Supporting Information

ZINClick v.18: Expanding Chemical Space of 1,2,3-Triazoles

*Doriana Levré, Chiara Arcisto, Valentina Mercalli, Alberto Massarotti**

Dipartimento di Scienze del Farmaco, Università degli Studi del Piemonte Orientale "A. Avogadro", Largo Donegani 2, 28100 Novara, Italy.

*Phone: +39.0321.375.753, Fax: +39.0321.375.821, E-mail: alberto.massarotti@uniupo.it

Table of Contents

S1. EXPERIMENTAL PROCEDURES

S1.1. Other subsets. To reduce the number of ZINClick compounds, a random selection (10%) of the compounds in ZINClick was generated. The 10%-subset (10S) consists of 1,690,778 compounds. To assess the structural diversity of ZINClick, a small diversity-subset (DS) was prepared using some dedicated functionalities implemented by RDKit. First, ECFP4[1] equivalent Morgan fingerprints were generated as bit vector for the compounds. This circular fingerprint comprises 2048 bits and represents atom connectivity and chemical features, in addition to taking account of the neighborhood of each atom. Finally, the MaxMin[2] algorithm was used to grab 2,500 different compounds. A diversity analysis was performed using both Principal Component Analysis (PCA) and Principal Moment of Inertia (PMI) techniques implemented in Python by RDKit. The diversity subset was exported in SMILES, 2D and 3D SDF formats.

S1.2. PAINS Analysis. PAINS filters reported in the literature in Sybyl Line Notation (SLN format (Tables S6, S7, and S9 in the Supporting Information from the work of Baell and Holloway)[3] were considered. The ZINClick database was mapped to individual PAINS substructure motifs using RDKit scripts based on the filters published by the Guha group.[4]

S1.3. External sources. We collected data from other databases of chemical entities (ZINC, ChEMBL, SureCHEMBL, PDB, CSD) using some python scripts based on RDKit functionalities. Each database was filtered out for 1,4-disubstituted-1,2,3-triazole, alkyne and azide substructures. Then, the outcomes were cross-checked with ZINClick items looking through canonical SMILES correlation. Any time two structures were found matched an ID tag was applied to the triazole entry in ZINClick to connect the external resource. More details about each source are presented below.

ZINC15.[5] "2*H*-triazole-substances" trance was downloaded as result of a ring-based substructure search (search performed on February 2018). It contained 3,405,900 randomly substituted triazoles, 1,033,111 were found bringing 1,4-substitution. Of these, 6,042 were found in ZINClick.

ChEMBL.[6] The entire 23rd version of ChEMBL was downloaded from the EMBL-EBI website and included 1,727,112 compounds. Among these molecules, 11863 structures of 1,4-disubstituted 1,2,3-triazole were identified. Of these, 571 structures of 1,4-disubstituted 1,2,3-triazole were found in ZINClick.

PDB.[7] From the whole Protein Data Bank ligand database 148 deposit containing a compound with a 1,4-disubstituted-1,2,3-triazole (search performed on February 2018). Of these, 9 were found in ZINClick.

CSD.[8] 710 crystal structures of triazoles were recovered from the Cambridge Structural Database (search performed on March 2018). Macrocycles bringing more than one triazole ring were discarded, 556 objects were kept eventually. Of these, 19 were found in ZINClick.

