosting with up to 500 weak learners, and logistic regression) were reported in 3. We note that there was a tradeoff between precision and recall, where higher recall and lower precision values corresponded with increasing under-sampling. For molecule screening, a tradeoff for a high-recall model might be useful with highly unbalanced targets. When the target class (feasible molecules) is the minority one, a high recall rate ($\approx 95\%$) ensures that the majority of 'feasible' molecules are kept. At the same time, a modest precision ($\approx 50\%$) score means the vast majority of 'infeasible' molecules are ruled out - leaving behind a much smaller set of molecules from which further pruning can be performed.

Table 3: Results of the 2 best L_{angle} models trained on dataset with (0.1 to 0.5) undersampling on the most populous class.

	0.1	0.2	0.3	0.4	0.5
Accuracy	95.70%	97.06%	97.66%	97.92%	98.05%
Precision	52.55%	63.09%	69.93%	74.32%	77.36%
Recall	95.25%	91.54%	88.56%	85.63%	83.06%
AUC-ROC	0.9549	0.9444	0.9333	0.9208	0.9092

(a) Performance of GB_{500} on the testing set, training on undersampled L_{angle} dataset.

	0.1	0.2	0.3	0.4	0.5
Accuracy	95.06%	96.43%	97.01%	97.31%	97.46%
Precision	48.87%	58.00%	63.61%	67.71%	70.59%
Recall	94.18%	89.33%	85.86%	82.54%	79.44%
AUC-ROC	0.9464	0.9305	0.9171	0.9029	0.8890

(b) Performance of LR on the testing set, training on undersampled L_{angle} dataset.

Comparison with GDB

GDB11, GDB13, and GDB17 uses a combination of conformer based and topological based pruning rules to select realistic molecules from the combinatorial graphs. Of the real molecule

sets considered (RDB = PubChem and ChEMBL), around 80% of RDB molecules (within the atom counts and types considered in GDB17's enumeration) were compatible with GDB17's enumeration rules.[4] This percentage approximately corresponds to the 'recall rate' GDB17's enumeration rules in this study.

For molecules chosen by our method, the RDB inclusion rate was dependent on 1) the atomic compositions used to generate the initial combinatorial molecule set, 2) the choice of classification targets, and 3) the recall rate of the model used for prediction - each steps are potential places where 'feasible' molecules can escape our selection rule. We have seen in 2 and 2 that ETKDG was able to generate at least one conformer for RDB molecules $\geq 98.88\%$ of the times ($\leq 1.12\%$ of RDB molecules lost). Moreover, we chose a threshold metric value of 99% out of the molecules with a conformer generated to form our classification targets. This altogether led to $\leq 2.11\%$ of RDB molecules being incorrectly classified on the training set. The recall rate of the models is the ability of them to recreate the training set's classification, where 91.87% was achieved by the best model for the more robust L_{energy} targets. Overall, this led to the inclusion of around 90% of PubChem and ChEMBL molecules (with atomic composition CHNOPS) by our screening method. This was higher than GDB17's recall rate, translating to the inclusion of more feasible molecules as well as molecules covering a broader spectrum of the 'physically realistic' chemical universe.

Of the molecules selected by GDB17's enumeration rules, a tiny proportion (0.6%) of molecules failed to have any conformers generated by ETKDG - which was in between the rates observed in ChEMBL and PubChem. 5.82% of the $MW \leq 125$ CHNOPS molecules classified by our best machine learning model for L_{energy} had values greater than the cut-off, r_{energy}. In comparison, 9.76% of GDB17 molecule conformers had $L_{energy} \geq r_{energy} = 10.55727$, a value rarely observed in PubChem (1.00%) and ChEMBL molecules (0.33%) (see 2). This may suggest that GDB17's enumeration rules 'leaks' more potentially unrealistic molecules.

