Supporting Information for:

Descriptor driven de novo design algorithms for DOCK6 using RDKit

Guilherme Duarte Ramos Matos,[§,1,2] Steven Pak,[§,3] and Robert C. Rizzo*[,1,4,5]

[1] *Department of Applied Mathematics & Statistics, Stony Brook University, Stony Brook, New York 11794, USA.*

[2] *Instituto de Química, Universidade de Brasília, Brasília, Distrito Federal, 70910-900, Brazil*

[3] *Department of Pharmacological Sciences, Stony Brook University, Stony Brook, New York, 11794, USA.*

[4] *Institute of Chemical Biology & Drug Discovery, Stony Brook University, Stony Brook, New York 11794, USA.*

[5] *Laufer Center for Physical & Quantitative Biology, Stony Brook University, Stony Brook, New York 11794, USA.*

* Corresponding author e-mail: rizzorc@gmail.com, phone: 631-632-9340, fax: 631-632-8490.

[§] These authors contributed equally to this work.

DrugC dataset and properties derived from molecules in the DrugCentral database. To generate a set of default descriptor values for D3N protocols, we downloaded 4,052 small molecules from the DrugCentral database[1,2] (drugcentral.org, downloaded 9/21/2021), containing approved small molecule drugs and active pharmaceutical agents, as a csv file containing identifiers and SMILES strings. We next searched these SMILES strings in the ZINC15 database (zinc15.docking.org)[3] and identified 3,069 matches for which a DOCK6-ready MOL2 file could be downloaded. We further refined the dataset in an attempt to retain only those compounds that would be orally available, with a specific mechanism of action, by removing entries if labeled as a radioisotope (indicated by a F18 or C13 in the name, N=13 removed) or the compound contained linear alkyl chains with 9 or more carbons (potential topical agents, N=38 removed). The remaining 3,018 compounds were processed with the DOCK6 database filter to remove those if their formal charge was lower than -3 or higher than +2, their LogP was less than -4.5 or greater than 7.0, or their MW was less than 100 or greater than 750 (N=183 removed). These ranges corresponded to roughly the upper and lower first percentiles with respect to the group of 3,018 molecules. Table S1 lists the mean, standard deviation (σ), and highest and lowest values for the final set of 2,835 compounds, which we termed the DrugC dataset, computed using the DOCK6/RDKit implementation.

Table S1. Mean, standard deviation, and highest and lowest values for seven drug-like descriptors derived from 2,835 molecules in the DrugC dataset.

Descriptor	Min to max range	Mean	Std dev (σ)	Highest	Lowest
QED	0 to 1	0.61	0.19	0.94	0.06
SynthA	1 to 10	3.34	0.90	7.48	1.11
TPSA[a]	0 to +	70.87	42.33	323.92	0.0
LogP	- to +	1.73	2.02	6.97	-4.40
LogS	- to +	-3.29	1.94	2.56	-9.67
#Stereo	0 to +	1.05	N/A	13	0
#PAINS	0 to +	0.06	N/A	3	0

[a] TPSA values in angstroms squared.

Descriptor distributions for the DrugC dataset and ZINC13M show similar trends. As noted in the main text, the DOCK_DN fragment library was derived from 13M molecules downloaded from ZINC (ZINC13M), which is orders of magnitude larger than the group of 2,835 compounds in the DrugC dataset. To assess if the underlying molecular properties between the datasets were similar, we compared descriptors distributions for nine properties as shown in Figure S1. Overall, although some DrugC distributions are somewhat broader in comparison to their ZINC13M counterparts, the plots for TPSA, LogP, LogS, #Aromatic, #Stereo, #Spiro, and #PAINS show remarkable similarity in terms of shape and peak location. For QED and SynthA however, although the general trends are similar, there are some differences. Interestingly, despite containing primarily approved small molecule drugs and active pharmaceutical agents, compounds from DrugC (Figure S1 red) yield a broader range of less favorable QED scores compared to ZINC13M (Figure S1 blue) which shows a sharper peak at 0.8. And, although the SynthA distributions for both datasets are similar, the distribution peak for ZINC13M (blue) is shifted left which indicates the compounds are somewhat more synthetically accessible. Despite these differences, the overall good correspondence between the two datasets provides strong support for using the standard DOCK_DN fragment library in conjunction with the default D3N-drugc ranges.

