JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

ZINClick: A Database of 16 Million Novel, Patentable, and Readily Synthesizable 1,4-Disubstituted Triazoles

Alberto Massarotti,* Angelo Brunco, Giovanni Sorba, and Gian Cesare Tron*

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

Supporting Information

ABSTRACT: Since Professors Sharpless, Finn, and Kolb first introduced the concept of "click reactions" in 2001 as powerful tools in drug discovery, 1,4-disubstituted-1,2,3-triazoles have become important in medicinal chemistry due to the simultaneous discovery by Sharpless, Fokin, and Meldal of a perfect click 1,3-dipolar cycloaddition reaction between azides and alkynes catalyzed by copper salts. Because of their chemical features, these triazoles are proposed to be *aggressive pharmacophores* that participate in drug−receptor interactions while maintaining an excellent chemical and metabolic profile. Surprisingly, no virtual libraries of 1,4-disubstituted-1,2,3-triazoles have been generated for the systematic investigation of the click-chemical space. In this manuscript, a database of triazoles called ZINClick is generated from literature-reported alkynes and azides that can be synthesized within three steps from commercially available products. This combinatorial database contains over 16 million 1,4-disubstituted-1,2,3-triazoles that are *easily synthesizable*, *new*, and *patentable*! The structural diversity of ZINClick (http://www.symech.it/ZINClick) will be explored. ZINClick will also be compared to other available databases, and its application during the design of novel bioactive molecules containing triazole nuclei will be discussed.

1. INTRODUCTION

Virtual screening[1] is an important tool during the search for novel bioactive compounds despite its recognized pitfalls.[2] This methodology has become important within numerous drug discovery programs in academia and industry. The opportunity to screen databases containing over a million compounds in a relatively short period is an added bonus for drug hunters and must be exploited to the fullest. Several review articles[3] and books[4] reported successful cases where virtual screening campaigns have unveiled novel hit compounds as starting points for classical lead optimization procedures. To date, the number of virtual libraries, both public and private,[5] available for computational chemists is growing exponentially, increasing the regions of chemical space that can be used against a specific biological target. Several libraries are already available for docking with 3D structures and include the correct protonation, tautomeric, and conformational states. One of the most important databases used worldwide is ZINC. This free database includes approximately 20 million commercially available compounds; it was developed by Stoichet[6] and improved by Irwin.[7] Another well-known library is the ACD (Available Chemical Directory) from Molecular Design Limit.[8] Although this library contains over 2.3 million compounds, it is not public, and it is expensive.

To expand the chemical space available for docking, other authors created novel combinatorial databases. For example, TIN is a combinatorial collection of compounds generated by employing a definite set of multicomponent reactions.[9] Our experience with multicomponent reactions[10] suggests that not all of the compounds present in that database are easily available due to the strong dependence upon multicomponent-derived products. Another published database enumerates the entire virtual chemical universe of molecules containing up to 11, 13, and 17 atoms formed by combining the following four elements: carbon, nitrogen, oxygen, and fluorine.[11] Although the types of molecules generated using this method are interesting, usually these compounds lack readily apparent synthetic routes. Two papers have demonstrated that medicinal chemists tend to use the same reliable and robust reactions to build their molecular scaffolds during drug discovery.[12]

Databases containing natural products have also been reported.[13] In this case, the retrieval of several natural compounds is difficult, limiting the usefulness of these databases.

Apart from the cost of the compounds, other problems may arise from the vendors. Sometimes, the desired molecules are no longer available or their synthesis has never been reported, rendering the use of these molecules impractical. In other cases, the amount of a compound available is only enough for preliminary screening; some vendors have also had resupply issues. This drawback has been highlighted by at least one paper.[14] The patentability of the compounds is another issue. Many compounds are not new, generating problems with

Received: September 9, 2013
Published: January 22, 2014

dx.doi.org/10.1021/ci400529h | J. Chem. Inf. Model. 2014, 54, 396−406

intellectual property. Finally, the same compounds can be ordered by competitors working on the same projects.

These considerations show one of the most important problems that computational chemists face when involved in virtual screening research: rapid access to the selected compounds to test their hypotheses. Because most computational chemists are not synthetic chemists, they must ask to an organic chemist to collaborate if the predicted compounds are not available. These undertakings are not easy, especially when the required synthesis is long and challenging; synthetic chemists are notoriously reluctant to undertake long syntheses for compounds predicted via computer.

The synthetic feasibility of the selected compounds is critical during the drug discovery process. Only a few reactions are commonly used by medicinal chemists; the most recurrent functional group in drugs is the amide due to its accessibility.[12b] Finally, the number of synthetic steps required to synthetize the molecules must be considered. With few exceptions, drug candidates should not require more than eight synthetic steps.[12a]

In 2001, Professors Sharpless, Finn, and Kolb introduced "click reactions"[15] as powerful tools for drug discovery. Afterward, 1,4-disubstituted-1,2,3-triazoles became important in synthetic chemistry thanks to the simultaneous discovery by Sharpless, Fokin, and Meldal of the perfect *click* 1,3-dipolar cycloaddition reaction between azides and alkynes catalyzed by copper salts.[16] Because of their electronic nature, these triazoles might be *aggressive pharmacophores*[17] that can actively participate in drug−receptor interactions while maintaining an excellent chemical and metabolic profile.

New technologies require time to mature and impact research, usually between 10 and 15 years.[18] Despite the interesting features of click chemistry, a 2008 paper revealed that only 14% of the published works on click chemistry were related to drug discovery,[19] indicating the reluctance of medicinal chemists to use this reaction despite the successful results.[20] We have considered using a library of 1,4-disubstituted triazoles for virtual screening since 2009. After a literature search, we noticed that no systematic attempt has been made to generate a feasible click library, and no virtual libraries of 1,4-disubstituted-1,2,3-triazoles existed to investigate the click-chemical space systematically. While conducting Internet research, we found different enumerating software programs (e.g., Chemaxon reactor,[21] Maestro,[22] MOE,[23] etc.) able to generate virtual libraries of molecules that could be constructed using click reactions. Another Web site provided a freely downloadable program that generates a library of 1,4-disubstituted triazoles after the insertion of azides and alkynes;[24] however, the overall combination output is limited to 10,000 triazoles, limiting its usefulness. In these programs, the operator must insert the desired starting materials, and the selection process for the azides and alkynes is too arbitrary. The same concept has been published, generating AutoClickChem.[25] Even in this case, a series of reactions is listed following the click criteria, and a library will be generated after inserting the preferred building blocks. These methods are useful, but they are limited by the number of building blocks and the process of selecting which building block to add.

In this manuscript, we describe the preparation of a novel library containing over 16 million 1,4-disubstituted-1,2,3-triazoles that can be synthesized using *existing* alkynes and azides within three synthetic steps starting from commercially available products while using an established synthetic procedure. Because of its public nature and the inspiration provided by the work of Stoichet,[6] we called this library ZINClick. Importantly, most of the molecules in this database are *new* and *patentable*. Because they are derived from known azides and alkynes, after identification using virtual screening and docking procedures, these compounds should be easily prepared, enabling tests of the docking hypotheses. Using azides and alkynes reported in the literature synthesized for a particular scope generates a library of triazoles as diverse as possible. Many medicinal chemists tend to create molecules that are feasible following their chemical knowledge and experience.