Strengths and Limitations

Our hybrid approach to chemical space screening is flexible, learning the pruning rules for the dataset on the go, and does not rely on any particular chemical properties of the dataset. The point at which chemical knowledge is involved is in the design of evaluation methods of the conformer's quality. Due to the small training sets used for the machine learning models, the evaluation metric is not heavily constrained by computation complexity, allowing room for high-quality evaluations (e.g. forcefields, or quantum simulations) to be computed without substantial impact on computation speed.

It should be noted that the overall accuracy of the method is dependent on the topological feature encoding used for each molecule, the quality of the conformer evaluation methods, and the learnability of the training sets generated. Care thus needs be taken to ensure that the feature vector encodes sufficient information about the topology to allow for accurate predictions, whilst not including unnecessary information, increasing computational demands. Since the model's accuracy can only be as good as that of the classifications targets, it is important to select high-quality conformer evaluation metrics (as seen in the comparison between $L_{length}, L_{angle}, L_{energy}$) which will effectively distinguish between feasible and infeasible conformers.

Conclusion

We have developed a new automated approach for fast and high-quality screening of small molecules. This approach, being $\approx 8\times$ faster and without significantly sacrificing accuracy, can be used as an extension of 3D based screening methods. From our investigations, we found that the neural network model provided a good balance between screening accuracy and training/prediction complexity. Moreover, the UFF-based L_{energy} metric appeared to be the most informative of the conformer evaluation metrics we tested. Compared to 2D-based pruning rules, this approach is more accurate, with better coverage of known feasible

molecules. As opposed to 2D pruning rules used for enumeration of combinatorial chemical spaces, our approach instead learns these pruning rules automatically based on 3D calculations of molecular feasibility. Once the topological features and conformer evaluation methods are established, the process can be fully automated, without significant expert chemistry knowledge.

Implementation & Availability

Our modified OMG, which adds the ability to enumerate chemical spaces and isomers under crowding constraints, is available on GitHub under the GNU AGPL v3 license. The code is based on the concurrent version of the OMG.[21] All Python scripts and Jupyter Notebooks used for the machine learning pipeline and analysis is also available on GitHub.

References

(1) Faradzev, I. Constructive enumeration of combinatorial objects. **1978**,

(2) Iványi, A. Degree sequences of multigraphs. ANNALES UNIVERSITATIS SCIENTIARUM BUDAPESTINENSIS DE ROLANDO EOTVOS NOMINATAE SECTIO COMPUTATORICA. 2012; pp 195–214.

(3) Faulon, J. Enumerating molecules. **2005**,

(4) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling* **2012**, *52*, 2864–2875.

(5) Meringer, M. *Handbook of chemoinformatics algorithms*; Chapman and Hall/CRC, 2010; pp 245–280.

(6) Reymond, J.-L.; Blum, L. C.; van Deursen, R. Exploring the chemical space of known and unknown organic small molecules at www. gdb. unibe. ch. *CHIMIA International Journal for Chemistry* **2011**, *65*, 863–867.

(7) Meringer, M.; Cleaves, H. J. Exploring astrobiology using in silico molecular structure generation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2017**, *375*, 20160344.

(8) Cleaves, H. J.; Meringer, M.; Goodwin, J. 227 Views of RNA: Is RNA unique in its chemical isomer space? *Astrobiology* **2015**, *15*, 538–558.

(9) Rücker, C.; Meringer, M. How Many Organic Compounds are Graph–Theoretically Nonplanar.

(10) Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. *Advances in mathematical chemistry and applications*; Elsevier, 2015; pp 113–138.

(11) Liberti, L.; Lavor, C.; Maculan, N.; Mucherino, A. Euclidean distance geometry and applications. *SIAM review* **2014**, *56*, 3–69.

(12) Schwab, C. H. Conformations and 3D pharmacophore searching. *Drug Discovery Today: Technologies* **2010**, *7*, e245–e253.

