

Chemical Space Mimicry for Drug Discovery

William Yuan,^{§†} Dadi Jiang,[†] Brandon Turner,[†] Quynh-The Le,[†] Robert Tibshirani,[†] Purvesh Khatri,[†] Mark G. Moloney,[‡] and Albert C. Koong[†]

[§] Trinity College, University of Oxford, Oxford, OX1 3BH, United Kingdom

[†] Stanford University School of Medicine, Stanford, CA 94305, United States

[‡] Chemistry Research Laboratory, University of Oxford, Oxford, OX1 3TA, United Kingdom

ABSTRACT: We describe a new methodology, Machine-based Identification of Molecules Inside Characterized Space (MIMICS) that generates chemical libraries inspired by a text-based input. MIMICS-generated libraries were found to preserve distributions of chemical properties while simultaneously increasing structural diversity. Newly identified MIMICS-generated compounds were found to be bioactive as inhibitors of the unfolded protein response in cell-based assays, confirming the applicability of MIMICS towards drug discovery. Wider application of MIMICS could facilitate efficient utilization of chemical space.

Effective enumeration of unknown and novel compounds has the potential to change the way discovery of new molecular entities is pursued. In the regime of drug design, these types of compounds can be used to populate libraries, providing an effective starting point for the identification of new leads and motifs. In particular, Vishrup and Rupakheti^{1,2} describe an iterative method to enumerate compounds over all of chemical space in a way that maximizes

1
2
3 structural diversity, and demonstrates the potential of this approach towards drug design
4
5 applications.
6
7

8
9 We show that novel compounds can be generated in a facile manner with minimal a priori
10 information and that compounds generated in this way can function in a bioactive manner. This
11 approach, called Machine-based Identification of Molecules Inside Characterized Space
12 (MIMICS), considers both the properties of a set of molecules rather than an individual molecule
13 and generates an inspired set with both increased structural diversity and chemical novelty. The
14 structures of the reference set are not needed for molecule generation, and instead only a partial
15 text-based representation is used for reference. Additionally, the particular physical property for
16 optimization does not need to be known: MIMICS can preserve multiple descriptors despite
17 limited initial information.
18
19
20
21
22
23
24
25
26
27
28
29
30

31 The simplified molecular input line entry system (SMILES) is used to encode molecules in a
32 linear, text-based format for use in MIMICS. SMILES lacks implicit hydrogens and
33 interpretation of SMILES strings as complete structures require the use of outside algorithms³.
34
35
36
37
38 The starting input information available to MIMICS is thus necessarily incomplete.
39
40
41

42 The creation of a set of molecules requires only two steps: character generation and filtration.
43 First, SMILES strings from an enumerated input set of molecules, whose physical properties
44 inform the resultant properties of the MIMICS molecules generated, are used to generate a
45 section of text. This is done using the character level Recurrent Neural Network⁴ (char-RNN),
46 freely available software that generates context-independent text based on analysis of character
47 sequences from an input. The characters of this generated text take the form of SMILES-encoded
48 molecules.
49
50
51
52
53
54
55
56
57
58
59
60

Second, filtration of generated characters allows the population of a library of molecules. Strings filtered out include those with syntax errors, complete strings copied from the input set, identical strings generated more than once, or strings representing invalid molecules (due to invalid valences, aromaticity, or ring strain errors)^{5,6}. There is no property or structure based filtration; all valid and unique SMILES strings are retained. The populated library represents the final output of MIMICS.