2. EXPERIMENTAL PROCEDURES

All molecular modeling studies were performed on a Tesla workstation equipped with two Intel Xenon X5650 2.67 GHz processors and Ubuntu 10.04 (www.ubuntu.com). A standard LAMP server was prepared. The Web site was programmed using PHP with the JSDraw structure editor (www.scilligence.com), and a MySQL database (www.mysql.com) was used to store the chemical structures. Different software was used to elaborate the chemical information: PerlMol and OpenEye (www.eyesopen.com) applications. The graphs were generated using R.[26] The protein structures and 3D chemical structures were generated in PyMOL.[27]

2.1. Compounds Source. The collection of azides and alkynes was retrieved using SciFinder.[28] This searchable database includes chemical structures available from chemical vendors, as well as published literature and patents. The structures were downloaded using the Web-based interface and manually added to the database. To assess whether the selected triazoles could be synthesized in a reasonable number of steps, we decided to use azides and alkynes with reported syntheses shorter than three steps when beginning from commercially available products. Consequently, the maximum number of synthetic steps required to synthesize the desired compounds could not exceed seven.

Additionally, we made 500 Da the upper limit for the molecular weights for the azides and the alkynes, facilitating compatibility with synthetic drugs. Therefore, the maximum molecular weight for the final compounds cannot exceed 1000 Da. For chiral molecules, if the absolute stereochemistry was not available, both enantiomers were generated. Using these criteria, 4627 alkynes and 3506 azides were retrieved (the database was updated December 31, 2012). The azides and alkynes were saved as canonical SMILES and stored in a MySQL database.

2.2. Generation of ZINClick. An *in house* Perl script based on PerlMol[29] was used to retrieve each azide and perform a virtual combinatorial reaction with all alkynes. The resulting SMILES for 1,4-triazoles were saved in a MySQL database. Each structure was rebuilt as a 3D structure using OMEGA2.[30] The molecular descriptors (listed in Table S1, Supporting Information) were calculated for each output. These calculations were performed with the FILTER[31] software from OpenEye. Finally, the database was exported in SMILES format. For internal use, the resulting SMILES strings were stored in a MySQL database to facilitate standard Structured Query Language (SQL) searches and to prepare for future Web server integration.

2.3. Data Curation for ZINClick. To prevent automation problems during the creation of ZINClick, the generation steps were carefully checked. The collections of alkynes and azides were checked for duplicate structures. For each generation process, a log file detailing any errors was produced. All errors were corrected manually; otherwise, the compound was deleted.

2.4. Reference Databases. Three different compound databases were considered: the ZINC "All Purchasable" subset

(19,607,982 compounds dated January 11, 2013), the ZINC "Natural product" subset (189,355 compounds), and the DrugBank "Approved Drugs" set (1486 compounds).[32] The molecular descriptors were subsequently calculated for all databases using FILTER. The descriptors were derived from the Lipinski "drug-like",[33] the Oprea "lead-like",[34] and the rule-of-three "fragment-like" rules:[35] molecular weight (MW), log P(o/w), number of hydrogen bond acceptors (HBAs), number of hydrogen bond donors (HBDs), total polar surface area (TPSA), and amounts of rotatable bonds (rotB), charges, and chiral atoms.

2.5. Properties Subsets. Three different collections of ZINClick compounds were prepared using the following criteria: (I) A drug-like subset (DL) was based on Lipinski's rules: MW ≤ 500 and MW ≥ 150, log P ≤ 5, rotB ≤ 7, PSAa < 150, HBD ≤ 5, and HBA ≤ 10.[33] (II) A lead-like subset (LL) was used: MW ≤ 350 and MW ≥ 250, log P ≤ 3.5, and rotB ≤ 7.[34] (III) A fragment-like subset (FL) was also used: log P ≤ 3.5, MW ≤ 250, and rotB ≤ 5.[35] The subset collections include 4,140,042, 744,235, and 22,104 compounds, respectively. The subsets are exported in both SMILES and SDF formats.

2.6. 10% Subset. To reduce the number of ZINClick compounds, a random selection (10%) of the compounds in ZINClick was generated. The 10% subset (10S) consists of 1,622,226 compounds and is available in both SMILES and SDF formats.

2.7. Diversity Subset. To assess the structural diversity of ZINClick, a diversity subset (DS) was prepared. First, the ChemFP Substruct fingerprints were generated for the compounds using chemfp v1.1[36] with the OEChem[37] and GraphSim[38] Python Toolkits. This fingerprint comprises 881 bits; they are not the same but were "heavily inspired" by the CACTVS 661-bit substructure fingerprints used in PubChem.[39] Finally, the SUBSET 1.0 algorithm[40] was used to select compounds that differ from those previously selected by at least a Tanimoto cutoff of 0.7; 2323 compounds in SMILES format were selected. To provide a representative group of molecules for comparison with the ZINClick diversity set, 27,442 structures were selected from ZINC using the same procedure. A short perl program was written to record the occurrence of each fingerprint for plotting. The PCA plots were generated using R.[26] The diversity subset was exported in both SMILES and SDF formats.

2.8. PAINS Analysis. PAINS filters present in the literature in SLN format (Tables S6, S7, and S9, Supporting Information, from the work of Baell and Holloway)[41] were considered. The ZINClick database was mapped to individual PAINS sub-structure motifs using in-house perl scripts containing the SMARTS form version of the filters published by the Guha group.[42]

3. RESULTS AND DISCUSSION

The ZINClick database provides 16,222,262 virtual 1,4-disubstituted 1,2,3-triazoles generated using *existing* alkynes and azides that require no more than three synthetic steps from commercially available products and have a molecular weight cutoff of 500 Da.[43] After the database was generated, common descriptive properties were computed to define the drug-like and lead-like compounds. The Lipinski[33] violations for the ZINClick compounds are reported in Figure 1. Of the 16 million compounds, 4,140,042 (25%) are drug-like and 744,235 are lead-like (5%) (the detailed count is reported in Table S2, Supporting Information). To understand these results, we

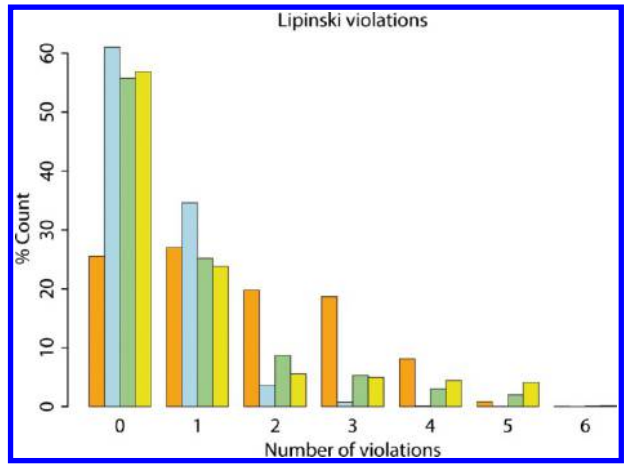


Figure 1. Histogram of the Lipinski violations as percentages of the ZINClick (orange), ZINC (cyan), ZINC "Natural product" (green), and DrugBank (yellow) databases. Details are reported in Table S2 of the Supporting Information.

computed these data using three other databases: the ZINC "All Purchasable" subset, the ZINC "Natural product" subset, and DrugBank, representing the known chemical space, the natural product chemical space, and the drug chemical space, respectively. Clearly, the drug-like character of our library is poor compared to the above libraries; in those libraries, the amounts of drug-like compounds account for, respectively, approximately 61%, 56%, and 57% of the total. This result includes only compounds that do not violate Lipinski's rules. In our case, the choice to retrieve azides and alkynes with a molecular weight ≤ 500 Da, dramatically reduced the percentage of triazoles with a drug-like molecular weight.[44] If the molecules with no more than one violation are included, about half of the ZINClick library is drug-like (52%; 8,522,007 compounds).