(13) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely available conformer generation methods: how good are they? *Journal of chemical information and modeling* **2012**, *52*, 1146–1158.

(14) Faulon, J.-L. Stochastic generator of chemical structure. 1. Application to the structure elucidation of large molecules. *Journal of Chemical Information and Computer Sciences* **1994**, *34*, 1204–1218.

(15) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B., et al. PubChem 2019 update: improved access to chemical data. *Nucleic acids research* **2019**, *47*, D1102–D1109.

(16) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B., et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, *40*, D1100–D1107.

(17) Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **1992**, *114*, 10024–10035.

(18) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *Journal of the American Chemical Society* **1985**, *107*, 3902–3909.

(19) Riniker, S.; Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *Journal of chemical information and modeling* **2015**, *55*, 2562–2574.

(20) Landrum, G. RDKit: Open-source cheminformatics.

(21) Jaghoori, M. M.; Jongmans, S.-S. T.; Boer, F. d.; Peironcely, J. E.; Faulon, J.-L.; Reijmers, T. H.; Hankemeier, T., et al. PMG: multi-core metabolite identification. *Electronic Notes in Theoretical Computer Science* **2013**, *299*, 8.

(22) Peironcely, J. E.; Rojas-Chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. OMG: open molecule generator. *Journal of cheminformatics* **2012**, *4*, 21.

(23) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002**, *42*, 1273–1280.

(24) Todeschini, R.; Consonni, V. *Handbook of Chemoinformatics*; John Wiley & Sons, Ltd, 2003; Chapter VIII.2, pp 1004–1033.

(25) Wiener, H. Structural determination of paraffin boiling points. *Journal of the American chemical society* **1947**, *69*, 17–20.

(26) Trinajstic, N. *Chemical graph theory*; Routledge, 2018.

(27) Kufareva, I.; Abagyan, R. *Homology Modeling*; Springer, 2011; pp 231–257.

(28) Crippen, G. M.; Havel, T. F., et al. *Distance geometry and molecular conformation*; Research Studies Press Taunton, 1988; Vol. 74.

(29) Nalin de Silva, K.; Goodman, J. M. What is the smallest saturated acyclic alkane that cannot be made? *Journal of chemical information and modeling* **2005**, *45*, 81–87.

(30) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems. 2017; pp 3146–3154.

(31) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; http://www.deeplearningbook.org.

(32) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(33) Haixiang, G.; Yijing, L.; Shang, J.; Mingyun, G.; Yuanyue, H.; Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* **2017**, *73*, 220–239.

(34) Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **2017**, *18*, 1–5.