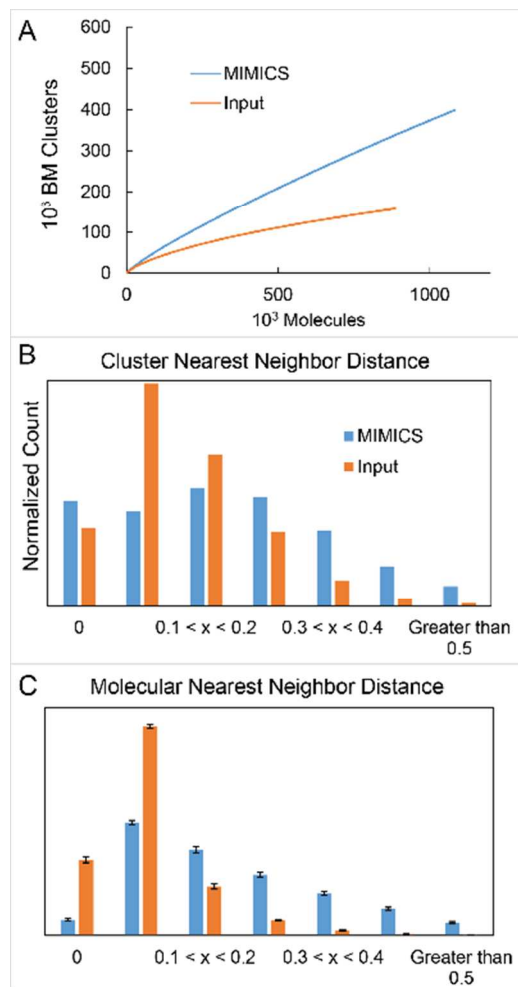


Figure 1. Structural Novelty Comparison. Generation of a set of MIMICS molecules results in a large increase in novel structures relative to the input set. (A) Bemis-Murcko clustering⁷ was conducted on the MIMICS and input molecule sets in order to assess the diversity and novelty of

1
2
3 central structural motifs. (B) The distance between a particular cluster and its nearest neighbor in
4 the input set was computed for samples of MIMICS and input molecules. MIMICS clusters
5
6 tended to have higher distances. (C) Nearest neighbor analysis for the molecules themselves
7
8 showed likewise that MIMICS molecules were farther on average from their nearest neighbors
9
10 than input molecules.
11
12
13
14

15
16 An input set was created using 880,000 molecules from the ChemBank⁸ database. Molecules
17
18 were selected based on adherence to Lipinski's Rule of Five, with the additional restriction that
19
20 no input molecules would have weight greater than 500 Da. 7.0×10^8 characters were generated
21
22 and processed into a library of 1.09×10^6 molecules using MIMICS that was then compared with
23
24 the input set. 9.2% of initially generated strings were filtered out as unusable due to repetition,
25
26 syntax errors, or invalidity and removed during processing. However, the percentage removed
27
28 for chemical invalidity was only 0.5%.
29
30
31
32

33
34 Generated molecules were first compared to the input set using Bemis-Murcko⁹ and nearest
35
36 neighbor analyses. In order to be chemically and medicinally useful, we hypothesized that the
37
38 generated set of compounds must contain both novelty and structural diversity. The 880,000
39
40 input set required 158,000 BM clusters to describe completely, while the generated set required
41
42 more than 340,000 (Figure 1A). Additionally, the number of MIMICS clusters was not observed
43
44 to converge, even when the generated set surpassed the input set in size. Nearest neighbor
45
46 analysis (Figure 1B) shows much higher density for input molecules on the higher scoring end of
47
48 the histogram. This implies that clusters that enumerate MIMICS molecules are more structurally
49
50 diverse than input molecule clusters.
51
52
53
54
55
56
57
58
59
60

Nearest neighbor analysis on samples of the molecules themselves confirms (Figure 1C) this and reinforces the lack of a one-to-one correlation between generated and input molecules. There were more than 19 times more MIMICS molecules with nearest neighbor distances higher than 0.50 than input molecules. 81% of input molecules had distances below than 0.10, compared to only 36% of MIMICS molecules. Overall, the generated set both contains novel structures and contains more structural diversity than the parent input set.