Lipinski's rules are depicted in Figure 2. After analysis, the molecular weight (MW) distributions show a peak value between 450 and 500 Da (Figure 2a) on a curve similar to a Gaussian distribution. The distributions of the log P values also show a Gaussian-shaped curve with a peak centered at 3.5 log P units (Figure 2b). The peaks for the hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD) are at 8 and 1, respectively, and both curves fall off rapidly at maximas of 20 and 12, respectively (Figure 2c,d). The polar surface area (PSA) and rotatable bonds count (rotB) distribution have peaks at approximately 100 and 7, respectively (Figure 2e,f). Finally, approximately 68% of the compounds are neutral, and 7,062,641 structures (43%) contain at least one stereogenic center (Figure 2g,h).

The distributions for each descriptor of compounds from ZINClick and the other databases are compared using a line graph (Figure 3). The molecular weight space covered by ZINClick is shifted toward higher values (Figure 3a) due to the MW cutoff used while preparing the database. The distributions of the log P values are more similar (Figure 3b). The clusters are equally populated between 2 and 5, while ZINC shows a preference for the region between 2 and 4. The number of hydrogen bond acceptors in ZINClick compounds is generally double that of compounds from the other databases (Figure 3c). However, for hydrogen bond donation capabilities, similar behavior was observed between all of the databases (Figure 3d). Consequently, for the higher MW and HBD counts in ZINClick, higher polar surface area values are generated relative to ZINC

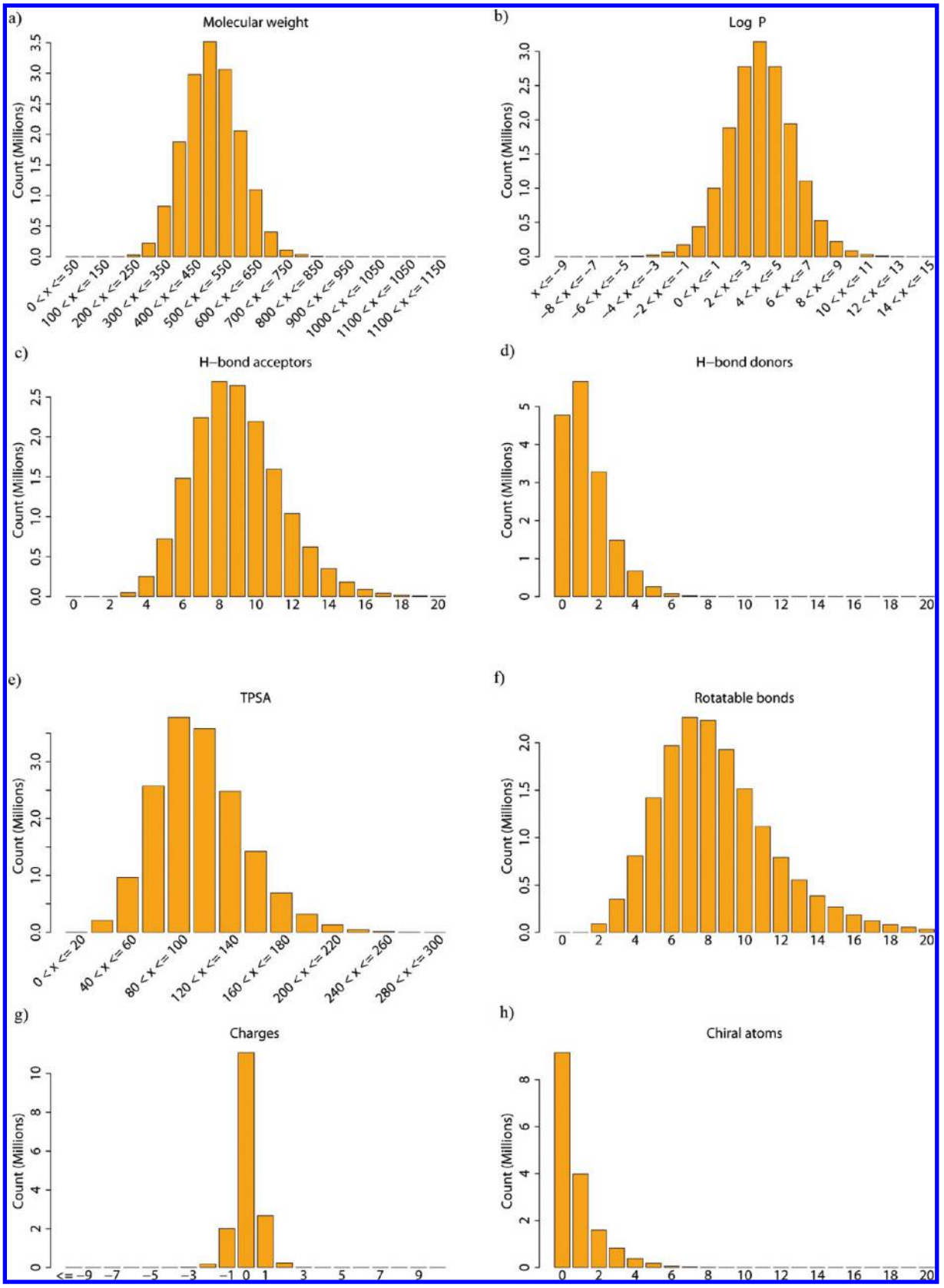


Figure 2. Distribution histograms for the descriptors (MW, log P, HBA, HBD, TPSA, rotB, charges, and chiral, respectively) of the compounds in ZINClick.

and the other databases (Figure 3e). Unsurprisingly, the ZINClick database is very different from the Natural products and DrugBank databases; for the latter, fewer rotatable bonds are preferred (Figure 3f). Finally, a similar net charge distribution (Figure 3g) is apparent, while chiral molecules are more common in the Natural products database (Figure 3h).