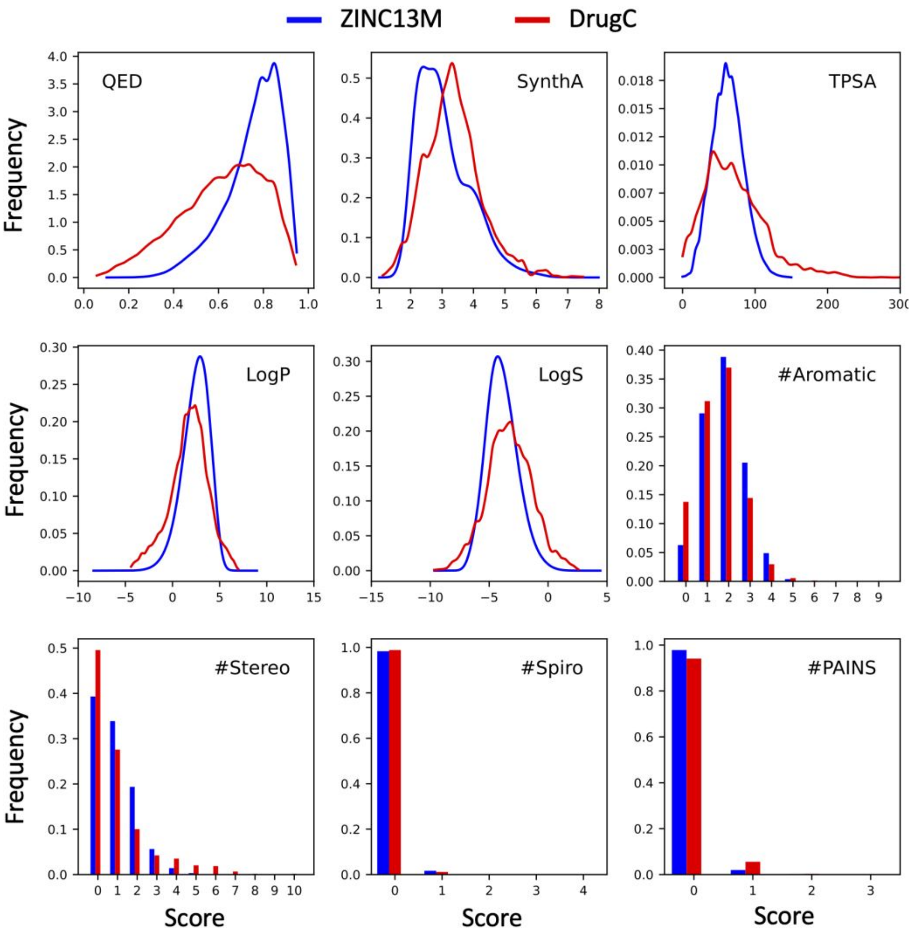

Figure S1. Cheminformatics descriptors calculated by DOCK6/RDKit for 13M molecules downloaded from ZINC (ZINC13M, blue)[3] and 2,835 active pharmaceutical agents curated from DrugCentral (DrugC, red).[1,2] See text for curation protocol. The first five distributions were normalized by using kernel density estimation. The last four bar plots were normalized manually. TPSA values in angstroms squared.

Comparison of descriptor distributions obtained using D3N-drugc and D3N-loose filtered. Figure S2 compares results obtained using the D3N-drugc protocol (on-the-fly pruning approach, red) vs the D3N-loose filtered protocol (build-all-then-filter approach, gray) which yield smooth-tailed vs hard-cut distributions.