Generated molecules were compared to the input set both descriptively and structurally. Note that character generation, and thus, molecule generation, was informed only by the SMILES strings of the input molecules. No other information was available to the neural network, including atomic masses and identities, bond lengths, implicit hydrogen position, ground state 3D conformations, or the metrics and descriptors that would later be used to generate the molecules in question. Out of the 1.09×10^6 compounds generated, only 37,000 independently generated input compounds (that is, a new SMILES that corresponded to an input molecule) were present (3.4%). Because MIMICS had no information regarding the existence or structure of compounds outside its input, the remainder of the generated molecules represent novel, independent creations.

Figure 2 compares the distributions of properties of the MIMICS and input sets. Filtering based on chemical properties on the generated molecules was not conducted, and so the property distributions reflect creations of MIMICS, rather than an artificial subset of molecules. A Principal Moment of Inertia (PMI) ratio plot¹⁰ (A) of each set shows that even with only the input SMILES to work with and no background knowledge of organic chemistry, MIMICS was able to generate a set of molecules that preserved the distribution of 3D conformation. Distributions of descriptor properties (B-I) show that the two sets are comparable. Distortion on

the heavier side of the molecular weight (I) histogram is attributed to the fact that no compounds with weight greater than 500 Da were present in the input set. The relative lack of compounds between 400-500 Da is offset by the population of compounds heavier than 500 Da.

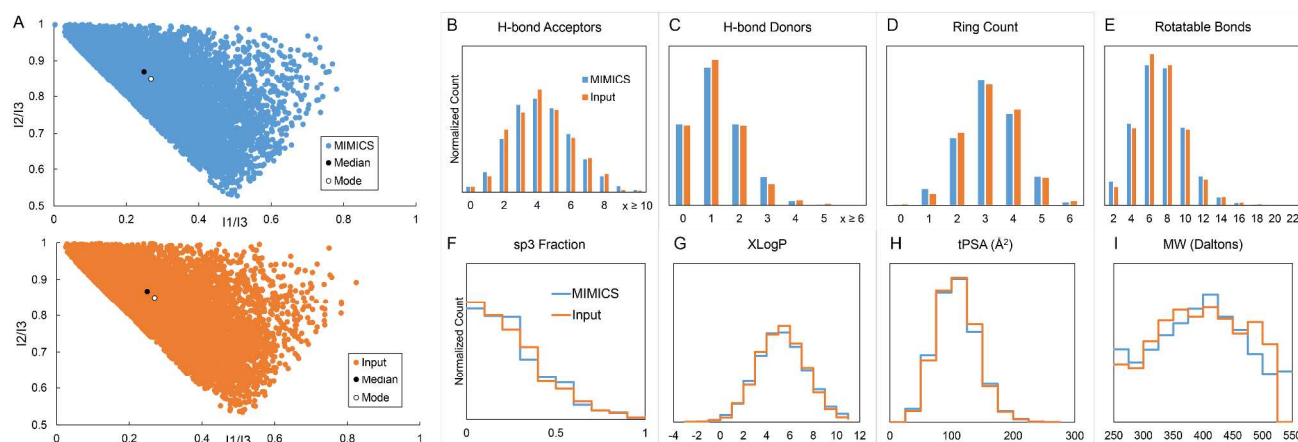


Figure 2. Comparison with input. MIMICS (blue) and input molecules (red) are compared structurally and descriptively. (A) Normalized PMI ratio plots for each set of compounds were computed. Descriptive properties computed using PaDEL-descriptor¹¹ include: (B-C) number of hydrogen bond acceptors and donors, (D) ring count, (E) rotatable bond count, (F) fraction of sp³-hybridized carbon, (G) XLogP, (H) topological polar surface area, and (I) Molecular weight (MW). For all computed descriptors, both average values and overall distributions were preserved from input set to the generated MIMICS set.