dx.doi.org/10.1021/ci400529h | J. Chem. Inf. Model. 2014, 54, 396−406

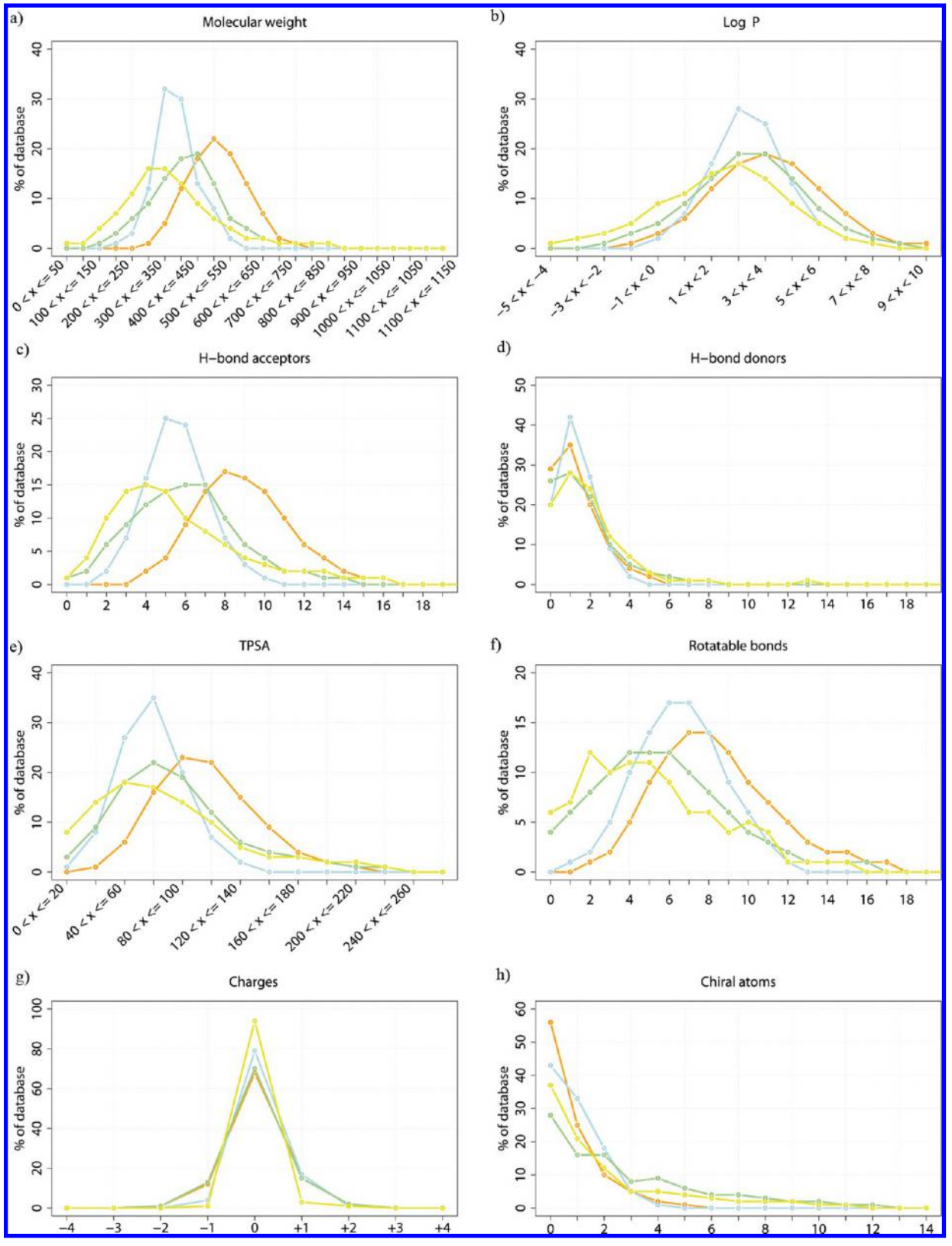


Figure 3. Comparison of the distributed properties for the ZINClick (orange), ZINC (cyan), ZINC "Natural product" (green), and DrugBank (yellow). The line graphs represent the percentages of MW, log P, HBA, HBD, TPSA, rotB, charges, and chiral molecules, respectively, in the database.

The diversity of the ZINClick database was analyzed versus ZINC. Molecular similarity is a key prerequisite when assessing molecular diversity.[45] Many different techniques can be used to measure whether two compounds are similar;[46] fingerprint methods are simple and commonly applied. The chemfp[36] software was used to generate a fingerprint from the GraphSim Python Toolkits[38] of OpenEye. The SUBSET algorithm[40] selected 2323 compounds using the calculated fingerprints.

dx.doi.org/10.1021/ci400529h | J. Chem. Inf. Model. 2014, 54, 396−406

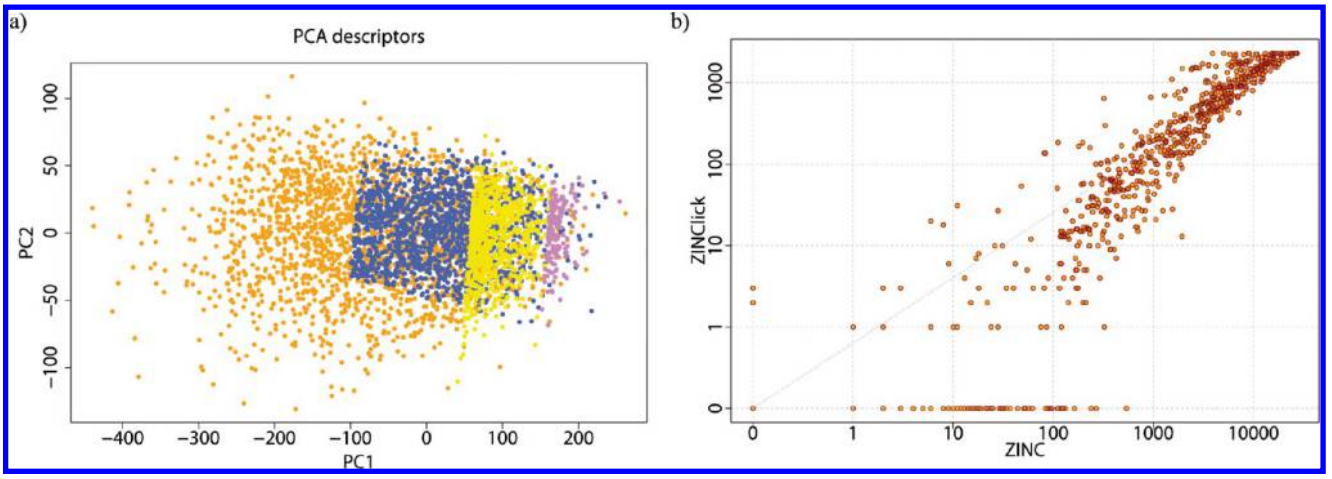


Figure 4. Diversity of ZINClick versus ZINC. (a) A principal component analysis (PCA) plot comparing the chemical space defined by the ZINClick databases: all compounds (orange), the "drug-like" subset (blue), the "lead-like" subset (yellow), and the "fragment-like" subset (violet). (b) Diversity of ZINClick versus ZINC. The substructural fingerprints were analyzed in each compound from both the ZINC and ZINClick diversity subsets using an 881-bit search. Each point in this logarithmic scatter plot corresponds to the number of compounds found with least one copy of the fingerprint in either the ZINC (*x*-axis) or ZINClick diversity subsets (*y*-axis). To simplify the graphical representation, "null" is represented by "0.1".

	All	Drug-Like	Lead-Like	Fragment-Like	10%-set	Diversity-set
Name	ZINClick_all	ZINClick_DL	ZINClick_LL	ZINClick_FL	ZINClick_10S	ZINClick_DS
Size	16,222,262	4,140,042	744,235	22,104	1,622,226	2,323
Updated	January 2013	January 2013	January 2013	January 2013	January 2013	January 2013
	Properties	Properties	Properties	Properties	Properties	Properties
Files	SMILES	SMILES	SMILES	SMILES	SMILES	SMILES
	.SDF (3D)	.SDF (3D)	.SDF (3D)	.SDF (3D)	.SDF (3D)	.SDF (3D)
Comment / Citations	-	Lipinski, C. A. *J Pharmacol Toxicol Methods* 2000, *44*, 235-249.	Teague, S. J. et al. *Angew Chem Int Ed Engl* 1999, *38*, 3743-3748.	Carr, R. A. et al. *Drug Discov Today* 2005, *10*, 987-992.	-	-
Filtering Critieria	-	MW <= 500 and MW >= 150, logP <= 5, rotB <=7, PSAa < 150, HBD <= 5 and HBA <= 10	MW <= 350 and MW >= 250, logP <= 3.5, rotB <= 7	logP<=3.5, MW <=250 and rotB <= 5	random 10%	Tanimoto <= 0.7

Figure 5. Detailed view of the ZINClick Web page showing the available subsets.

