Supporting Information

One Among Millions: The Chemical Space of Nucleic Acid-like Molecules

Henderson James Cleaves II[1,2,3*†], Christopher Butch[1,3,4†], Pieter Buys Burger[4], Jay Goodwin[4],

Markus Meringer[5†]

1. Earth-Life Science Institute, Tokyo Institute of Technology, 2-12-IE-1 Ookayama, Meguro-ku, Tokyo 152-8551, Japan

2. Institute for Advanced Study, Princeton, NJ 08540

3. Blue Marble Space Institute for Science, 2800 Woodley Rd. NW, Washington, DC 20008

4. Department of Chemistry, Emory University, 1515 Dickey Dr, Atlanta, GA 30322 USA

5. German Aerospace Center (DLR), Earth Observation Center (EOC), Münchner Straße 20, 82234 Oberpfaffenhofen-Wessling, Germany

*To whom correspondence should be addressed, Email: henderson.cleaves@gmail.com

†HJC, MM and CB contributed equally to this work

Additional files include:

1. The “Badlist” used for structure enumeration: BadAaNucList.sdf
2. .txt files containing the generated structures in SMILES format
3. Post-processing substructures NucPostProc.pdf
4. High-resolution heatmaps of the most common substructures for CHO and CHON sets

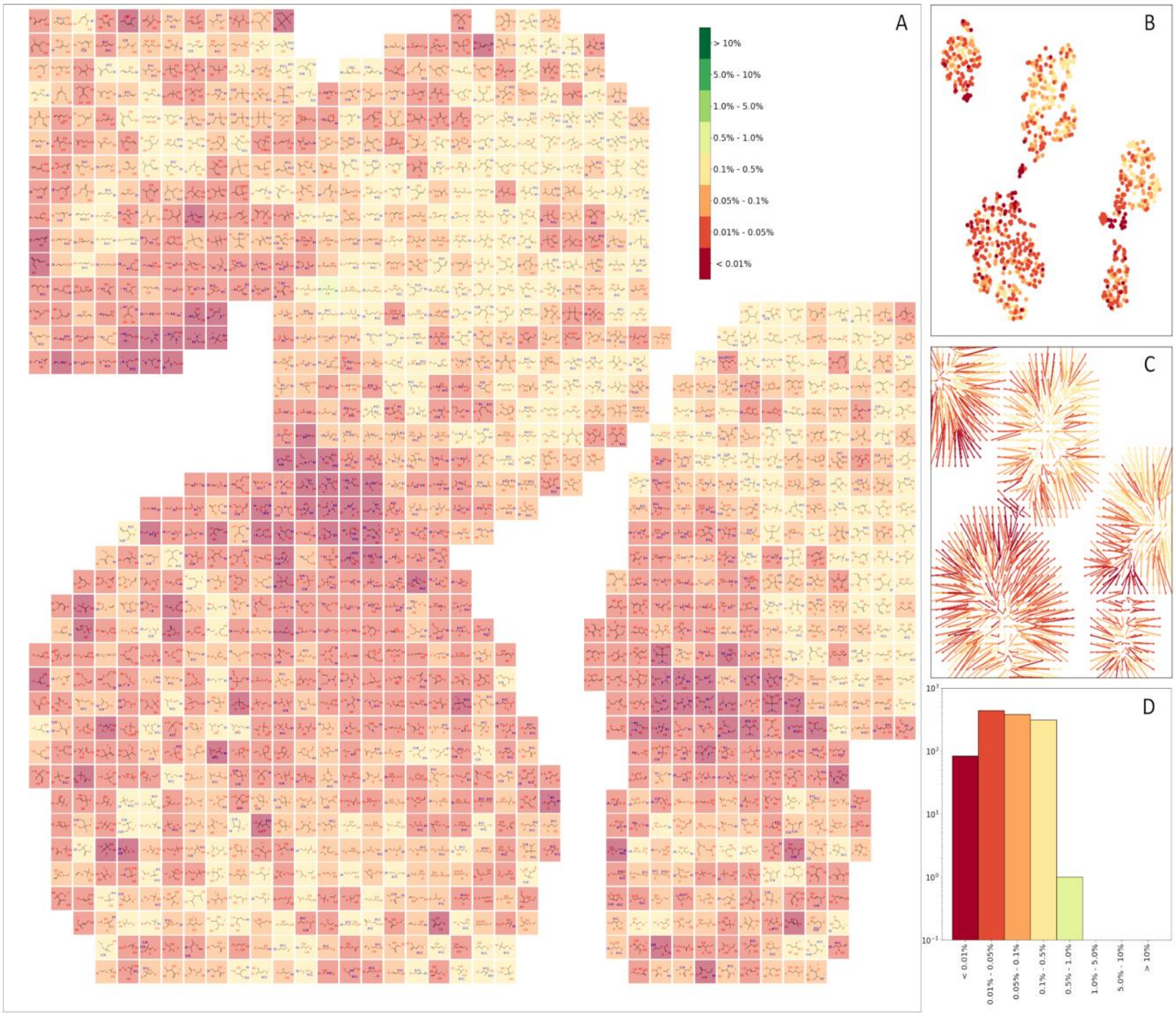