To identify compounds with potential for bioactivity, and confirm the potential of the MIMICS methodology, a subset of MIMICS that were identified as commercially available was tested at a single dose for inhibitory effect on the Unfolded Protein Response, the IRE1 α /XBP1 and PERK/ATF4 branches in particular^{12,13}. In total, of 23 MIMICS generated molecules tested, twelve molecules with inhibitory activity towards the UPR were identified (Figure 3, 2 examples), none of which had been previously identified as such. While the identified

compounds were commercially available, none were present in the input set. Because only the input molecules were available as reference during compound generation, *de novo* synthesis of a new molecule was not necessary for biologic confirmation. Note that within the MIMICS set, there were 19 independently enumerated FDA approved drugs.

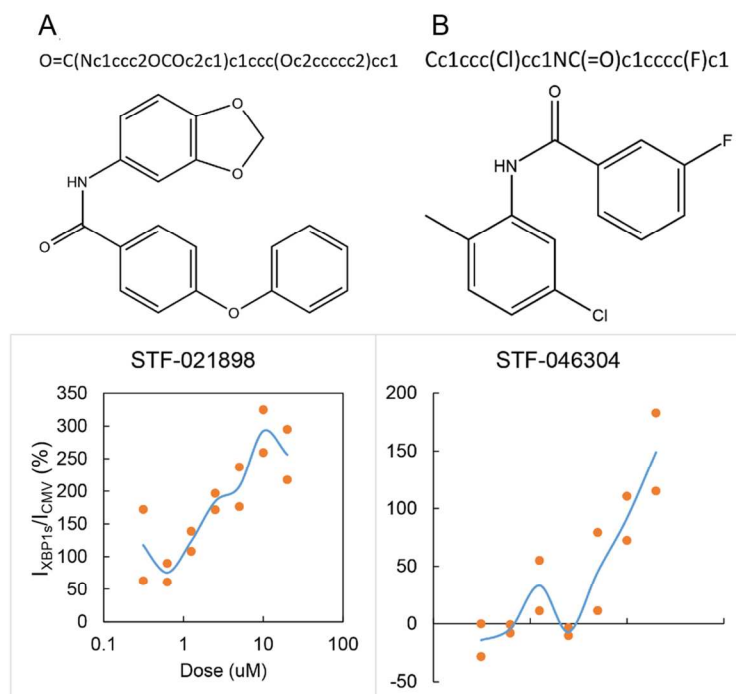


Figure 3: Confirmation of Bioactivity against the IRE1 α /XBP1 pathway, a branch of the UPR. HT1080 (human fibrosarcoma) cell line was stably transduced with an XBP1-luciferase reporter construct. Cells were treated with 300 nM thapsigargin to activate XBP1. Varying concentrations of each molecule were tested in the presence of thapsigargin to determine bioactivity. Generated SMILES expressions (top), structures of two generated inhibitors (center), and dose response curves (bottom) showing inhibitory action relative to CMV control towards IRE1 α /XBP1 for two identified inhibitors, STF-021898 (A) and STF-046304 (B).

MIMICS thus represents a unique methodology for identifying drug-like molecules, particularly in terms of the way in which compounds are generated. Rather than manipulating input

1
2
3 molecules directly, MIMICS generates molecules informed by the properties of the whole input
4 set. This results in the generation of similar sets of molecules, rather than molecules informed by
5 a single parent. While current approaches to enumerate “maximally diverse” libraries from a
6 parent currently exist, doing so invariably changes the physical properties of the resultant library.
7
8 Daughter libraries generated in such a manner are necessarily smaller than their parent: MIMICS
9
10 allows the generation of much larger libraries. The ability to direct library generation towards a
11 set of properties allows a middle ground between randomly screening whatever libraries may be
12 available or committing to a particular set of scaffolds in a combinatorial approach. By first
13 defining a chemical space, MIMICS can be directed to generate novel, structurally diverse,
14
15 libraries within those spaces.
16
17

18
19 MIMICS’s further contribution to the chemical regime of drug discovery is its ease of use. The
20 core component, char-RNN, is freely available and was popularized by its ability to replicate
21
22 Shakespearian prose and political speeches. Its ability to generate meaningful, chemically
23
24 useful, structures with minimal a priori information makes it an effective method for generating
25
26 novel molecules.
27
28