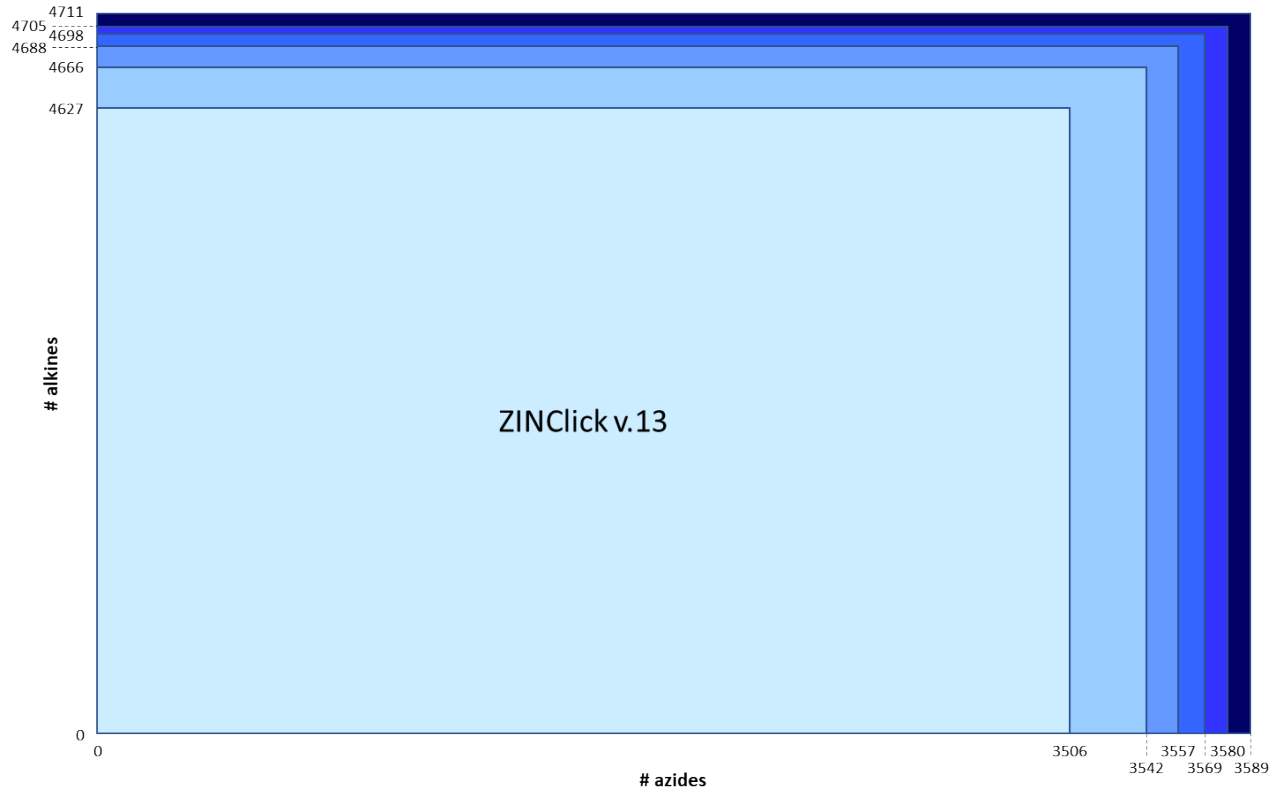


Figure S1. The increment of the ZINClick size (n. of molecules) over the years. The new azides and alkynes available every year determine the increment of triazole numbers available in ZINClick v.13 (cyan square) compared to the newest versions (v.18 is in dark blue). A small increment in the number of azides and alkynes available induce a remarkable effect on the total number of triazole generated.