Figure S1. Clustering of acyclic core structures as described in Figure 4. Before stereoisomerization, acyclic backbones represent 48.5% of the generated scaffold set. Backbone similarities were calculated using Tanimoto distance based on extended connectivity fingerprints of length four. Unlike the Bemis-Murcko scaffolds, the number of possible clusters generated for acyclic scaffolds is arbitrary. Therefore, for parity to the 1225 Bemis-Murcko clusters 1225 clusters of acyclic scaffolds were generated. The cluster centers were then reduced to a two-dimensional mapping using t-distributed stochastic neighbor embedding based on the Tanimoto Similarities. The two dimensional mapping was then aligned to a uniform grid using the Jonker-Volgenant Algorithm to improve readability while minimally disturbing group associations (original grouping and JVA mapping depicted in insert). This view demonstrates that unlike the cyclic species wherein several clusters represent 1-15% of the population, acyclic cluster populations are much more homogeneous with no populations greater than 1%.

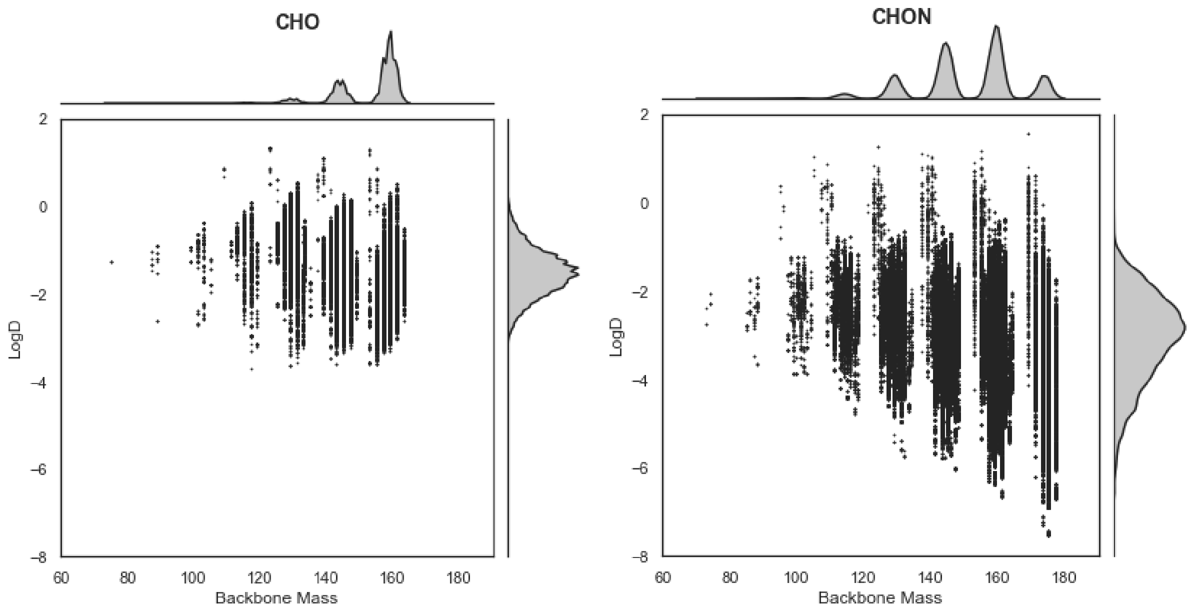

Figure S2: Predicted LogD values of the adenylated constituents of the CHO and CHON libraries plotted against backbone mass. The CHO backbones exist in a more narrow, more water insoluble range, while the CHON library tends toward greater water solubility due to the contributions of potentially ionizable substructures.