29
30 The implementation of MIMICS allows it to be used as a general purpose analytics tool. If given
31
32 a sample of a “chemical universe” and sufficient computing power, it could be used to populate
33
34 areas around compounds of interest, analogues created not by the substitution of an R-group or
35
36 heteroatom, but informed by the universe of bioactivity around it. Alternately, gaps in chemical
37
38 space already shown to exist could be filled with MIMICS compounds that not only occupy the
39
40 same space, but that also have desired physical or structural properties. MIMICS’s ability to
41
42 mimic sets of molecules in an efficient, facile manner can make practical utilization of the
43
44 vastness of chemical space possible.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ASSOCIATED CONTENT

Supporting Information. MIMICS schematic. Description of characterization methodology. Structures of 12 identified UPR inhibitors. Structures of MIMICS generated FDA approved drugs. This material is available free of charge via the Internet at <http://pubs.acs.org>.

Corresponding Author

william.yuan@trinity.ox.ac.uk; akoong@stanford.edu

ACKNOWLEDGMENTS

The authors would like to acknowledge grant support from P01 CA67166 (QTL, ACK) and the award of a Sarah and Nadine Pole Scholarship (WY) as well as support from ChemAxon for providing an academic license.

ABBREVIATIONS

MIMICS, Machine-based Identification of Molecules Inside Controlled Space; UPR, Unfolded Protein Response, SMILES, Simplified Molecular Input Line Entry System; char-RNN, Character-level Recurrent Neural Network; BM, Bemis-Muckro; PMI, Principal Moment of Inertia; IRE1 α /XBP1, Inositol-requiring enzyme 1/X-box Binding Protein 1; PERK/ATF4, PKR-like eukaryotic initiation factor 2 α kinase/Activating transcription factor 4

REFERENCES

(1) Virshup, A. M., Contreras-García, J., Wipf, P., Yang, W., Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2003**, 125(19), 7296-7303.

(2) Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. J. Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe, Chem. Inf. Model. **2015**, 55(3), 529–537.

(3) Anderson, E.; Veith, G. D.; Weininger, D. SMILES: A line notation and computerized interpreter for chemical structures. **1987**. U.S. EPA,. Report No. EPA/600/M-87/021.

(4) Karpathy, A. Multi-layer Recurrent Neural Networks for character-level language models in Torch. **2015**. Retrieved from <https://github.com/karpathy/char-rnn>

(5) O'boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. **2011**. J Cheminf. 3–33.

(6) Chemical Validity filtration computed using Marvin version 15.10.5.0 (<http://www.chemaxon.com>)

(8) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; Ferraiolo, P.; Tolliday, N. J.; Schreiber, S. L.; Clemons, P. A. ChemBank: a small-molecule screening and cheminformatics resource database., **2007**. Nucleic Acids Research, D351-9