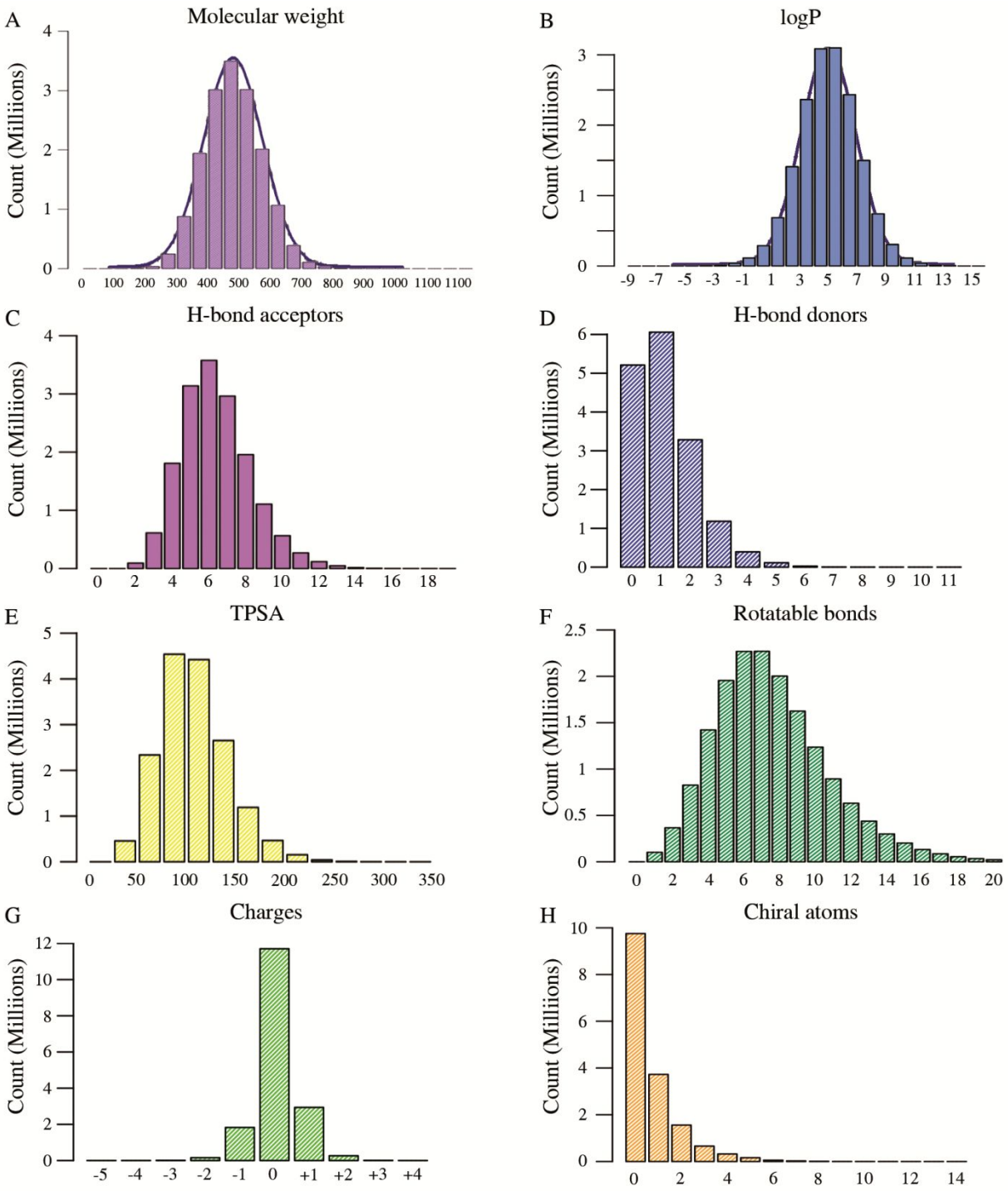


Figure S2. Distribution histograms of the descriptors (MW, logP, HBA, HBD, TPSA, rotB, charges and chiral centers respectively) for the compounds in the ZINClick database v.18.

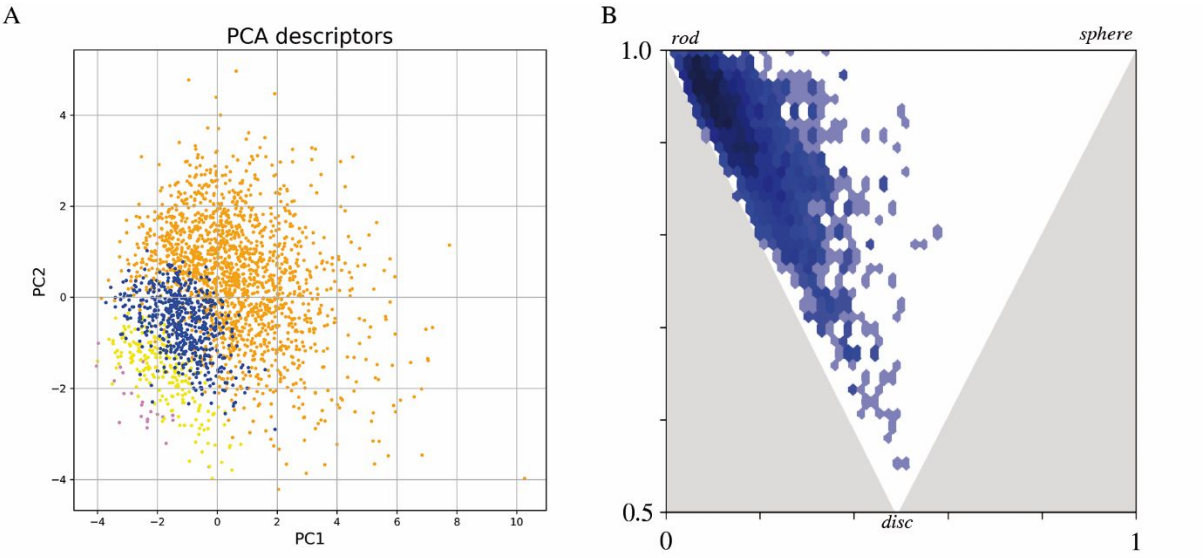
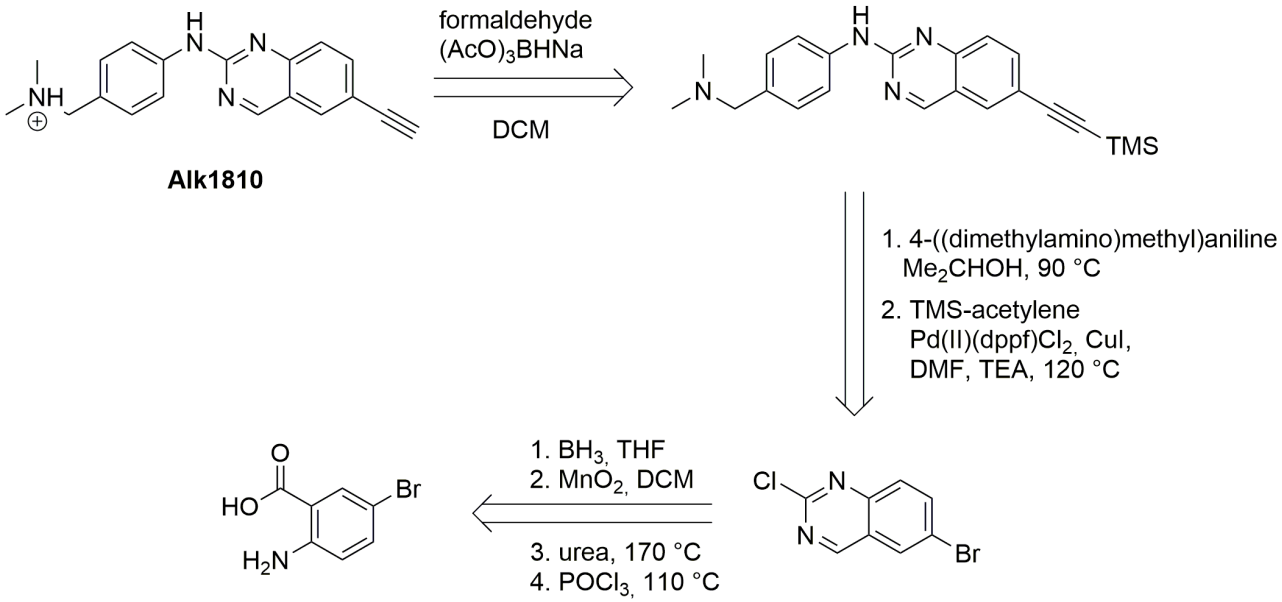