This diversity subset contains representative compounds that possess a Tanimoto index below 0.7 with any selected and nonselected compounds in the database. When using this representative subset, a principal component analysis (PCA) of the chemical space represented by ZINClick was performed. The first two principal components accounted for 92.1% and 7.4% of the X-variance. Figure 4a shows that all of the ZINClick compounds are distributed across the entire chemical space. The projection of ZINC compounds in the PCA plot (Figure S4, Supporting Information) demonstrates that the two databases do not completely overlap, indicating that ZINClick explores new regions of chemical space. Figure 4a shows the four distinctive regions related to the principal ZINClick subset.

To explore the structural diversity in ZINClick, the diversity subset selection was applied to the ZINC database, generating a ZINC diversity subset containing 27,442 compounds. The structural fingerprints used during the selection process comprised 881 bits, each representing a discrete substructure. A table reporting how many molecules from ZINClick and ZINC contained the substructures is included in Table S4 of the Supporting Information. Of the 881 fingerprint substructure features, 737 appear in at least one ZINC molecule, and 649 appear in at least one ZINClick compound. In addition, 633 of these features are common to at least one molecule from ZINC and ZINClick. Although ZINClick contains two substructures that do not appear in ZINC, ZINC contains 81 substructures that do not appear in ZINClick. The distribution of the fingerprints is shown in Figure 4b. Each point in the figure represents a single substructural feature. Clearly ZINClick shares several substructure fingerprints with ZINC at a comparable frequency.

To investigate whether any ZINClick compounds can act against protein targets in vitro, a systematic search within the 16 million compounds was carried out using the open access ChEMBL v.16 databases.[47] This database contains published data regarding the biological activity (without discrimination between active or inactive compounds) of approximately 1.6 million small molecules (1.3 million compounds are reported as a chemical structure). In total, 320 1,4-disubstituted triazoles are present, demonstrating the potential for finding novel bioactive compounds using ZINClick; 320 triazoles represent only 0.002% of the ZINClick database, confirming that ZINClick represents an underexplored region of chemical space.

Finally, the presence of nonspecific frequent hitters within screening libraries was evaluated; this problem is commonly associated with false positives during screening campaigns.[48] The compound collections provided for virtual screening often include molecules that contain chemically reactive groups or other undesirable functionalities that may interfere with the HTS

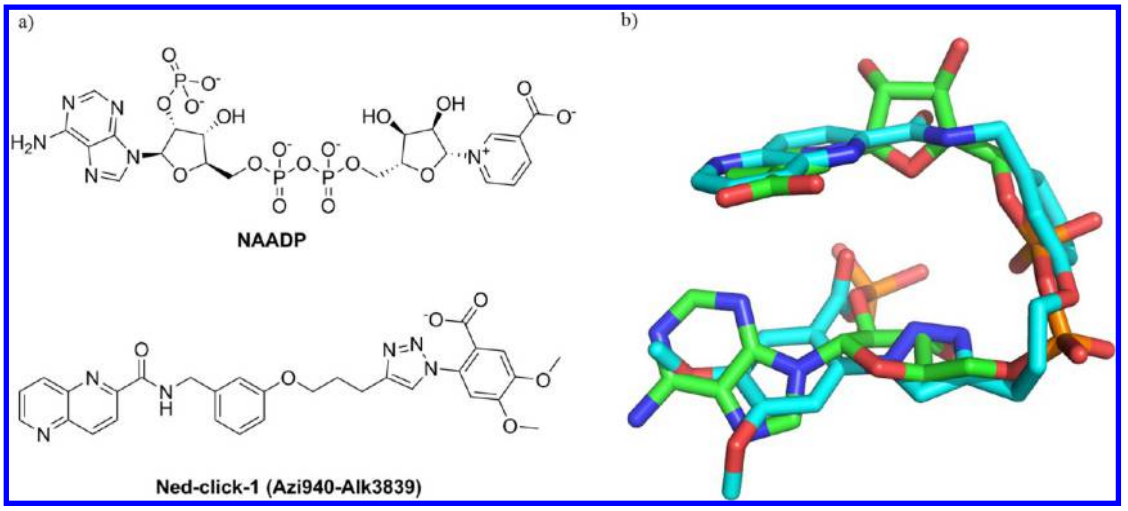
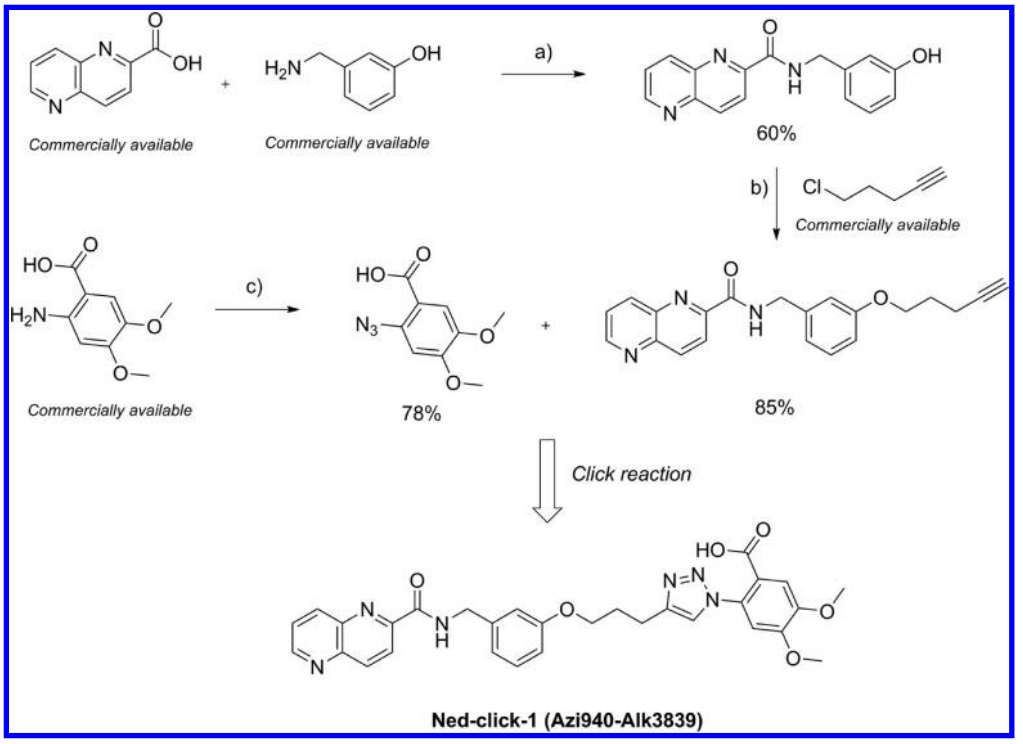


Figure 6. Top-scored compounds in the case studies. (a) Two-dimensional chemical structures of NAADP and Ned-click-1. (b) Overlay of NAADP (green carbons) and Ned-click-1 (cyan carbons).