Supplementary Information

Angle Profiles Extracted From RDB Molecules

Table of Contents Graphic

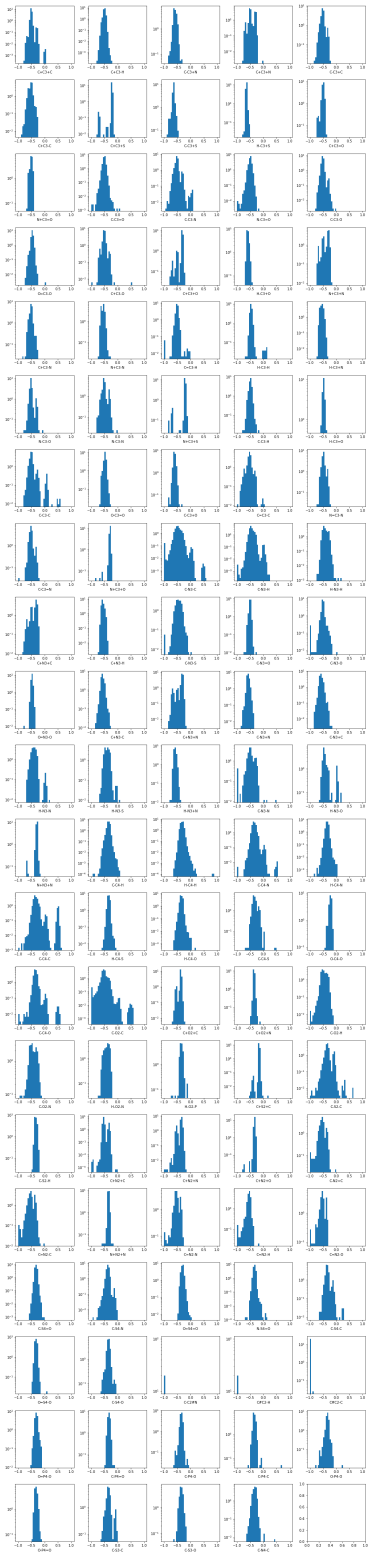

Figure 3: The histograms of the observed cosine-angles, for the 104 groups where a probability mass function is extracted. The groups are labelled according to the format $A_x b_{xc} A_c d_c b_{cy} A_y$ - where A_x, A_c, A_y are the 3 constituent atoms involved in the angle, with central atom A_c; d_c is the degree of A_c; b_{xc}, b_{cy} are the 2 bonds forming the angle, with $-, +, =, \#$ denoting single, aromatic, double, and triple bonding, respectively.

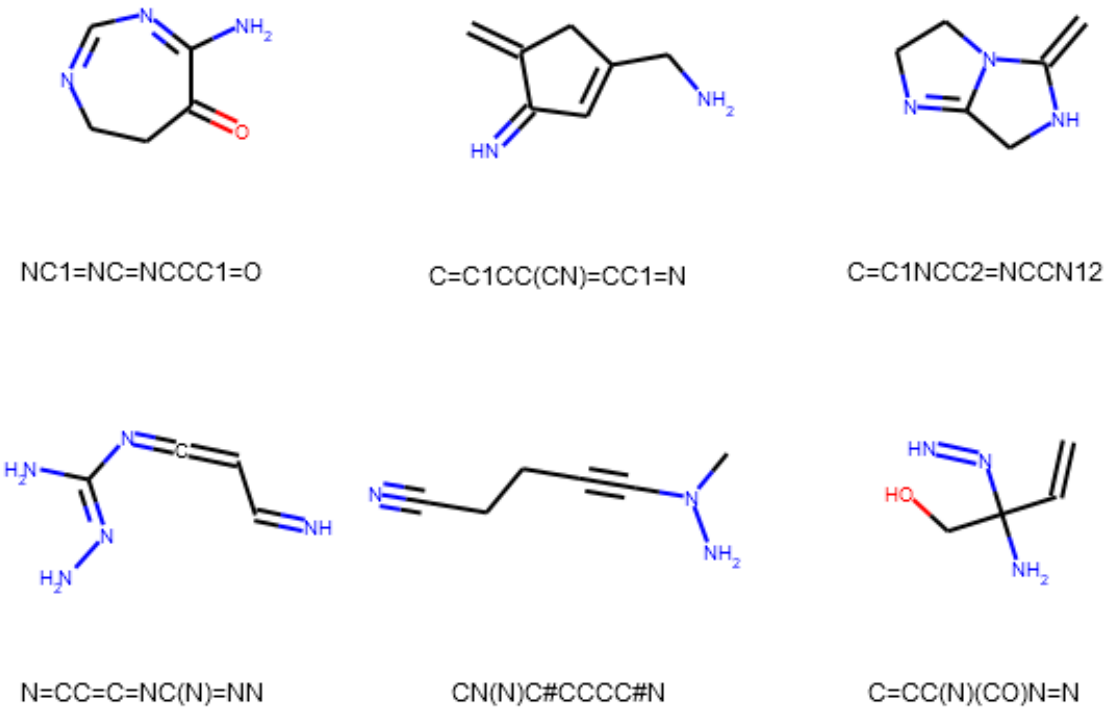

Figure 4: Table of contents graphic.