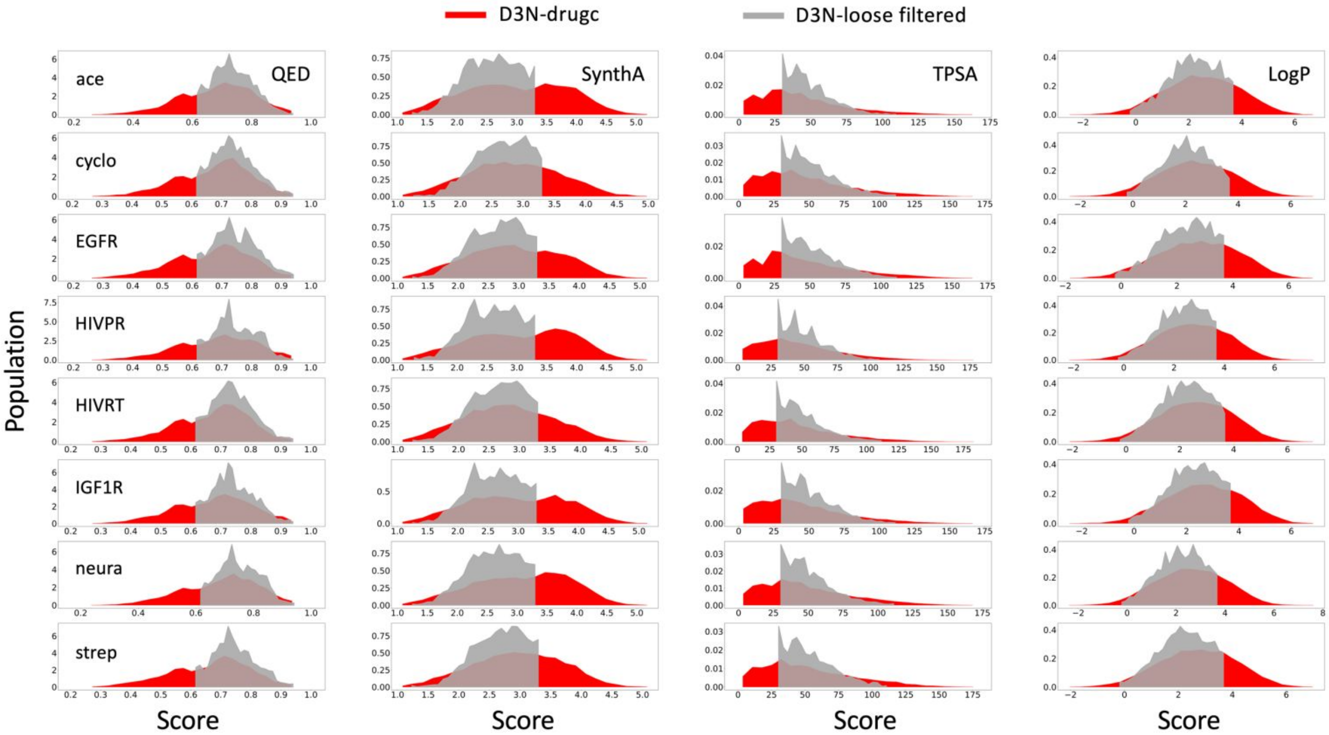

Figure S2. Descriptor distributions (QED, SynthA, TPSA, LogP) for molecules constructed using the D3N-drugc protocol (red area) versus the D3N-loose protocol hard-filtered to conform to the D3N-drugc target ranges (gray area). All distributions normalized. TPSA values in angstroms squared. D3N-drugc target ranges in Table 3.

References

(1) Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C. G.; Yang, J. J.; Mathias, S. L.; Nelson, S. J.; Oprea, T. I. DrugCentral: online drug compendium. *Nucleic Acids Res* **2017**, *45*, D932-d939.

(2) Avram, S.; Bologa, C. G.; Holmes, J.; Bocci, G.; Wilson, T. B.; Nguyen, D.-T.; Curpan, R.; Halip, L.; Bora, A.; Yang, J. J.; Knockel, J.; Sirimulla, S.; Ursu, O.; Oprea, T. I. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* **2021**, *49*, D1160-D1169.

(3) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 177-182.