Scheme 1. Synthesis of Compound Ned-click-1[a]



[a]Reagents and conditions: (a) EDC, DMF, 3 h, r.t., (b) NaH, DMF, 2 h at 80 °C than 4 h at 100 °C, and (c) t-BuONO, Me3SiN3, acetonitrile, 0 °C; 2 h, r.t.

detection techniques; these compounds are often referred to as PAINS (pan-assay interfering substances) or frequent hitters. We applied the PAINS substructure filters published by Baell and Holloway[41] to flag any compounds present in the ZINClick database with PAINS structural motifs. ZINClick contains only eight different PAINS scaffolds (Table S3 and Figure S5, Supporting Information). In total, 1,281,422 (8%) compounds have at least one PAINS substructure. In the current version of ZINClick, these molecules were annotated but not removed, warning the operator while allowing him/her complete control of the use of the ZINClick database.

3.1. Availability and Future Directions. The ZINClick database is freely available online via our Web site (http://www.symech.it/ZINClick). The current database is version 13; the number refers to the year of release (January 2013). The database is updated with azides and alkynes up the end of December 2012. A new version will be released every January.

The large size of the ZINClick database restricts users from downloading the entire database in 3D multiconformer format, but all 16 million structures and numerous subsets (drug-like, lead-like, fragment-like, diversity, and 10% subset) are freely available in the canonical SMILES and/or sdf formats (minimized 3D structure) (Figure 5). The files can be downloaded in a compressed format; they are also immediately available for docking or virtual screening studies. Currently, a basic search functionality (by ZINClick id, physical properties,

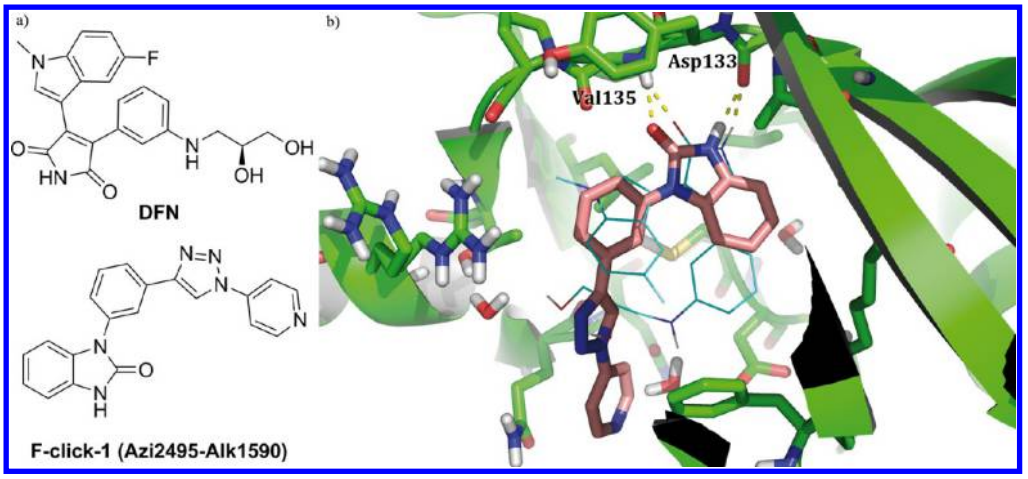
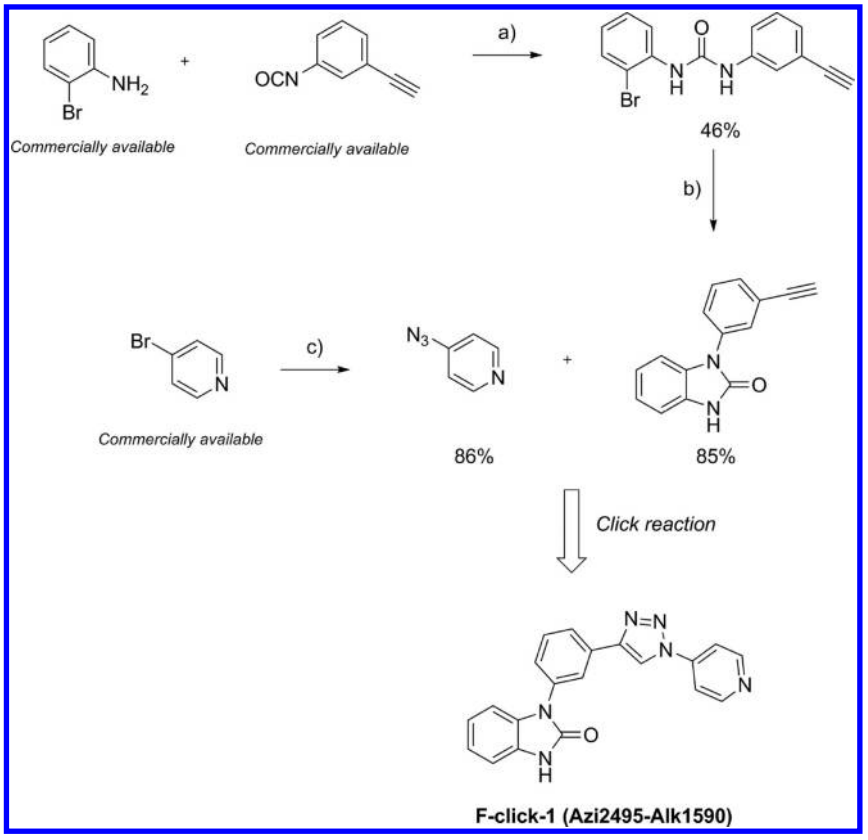


Figure 7. Best predicted GSK-3β inhibitor. (a) Two-dimensional chemical structures of DFN and F-click-1. (b) Docked structure of F-click-1 (pink sticks) in the GSK-3β binding site (green cartoon), with the crystal structure DFN superposed as cyan lines.

Scheme 2. Synthesis of Compound F-click-1[a]



[a]Reagents and conditions: (a) triphosgene, TEA, CH2Cl2, 20 min, 120 °C, (b) DBU, CuI, (S)-Proline, DMSO, 20 min, 120 °C, and (c) NaOH, NaN3, H2O, EtOH, r.t.; 2 h, 110 °C.

and chemical structure) is available for searching only the 10% subset.

4. PRACTICAL USE: CASE STUDIES

Although, it is beyond of the scope of this work to test and compare the performances of commonly used programs associated with virtual screening, we replicated some recent studies to illustrate the potential utility of the ZINClick database.

4.1. Identification of NAADP Analogues. NAADP (nicotinic acid adenine dinucleotide phosphate) is a Ca^{2+}-releasing second messenger that acts in various organisms ranging from plants to mammals. Reportedly, this molecule interacts through an independent receptor that differs from the others involved in Ca^{2+} release from storage. Furthermore, classical chemical agents able to induce Ca^{2+} release do not affect NAADP-induced Ca^{2+} release. Therefore, a small organic molecule able to mimic NAADP is an interesting opportunity for medicinal chemists. In 2009, Naylor et al. identified an NAADP antagonist through a three-dimensional shape and electrostatic map-based virtual screening using the ZINC database.[49] After following the reported protocol, we performed a three-dimensional shape comparison between NAADP and the

dx.doi.org/10.1021/ci400529h | J. Chem. Inf. Model. 2014, 54, 396−406

molecules in the ZINClick library using ROCS[50] to calculate an overlay of the two molecules. The compounds selected via ROCS were ranked using their shape Tanimoto score. The top 500 "hits" from ZINClick had a Tanimoto score between 0.69 and 0.77; in comparison, the values from ZINC were 0.66 and 0.73. We reranked the top 500 compounds for similarity to NAADP based on the electrostatic Tanimoto score while using the EON program.[51] The top 500 "hits" had an electrostatic Tanimoto score between 0.00 and 0.64, whereas the values of the original paper were −0.31 and 0.85. We selected the top 10 Nrd-click hits (NAADP discovered by ROCS) and the top 15 Ned-click hits (NAADP discovered by EON). The details regarding the selected hits are reported in Table S5 of the Supporting Information. Remarkably, the best compounds selected by ROCS show a shape Tanimoto score comparable to or even better than the best compounds reported by Naylor, while the electrostatic Tanimoto scores are lower. The best compound was Ned-click-1, as depicted in Figure 6, as it showed the best overlay with NAADP. This compound was never reported and can be prepared in four synthetic steps (Scheme 1).