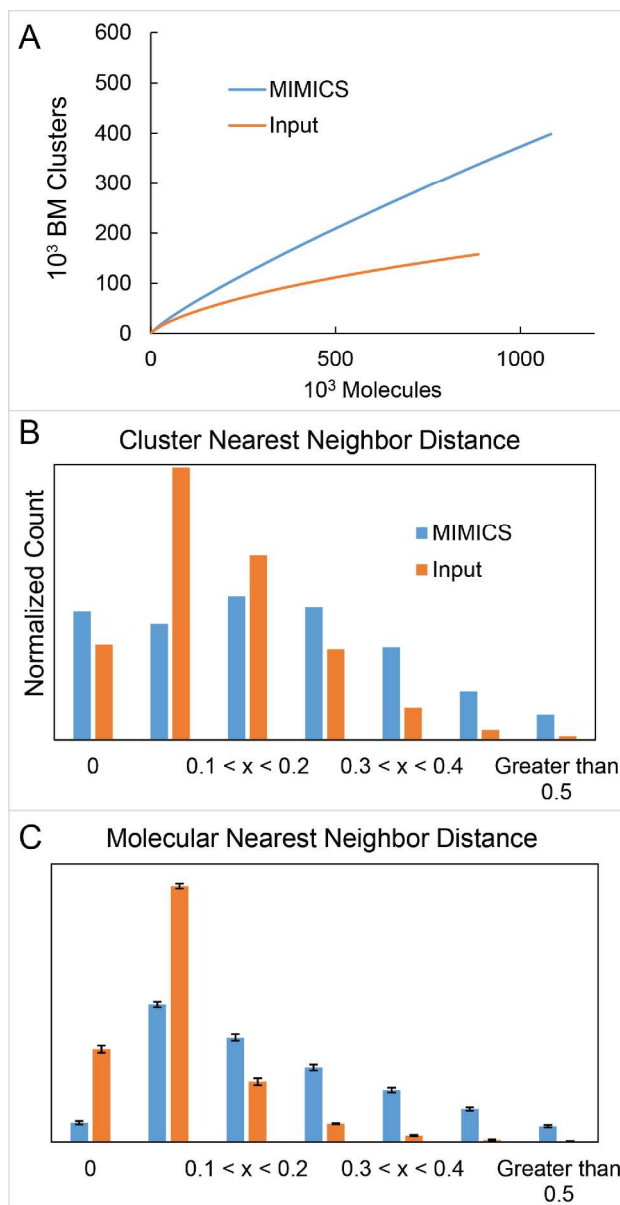
(9) Bemis Murcko clustering computed using JKlustor version 15.10.5.0 (<http://www.chemaxon.com>)

(10) Sauer W. H. B.; Schwarz M. K. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. **2003** J. Chem. Inf. Comput. Sci., 43, 987–1003

(11) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, **2011**. J. Comput. Chem., 32: 1466–1474

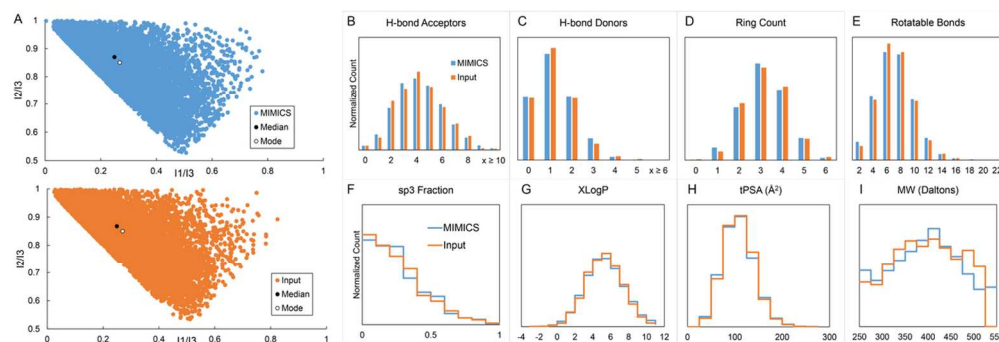
(12) Spiotto, M. T.; Banh, A.; Papandreou, I.; Cao, H.; Galvez, M. G.; Gurtner, G. C.; Denko, N. C.; Le, Q. T.; Koong, A. C. Imaging the unfolded protein response in primary tumors reveals microenvironments with metabolic variations that predict tumor growth., 2009, *A. C. Cancer Research.*, 78–88.

(13) Jiang, D.; Niwa, M.; Koong, A. C. Targeting the IRE1–XBP1 branch of the unfolded protein response in human diseases, **2015**. *Seminars in Cancer Biology.* 48–56.