Table S1. Numbers of isomers by molecular formula in the CHO library.

C3H5O3B	3
C3H7O2B	2
C4H5O3B	5
C4H5O4B	19
C4H7O2B	4
C4H7O3B	25
C4H7O4B	15
C4H9O2B	9
C4H9O3B	9
C5H5O3B	3
C5H5O4B	50
C5H7O2B	5
C5H7O3B	84
C5H7O4B	225
C5H9O2B	38
C5H9O3B	169
C5H9O4B	194
C5H11O2B	35
C5H11O3B	65
C5H11O4B	20
C6H5O2B	6
C6H5O3B	14

C6H5O4B	80
C6H7O2B	16
C6H7O3B	180
C6H7O4B	989
C6H9O2B	107
C6H9O3B	923
C6H9O4B	2332
C6H11O2B	225
C6H11O3B	1060
C6H11O4B	1786
C6H13O2B	120
C6H13O3B	349
C6H13O4B	272
C7H5O3B	33
C7H5O4B	192
C7H7O2B	37
C7H7O3B	531
C7H7O4B	3013
C7H9O2B	344
C7H9O3B	3401
C7H9O4B	14488
C7H11O2B	1031
C7H11O3B	7761
C7H11O4B	21690
C7H13O2B	1097
C7H13O3B	5930
C7H13O4B	13007
C7H15O2B	396
C7H15O3B	1559
C7H15O4B	2059

Table S2. Numbers of isomers by molecular formula and in the CHNO library.

C3H6NO2B	4
C3H7N2OB	1
C3H8NOB	3
C3H9N2B	2
C4H5N2O4B	1
C4H6NO2B	5
C4H6NO3B	22
C4H6NO4B	15
C4H7N2O2B	4
C4H7N2O3B	10
C4H8NOB	8
C4H8NO2B	41

$C_4H_8NO_3B$	28
$C_4H_9N_2OB$	16
$C_4H_9N_2O_2B$	21
$C_4H_9N_2B$	5
$C_4H_{10}NOB$	18
$C_4H_{10}NO_2B$	17
$C_4H_{11}N_2OB$	14
$C_4H_{11}N_2B$	11
$C_5H_5N_2OB$	9
$C_5H_5N_2O_2B$	24
$C_5H_5N_2O_3B$	22
$C_5H_5N_2O_4B$	11
C_5H_6NOB	3
$C_5H_6NO_2B$	9
$C_5H_6NO_3B$	43
$C_5H_6NO_4B$	148
$C_5H_7N_2OB$	3
$C_5H_7N_2O_2B$	12
$C_5H_7N_2O_3B$	70
$C_5H_7N_2O_4B$	84
$C_5H_7N_2B$	6
C_5H_8NOB	15
$C_5H_8NO_2B$	147
$C_5H_8NO_3B$	379
$C_5H_8NO_4B$	375
$C_5H_9N_2OB$	60
$C_5H_9N_2O_2B$	197
$C_5H_9N_2O_3B$	287
$C_5H_9N_2O_4B$	97
$C_5H_9N_2B$	12
$C_5H_{10}NOB$	86
$C_5H_{10}NO_2B$	335
$C_5H_{10}NO_3B$	427
$C_5H_{10}NO_4B$	114
$C_5H_{11}N_2OB$	179
$C_5H_{11}N_2O_2B$	342
$C_5H_{11}N_2O_3B$	137
$C_5H_{11}N_2B$	59
$C_5H_{12}NOB$	73
$C_5H_{12}NO_2B$	151
$C_5H_{12}NO_3B$	52
$C_5H_{13}N_2OB$	125
$C_5H_{13}N_2O_2B$	63
$C_5H_{13}N_2B$	44

```
C6H5N2OB 2
C6H5N2O2B 78
C6H5N2O3B 290
C6H5N2O4B 282
C6H5N2B 3
C6H6NOB 10
C6H6NO2B 120
C6H6NO3B 231
C6H6NO4B 519
C6H7N2OB 147
C6H7N2O2B 321
C6H7N2O3B 567
C6H7N2O4B 1165
C6H7N2B 12
C6H8NOB 62
C6H8NO2B 450
C6H8NO3B 1782
C6H8NO4B 4103
C6H9N2OB 272
C6H9N2O2B 1100
C6H9N2O3B 3329
C6H9N2O4B 3824
C6H9N2B 66
C6H10NOB 315
C6H10NO2B 2006
C6H10NO3B 5140
C6H10NO4B 6606
C6H11N2OB 1174
C6H11N2O2B 3842
C6H11N2O3B 5886
C6H11N2O4B 3528
C6H11N2B 240
C6H12NOB 554
C6H12NO2B 2428
C6H12NO3B 4326
C6H12NO4B 2601
C6H13N2OB 1668
C6H13N2O2B 3698
C6H13N2O3B 3037
C6H13N2O4B 619
C6H13N2B 380
C6H14NOB 270
C6H14NO2B 873
C6H14NO3B 761
```