4.2. Identification of Glycogen Synthase Kinase 3β Inhibitors. The second case study uses ZINClick to identify novel glycogen synthase kinase 3β (GSK-3β) inhibitors. GSK-3β is a major protein kinase that is involved in the regulation of glucose metabolism, attracting significant attention as a therapeutic target. In a recent study conducted by Osolodkin et al.,[54] a cross-docking study was used to identify the X-ray structure and combination of constraints leading to the best reproduction of the binding positions for different inhibitors. While following their inputs, the crystal structure of Glycogen Synthase Kinase 3β (PDB id: 1R0E) was used, and two constraints were applied due the presence of hydrogen bonds between the ligand and hinge residues (Asp133, backbone carbonyl oxygen atom, and Val135, backbone amide hydrogen atom). FRED[55] software was used with the Chemgauss3 scoring function. The details for the 10 best hits (F-click) are reported in Table S6 of the Supporting Information. The predicted binding pose of the best-scoring ligand is plausible (Figure 7), and the binding affinities should be −15.18 kcal/mol. This compound has never been reported and can be prepared in two synthetic steps (Scheme 2). Studies to validate these two hypotheses are ongoing in our laboratory, and the results will be reported in due course.

5. CONCLUSIONS

As discussed above, one problem that cannot be underestimated with virtual screening is that most molecules present in several databases are not new, and therefore, many research groups can discover the same molecules. Furthermore, the chemical inaccessibility of some active compounds is limited due to their long and/or difficult synthesis, frustrating the operator.

In this manuscript, we have constructed a database containing 16 million 1,4-disubstituted triazoles that are new, patentable, and synthetically feasible. The chemistry used to create the required azides and alkynes has already been published, and the final click chemistry reaction is robust and tolerant of numerous functional groups, failing only in rare cases.[16d] Finally, we used our ZINClick database with two previously published cases for comparison.

This database is freely downloadable online via our Web site (http://www.symech.it/ZINClick), and it will be enriched every year with novel azides and alkynes reported in the literature. We encourage others to bring azides and alkynes (MW < 500 Da, synthesizable using no more than three synthetic steps) to our attention. We hope this novel database can inspire drug hunters and find routine use by medicinal and computational chemists searching for novel hits for their biological targets. Even if only one successful case is reported, we could consider our work successful.

■ ASSOCIATED CONTENT

Supporting Information

Tables listing the molecular descriptors considered in the present work, a table listing the PAINS structural motifs that were present in ZINClick, and figures showing the distribution of molecular descriptors for all the reference databases. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*Phone: +39.0321.375.753 (A.M.). Fax: +39.0321.375.621 (A.M.). E-mail: alberto.massarotti@unipmn.it (A.M.).
*Phone: +39.0321.375.857 (G.C.T.). Fax: +39.0321.375.621 (G.C.T.). E-mail: giancesare.tron@unipmn.it (G.C.T.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.M. and G.S. gratefully acknowledge OpenEye Scientific Software, Inc. for providing an academic license for their software packages. Financial support from the Università del Piemonte Orientale (Italy) and FIRB 2012 "Infiammazione e cancro: approcci innovativi basati su nanotecnologie" MIUR Italy is gratefully acknowledged. The authors thank NASA and The Hubble Heritage Team (STScI) for the free use of the picture adapted for use in the TCG (http://hubblesite.org/gallery/album/galaxy/pr1999041a/).

■ REFERENCES

(1) (a) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening—An overview. *Drug Discovery Today* **1998**, *3*, 160−178. (b) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862−865.

(2) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing pitfalls in virtual screening: A critical review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.

(3) (a) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169−178. (b) Schneider, G. Virtual screening: An endless staircase? *Nat. Rev. Drug Discovery* **2010**, *9*, 273−276.

(4) (a) Alvarez, J.; Shoichet, B. *Virtual Screening in Drug Discovery*, 1 ed.; CRC Press: Boca Raton, FL, 2005; p 470. (b) Sotriffer, C. *Virtual Screening: Principles, Challenges, and Practical Guidelines*, 1 ed.; Wiley-VCH: Weinheim, DE, 2011; Vol. *48*, p 519.

(5) Chemistry Databases and Search Services on the Web. http://cactus.nci.nih.gov/links/chem_www.html (accessed August 31, 2013).

(6) Irwin, J. J.; Shoichet, B. K. ZINC—A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(7) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(8) Available Chemicals Directory (ACD). http://www.akosgmbh.de/Symyx/software/databases/acd.htm.

(9) Dorschner, K. V.; Toomey, D.; Brennan, M. P.; Heinemann, T.; Duffy, F. J.; Nolan, K. B.; Cox, D.; Adamo, M. F.; Chubb, A. J. TIN: A