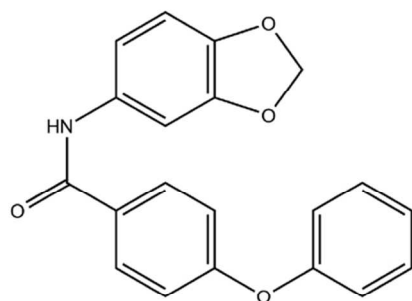
Structural Novelty Comparison. Generation of a set of MIMICS molecules results in a large increase in novel structures relative to the input set. (A) Bemis-Murcko clustering⁷ was conducted on the MIMICS and input molecule sets in order to assess the diversity and novelty of central structural motifs. (B) The distance between a particular cluster and its nearest neighbor in the input set was computed for samples of MIMICS and input molecules. MIMICS clusters tended to have higher distances. (C) Nearest neighbor analysis for the molecules themselves showed likewise that MIMICS molecules were farther on average from their nearest neighbors than input molecules.

228x441mm (300 x 300 DPI)

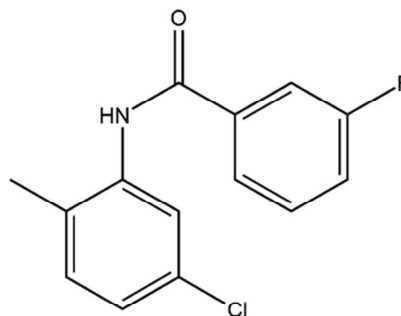


Comparison with input. MIMICS (blue) and input molecules (red) are compared structurally and descriptively. (A) Normalized PMI ratio plots for each set of compounds were computed. Descriptive properties computed using PaDEL-descriptor11 include: (B-C) number of hydrogen bond acceptors and donors, (D) ring count, (E) rotatable bond count, (F) fraction of sp³-hybridized carbon, (G) XLogP, (H) topological polar surface area, and (I) Molecular weight (MW). For all computed descriptors, both average values and overall distributions were preserved from input set to the generated MIMICS set.
57x19mm (600 x 600 DPI)

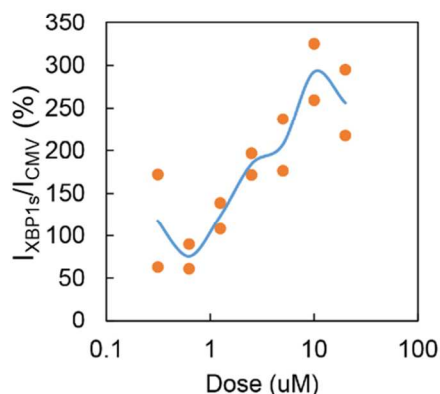
A

O=C(Nc1ccc2OCOc2c1)c1ccc(Oc2ccccc2)cc1


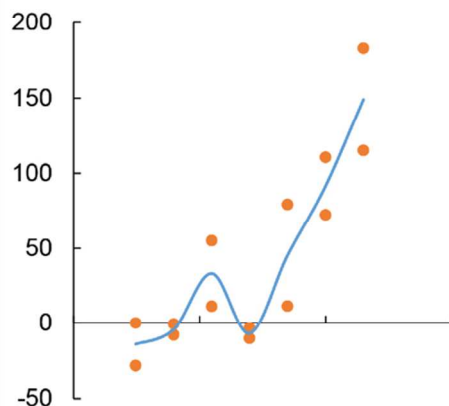
B

Cc1ccc(Cl)cc1NC(=O)c1cccc(F)c1


STF-021898



STF-046304



Confirmation of Bioactivity against the IRE1 α /XBP1 pathway, a branch of the UPR. HT1080 (human fibrosarcoma) cell line was stably transduced with an XBP1-luciferase reporter construct. Cells were treated with 300 nM thapsigargin to activate XBP1. Varying concentrations of each molecule were tested in the presence of thapsigargin to determine bioactivity. Generated SMILES expressions (top), structures of two generated inhibitors (center), and dose response curves (bottom) showing inhibitory action relative to CMV control towards IRE1 α /XBP1 for two identified inhibitors, STF-021898 (A) and STF-046304 (B).

173x158mm (150 x 150 DPI)