C6H14NO4B 155
C6H15N2OB 749
C6H15N2O2B 863
C6H15N2O3B 245
C6H15N2B 164

Table S3. Hits within Reaxys and PubChem and overlap with the CHO and CHON libraries without taking stereochemistry into account.

	Reaxys				PubChem			
	Core without N		Core with N		Core without N		Core with N	
Base	Hits	Overlap (%)	Hits	Overlap (%)	Hits	Overlap (%)	Hits	Overlap (%)
A	725	333	223	106	795	325	338	80
C	346	168	99	37	463	195	145	32
G	353	152	75	34	963	182	230	34
T	570	231	246	68	578	201	350	59
U	485	194	151	56	492	179	313	55
Sum	2479	1078	794	301	3291	1082	1376	260

Table S4. Hits within Reaxys and PubChem and overlap with CHO and CHON libraries stereochemistry taken into account.

	Reaxys				PubChem			
	Core without N		Core with N		Core without N		Core with N	
Base	Hits	Overlap (%)	Hits	Overlap (%)	Hits	Overlap (%)	Hits	Overlap (%)
A	1463	571	367	146	1909	441	606	105
C	658	260	154	46	1202	260	282	29
G	583	200	99	41	1386	216	275	35

T	1003	337	316	90	1069	259	529	78
U	809	307	217	73	927	242	414	54

Commands used for operating MOLGEN 5

These are the two commands used for structure generation with MOLGEN, executed in the Windows PowerShell:

```
Measure-Command{C:/Programs/Molgen5.0/mgen.exe C2-7H5-15O[h=0]0-2O[h=1]2-4Cl
-sum O=2-4 -badlist ../sdf/BadAaNucList.sdf -badlist ../sdf/BadCl-Het.sdf -
badlist ../sdf/BadAromaticsList.sdf -badlist ../sdf/BadRingList.sdf -ringsize
5-10 -maxbond 2 -v -o C2-7Ox.mb4 2> C2-7Ox.number.txt} > C2-7Ox.time.txt
```

```
Measure-Command{C:/Programs/Molgen5.0/mgen.exe  C1-6H5-15N[h=0]0-2N[h=1]0-
2N[h=2]0-2O[h=0]0-4O[h=1]0-4Cl -sum N[h=1]+N[h=2]+O[h=1]=2-6 -sum N=1-2 -sum
O=0-4 -badlist ../sdf/BadAaNucList.sdf -badlist ../sdf/BadCl-Het.sdf -badlist
../sdf/BadAromaticsList.sdf -badlist ../sdf/BadRingList.sdf -ringsize 5-10 -
maxbond 2 -v -o C1-6NxOy.mb4 2> C1-6NxOy.number.txt} > C1-6NxOy.time.txt
```

The first command is for generating the CHO library, the second for the CHNO library. Measure-Command is a PowerShell command to measure the time required for the execution of a program. Its output is redirected to `C2-7Ox.time.txt` and `C1-6NxOy.time.txt`. `C:/Programs/Molgen5.0/mgen.exe` is the path to the MOLGEN 5 executable.