Figure S3. Diversity of ZINClick v.18. (A) A principal component analysis (PCA) plot depicting the chemical space defined by the ZINClick database: all compounds (orange), the "drug-like" subset (blue), the "lead-like" subset (yellow) and the "fragment-like" subset (violet). The variance covered by PC1 and PC2 is 40% and 22%, respectively. (B) A 3D shape analysis of the diversity-subset of ZINClick.

Table S1. PAINS structural motifs present in ZINClick v.18. Their frequency of occurrence in the original reactants and in the final database. The REGID matches the structural motifs in Tables S6, S7 and S9 in Supporting Information from Baell and Holloway.[3] Only rules with non-zero values are reported.

REGID	Number of flagged alkynes	Number of flagged azides	Number of flagged ZINClick compounds
anil_di_alk_A	5	47	239127
quinone_A	3	29	147299
dyes5A	25	7	122527
azo_A	26	0	93314
anil_no_alk	8	6	56930
anil_di_alk_C	16	0	57424
ene_six_het_A	3	4	29599
anil_di_alk_E	5	3	32063
catechol_A	3	3	24891
indol_3yl_alk	3	3	24891
pyrrole_B	0	5	23555
anil_di_alk_B	1	4	22429
ene_five_het_G	3	3	24891
anthranil_one_A	1	2	13009
sulfonamide_B	0	4	18844
thiophene_hydroxy	0	2	9422
mannich_A	5	1	22651
amino_acridine_A	1	2	13009
acyl_het_A	2	0	7178
anil_di_alk_D	3	0	10767
ene_cyano_A	0	1	4711
keto_keto_beta_A	1	1	8299
imidazole_A	0	1	4711
anil_di_alk_K	0	1	4711
imine_one_A	1	0	3589
imine_one_isatin	1	0	3589
het_thio_5_B	1	0	3589
tert_butyl_A	1	0	3589
anil_OC_alk_E	1	0	3589
anil_NH_alk_C	1	0	3589
Total	120	129	1,037,786

Scheme S1. Synthesis of compound Alk1810.[9]



References

1. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010,** *50*, 742-754.
2. Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets Using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct.-Act. Relat.* **2002,** *21*, 598-604.
3. Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (Pains) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010,** *53*, 2719-2740.
4. Guha, R. Pains Substructure Filters as Smarts. http://blog.rguha.net/?p=850 (accessed 04-07-2013).
5. Sterling, T.; Irwin, J. J. Zinc 15--Ligand Discovery for Everyone. *J. Chem. Inf. Model* **2015,** *55*, 2324-2337.
6. Fechner, N.; Papadatos, G.; Evans, D.; Morphy, J. R.; Brewerton, S. C.; Thorner, D.; Bodkin, M. Chemblspace--a Graphical Explorer of the Chemogenomic Space Covered by the Chembl Database. *Bioinformatics* **2013,** *29*, 523-524.
7. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000,** *28*, 235-242.
8. Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Cryst.* **2016,** *B72*, 171-179.
9. McArthur, S.; Hertel, C.; Nettekoven, M.; Raab, S.; Roche, O.; Rodriguez-Sarmiento, R.; Schuler, F.; Plancher, J. M. Indole Derivatives as H3 Inverse Agonists. US 20050282864 A1, **2005**.