combinatorial compound collection of synthetically feasible multicomponent synthesis products. *J. Chem. Inf. Model.* **2011**, *51*, 986−995.
(10) Tron, G. C. Off the beaten track: The use of secondary amines in the Ugi reaction. *Eur. J. Org. Chem.* **2013**, *2013*, 1849−1859.
(11) (a) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47*, 342−353. (b) Blum, L. C.; Reymond, J. L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732−8733. (c) 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. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(12) (a) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the reactions used for the preparation of drug candidate molecules. *Org. Biomol. Chem.* **2006**, *4*, 2337−2347. (b) Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54*, 3451−3479.
(13) Fullbeck, M.; Michalsky, E.; Dunkel, M.; Preissner, R. Natural products: Sources and databases. *Nat. Prod. Rep.* **2006**, *23*, 347−356.
(14) Chuprina, A.; Lukin, O.; Demoiseaux, R.; Buzko, A.; Shivanyuk, A. Drug- and lead-likeness, target class, and molecular diversity analysis of 7.9 million commercially available organic compounds provided by 29 suppliers. *J. Chem. Inf. Model.* **2010**, *50*, 470−479.
(15) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click chemistry: Diverse chemical function from a few good reactions. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 2004−2021.
(16) (a) Rostovtsev, V. V.; Green, L. G.; Fokin, V. V.; Sharpless, K. B. A stepwise huisgen cycloaddition process: Copper(I)-catalyzed regioselective "ligation" of azides and terminal alkynes. *Angew. Chem., Int. Ed. Engl.* **2002**, *41*, 2596−2599. (b) Tornoe, C. W.; Christensen, C.; Meldal, M. Peptidotriazoles on solid phase: [1,2,3]-Triazoles by regiospecific copper(i)-catalyzed 1,3-dipolar cycloadditions of terminal alkynes to azides. *J. Org. Chem.* **2002**, *67*, 3057−3064. (c) Wu, P.; Fokin, V. V. Catalytic azide−alkyne cycloaddition: Reactivity and applications. *Aldrichim. Acta* **2007**, *40*, 7−17. (d) Bock, V. D.; Hiemstra, H.; van Maarseveen, J. H. CuI-catalyzed alkyne−azide "click" cycloadditions from a mechanistic and synthetic perspective. *Eur. J. Org. Chem.* **2006**, *2006*, 51−68. (e) Meldal, M.; Tornoe, C. W. Cu-catalyzed azide-alkyne cycloaddition. *Chem. Rev.* **2008**, *108*, 2952−3015. (f) Hein, J. E.; Fokin, V. V. Copper-catalyzed azide-alkyne cycloaddition (CuAAC) and beyond: New reactivity of copper(I) acetylides. *Chem. Soc. Rev.* **2010**, *39*, 1302−1315.
(17) Sharpless, K. B. *Secret Life of Enzymes: An Aggressive Strategy for Drug Discovery*. Abstract of Papers, 229th ACS National Meeting, American Chemical Society: San Diego, CA, United States, 2005; pp ORGN-611.
(18) Rydzewski, R. M. *Real World Drug Discovery*. 1 ed.; Elsevier Ltd.: 2008.
(19) Hein, C. D.; Liu, X. M.; Wang, D. Click chemistry, a powerful tool for pharmaceutical sciences. *Pharm. Res.* **2008**, *25*, 2216−2230.
(20) (a) Kolb, H. C.; Sharpless, K. B. The growing impact of click chemistry on drug discovery. *Drug Discovery Today* **2003**, *8*, 1128−1137. (b) Moses, J. E.; Moorhouse, A. D. The growing applications of click chemistry. *Chem. Soc. Rev.* **2007**, *36*, 1249−1262. (c) Moorhouse, A. D.; Moses, J. E. Click chemistry and medicinal chemistry: A case of "cycloaddiction". *ChemMedChem* **2008**, *3*, 715−723. (d) Tron, G. C.; Pirali, T.; Billington, R. A.; Canonico, P. L.; Sorba, G.; Genazzani, A. A. Click chemistry reactions in medicinal chemistry: applications of the 1,3-dipolar cycloaddition between azides and alkynes. *Med. Res. Rev.* **2008**, *28*, 278−308. (e) Agalave, S. G.; Maujan, S. R.; Pore, V. S. Click chemistry: 1,2,3-Triazoles as pharmacophores. *Chem. Asian J.* **2011**, *6*, 2696−2718. (f) Thirumurugan, P.; Matosiuk, D.; Jozwiak, K. Click chemistry for drug development and diverse chemical−biology applications. *Chem. Rev.* **2013**, *113*, 4905−4979.
(21) Reactor, version 6.0.4, ChemAxon, 2013. http://www.chemaxon.com.
(22) Suite 2012: Maestro, version 9.3; Schrödinger, LLC: New York, 2012.
(23) Molecular Operating Environment (MOE), 2012.10; Chemical Computing Group Inc.: Montreal, QC, Canada, 2012.
(24) (a) e-LEA3D: ChemInformatic Tools and Databases. http://chemoinfo.ipmc.cnrs.fr/eDESIGN/reagent.html (accessed September 2, 2013). (b) Douguet, D. e-LEA3D: A computational-aided drug design Web server. *Nucleic Acids Res.* **2010**, *38*, W615−W621.
(25) Durrant, J. D.; McCammon, J. A. AutoClickChem: Click chemistry in silico. *PLoS Comput. Biol.* **2012**, *8*, e1002397.
(26) R Core Team. R: A Language and Environment for Statistical Computing, version 3.0.1; R Foundation for Statistical Computing: Vienna, Austria, 2013. http://www.R-project.org.
(27) The PyMOL Molecular Graphics System, version 1.3; Schrödinger LLC: 2010.
(28) SciFinder®. http://scifinder.cas.org.
(29) (a) Tubert-Brohman, I. Perl and chemistry. *Perl J.* **2004**, *8*, 3−5. (b) Tubert-Brohman, I. PerlMol—Perl modules for molecular chemistry, 2004. http://www.perlmol.org.
(30) (a) OMEGA, version 2.4.6; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com. (b) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584. (c) Hawkins, P. C. D.; Nicholls, A. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919−2936.
(31) FILTER, version 2.0.2; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.
(32) Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S. DrugBank 3.0: A comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.* **2011**, *39*, D1035−1041.
(33) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(34) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 3743−3748.
(35) Carr, R. A.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-based lead discovery: Leads by design. *Drug Discovery Today* **2005**, *10*, 987−992.
(36) chemfp, version 1.1; Dalke Scientific: Sweden, 2013. http://chemfp.com.
(37) OEChem TK, Python version 1.7.7; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.
(38) GraphSim TK, Python version 2.0.1; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com.
(39) ChemFP Substruct keys. http://code.google.com/p/chem-fingerprints/wiki/Substruct (accessed September 4, 2013).
(40) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(41) 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.
(42) Guha, R. PAINS Substructure Filters as SMARTS. http://blog.rguha.net/?p=850 (accessed July 4, 2013).
(43) In this way, the maximum number of synthetic steps required to synthesize the desired compounds cannot be more than seven.
(44) Two reasons drove us to use this cutoff: (a) the possibility to expand the chemical complexity of the starting materials and (b) to allow the operator to identify triazoles with similar features to the reference molecule under study. See case study 1.
(45) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.
(46) (a) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45*,

4350−4358. (b) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comp. Sci.* **1998**, *38*, 983−996.
(47) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput.-Aided. Mol. Des.* **2009**, *23*, 195−198.
(48) (a) Sink, R.; Gobec, S.; Pecar, S.; Zega, A. False positives in the early stages of drug discovery. *Curr. Med. Chem.* **2010**, *17*, 4231−4255. (b) Baell, J. B. Observations on screening-based research and some concerning trends in the literature. *Fut. Med. Chem.* **2010**, *2*, 1529−1546.
(49) Naylor, E.; Arredouani, A.; Vasudevan, S. R.; Lewis, A. M.; Parkesh, R.; Mizote, A.; Rosen, D.; Thomas, J. M.; Izumi, M.; Ganesan, A.; Galione, A.; Churchill, G. C. Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **2009**, *5*, 220−226.
(50) (a) ROCS, version 2.4.1; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com. (b) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.
(51) (a) EON, version 2.0.1; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com. (b) Muchmore, S. W.; Souers, A. J.; Akritopoulou-Zanze, I. The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chem. Biol. Drug Des.* **2006**, *67*, 174−176.
(52) Kubota, M.; Sakaguchi, H.; Kandoh, Y. Preparation of naphthyridinecarboxamides for plant disease control. Patent WO2009093640 (A1), 2009.
(53) Barral, K.; Moorhouse, A. D.; Moses, J. E. Efficient conversion of aromatic amines into azides: A one-pot synthesis of triazole linkages. *Org. Lett.* **2007**, *9*, 1809−1811.
(54) Osolodkin, D. I.; Palyulin, V. A.; Zefirov, N. S. Structure-based virtual screening of glycogen synthase kinase 3β inhibitors: Analysis of scoring functions applied to large true actives and decoy sets. *Chem. Biol. Drug Des.* **2011**, *78*, 378−390.
(55) (a) FRED, version 3.0.0; OpenEye Scientific Software: Santa Fe, NM. http://www.eyesopen.com. (b) McGann, M. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578−596.
(56) Li, Z.; Sun, H.; Jiang, H.; Liu, H. Copper-catalyzed intramolecular cyclization to N-substituted 1,3-dihydrobenzimidazol-2-ones. *Org. Lett.* **2008**, *10*, 3263−3266.
(57) Jia, Z.; Zhu, Q. 'Click' assembly of selective inhibitors for MAO-A. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 6222−6225.