`C2-7H5-15O[h=0]0-2O[h=1]2-4Cl` is the fuzzy molecular formula used for the CHO library. It defines for each chemical element involved a range for the number of atoms, e.g. 2-7 carbon atoms, 5-15 hydrogen atoms or exactly one chlorine atom, which represents the nucleobase. For the oxygen atoms there is additionally the number of adjacent hydrogen atoms specified. We define a range of 0-2 oxygen atoms without adjacent hydrogen atoms and 2-4 oxygen atoms with one adjacent H atom. This way we get at least two OH, which are required as attachment points. The next arguments, `-sum O=2-4` , tell MOLGEN that the total number of oxygen atoms must be in the range 2-4.

C1-6H5-15N[h=0]0-2N[h=1]0-2N[h=2]0-2O[h=0]0-4O[h=1]0-4Cl is the fuzzy formula for the CHNO library. The next arguments, -sum N[h=1]+N[h=2]+O[h=1]=2-6, guarantee that we have at least two attachment points, which may be NH, NH_2 or OH.

Option -badlist points MOLGEN to a list of forbidden substructures given as SDfile. Here four such 'badlists' are used:

- BadAaNucList.sdf originated from previous projects on libraries of amino acids (1) and ribose isomers (2), and has been extended during this study for general nucleoside analogs. The SDfile is part if this SI.
- BadCl-Het.sdf has just two entries, N-Cl and O-Cl, which prohibit the nucleobase being adjacent to a hetero atom.
- BadAromaticsList.sdf contains 'bad' bridged aromatic substructures,
- BadRingList.sdf consists of 'bad' cyclic and unsaturated substructures.

The latter two files are part of the MOLGEN 5 delivery, the substructures are depicted in (3) and in the software user manual of MOLGEN 5, which is provided at http://molgen.de/documents/manual_molgen50.pdf.

The command line parameters -ringsize 5-10 constrain the length of rings to the range from five to ten. The lower limit excludes 3- and 4-rings, the upper limit is just an arbitrary threshold which cannot be exceeded due to the number of atoms available to be part of rings. Parameters -maxbond 2 set the maximum bond multiplicity to 2, *i.e.*, no triple bonds are allowed. Option -v enables verbosity on the lowest level, *i.e.*, MOLGEN writes a summary of the input and the number of generated structures to the standard output, which is redirected to a text file named C2-7Ox.number.txt. Parameters -o C2-7Ox.mb4 tell the program to write the generated structures to a file named C2-7Ox.mb4. The file type 'mb4' denotes a binary format (Molgen Binary 4) for storing chemical structures, specified in (4). This binary format enables faster writing and reading compared to text-based formats, which is particularly useful for transferring the generated structures to the post-processing step.

Substructures used for the post-processing step

Some undesired and required structural motifs could not be formulated in the syntax available for MOLGEN 5 substructures. These were classified into five groups:

- substructures that must not occur,
- substructures that must not occur more than once,
- substructures that must not occur after MOLGEN-QSPR aromaticity perception,
- substructures that must not occur more than once after MOLGEN-QSPR aromaticity perception,
- substructures contributing to the number of attachment points, which must be at least two.

Graphical representations of these substructures are depicted in NucPostProc.pdf, which is part of the SI. Such substructures with so-called substructure restrictions (4) can be handled by MOLGEN-QSPR (5). For the post-processing step a command line interface to MOLGEN-QSPR was used.

References

1. Meringer, M.; Cleaves, H.; Freeland, S., Beyond Terrestrial Biology: Charting the Chemical Universe of α-Amino Acid Structures. *J. Chem. Inf. Mod.* **2013**, *53*, 2851-2862.

2. Cleaves, H.; Meringer, M.; Goodwin, J, 227 Views of RNA: Is RNA Unique in Its Chemical Isomer Space? *Astrobiology* **2015**, *15*, 538-558.

3. Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M; Rücker, C.; Wassermann, A., MOLGEN 5.0, a Molecular Structure Generator. In *Advances in Mathematical Chemistry and Applications: Revised Edition* **2016**, *Vol. 1*, 113-138, Elsevier.

4. Kerber, A.; Laue, R.; Meringer, M.; Rücker, C.; Schymanski, E., *Mathematical Chemistry and Chemoinformatics: Structure Generation, Elucidation and Quantitative Structure-Property Relationships.* **2013**. Walter de Gruyter.

5. Kerber, A.; Laue, R.; Meringer, M.; Rücker, C., MOLGEN-QSPR, a Software Package for the Study of Quantitative Structure-Property Relationships. *MATCH Commun. Math. Comput. Chem.* **2004**, *51*, 187-204.