

Correlation of binding site properties with chemistries used for generation of ultra-large virtual libraries.

Robert X. Song¹, Marc C. Nicklaus², Nadya I. Tarasova^{1}*

¹Cancer Innovation Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick Maryland 21702, USA

²Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, NIH, Frederick, Maryland 21702, USA

** Correspondence:*

Nadya.tarasova@nih.gov

.

SUBJECTS: Chemical reactions, virtual screens, protein pockets, druggability, virtual libraries, ultra-large databases, transforms

ABSTRACT

Although the size of virtual libraries of synthesizable compounds is growing rapidly, we are still enumerating only tiny fractions of the drug-like chemical universe. At the same time, our ability to mine these newly generated libraries also lags their growth. That is why fragment-based approaches that utilize on-demand virtual combinatorial libraries are gaining popularity. These *à la carte* libraries utilize synthetic blocks that have been shown to be effective binders in parts of target protein pockets. There is, however, no data on the potential impact of the chemistries used for making on-demand libraries on the hit rates during virtual screening. There are also no rules to guide in selection of these synthetic methods for libraries production. We have used the SAVI (Synthetically Accessible Virtual Inventory) library, constructed using 53 reliable reaction types (transforms), to test for correlations between these chemistries and docking hit rates for 39 well-characterized protein pockets. The data shows that the hit rate depends on the chemistry used and that chemistry selection can be optimized based on pocket properties.

Introduction

Screening of virtual libraries of synthesizable compounds has become an increasingly important step in drug discovery¹. The surge in utilization of computational approaches has been stimulated by improvements in binding energy calculations, the growth of computational resources, advances in protein structures determination and availability of large and diverse virtual libraries of compounds²⁻⁹. However, our ability to access the vast druggable chemical space is still limited and will be impacted by the availability of sufficient computing resources for the foreseeable future^{10, 11}. We have the potential of generating billions of virtual synthesizable molecules, but enumerating these chemical spaces, and thus converting them into screenable files is impractical.

That is why fragment-based approaches that enumerate only parts of the chemical spaces, thus generating on-demand virtual combinatorial libraries are widely used^{6, 12-14}. Most fragment-based methods identify synthetic blocks binding to sub-pocket(s) of a larger protein pocket first and generate libraries containing these blocks^{3, 6, 15}. We have recently generated and made publicly available the Synthetically Accessible Virtual Inventory (SAVI), which comprises nearly 1.75 billion virtual molecules, each with a proposed synthesis scheme. It was constructed from 155,129 building blocks provided by Enamine (Kyiv, Ukraine, enamine.net) using robust chemistries encoded in 53 transforms¹⁶. SAVI transforms were written into rules based on an adaptation and extension of the CHMTRN/PATRAN programming languages describing chemical synthesis expert knowledge¹⁷. We note the terminology used here: We call the general reaction type (typically a "named reaction") a "chemistry" in the context of SAVI, whereas the individual CHMTRN/PATRAN rules are called "transforms." For example, SAVI uses the Suzuki-Miyaura cross-coupling chemistry that expressed in 6 different transforms (bromo, iodo, alkene cross-coupling etc.). Transforms have a descriptive name but also a four-digit number, which will frequently be used in the following. All 53 transforms can be downloaded from https://cactus.nci.nih.gov/download/savi_download/savi_2020_transforms_src_and_clb.tar).

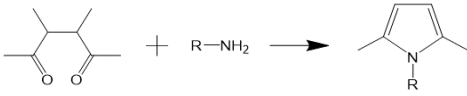
The cheminformatics toolkit CACTVS¹⁸ was used to apply these rules for the virtual synthesis of the entries. By now, 168 SAVI compounds have been synthesized. SAVI's predictions of synthetic accessibility were found to be 97.6% accurate. Enamine has a database called REAL, which could be called a "sister" of SAVI since it is constructed from essentially the same set of building blocks. The overlap between these two ultra-large databases is only about 10% because of the difference in chemistries applied for the generation of the entries¹⁶. We found significant differences in the number of hits when docking equally sized SAVI or REAL diversity sets into

the same protein pocket. For some pockets, the SAVI set was much more productive than the REAL one, whereas for others, REAL produced more hits than SAVI under the same docking conditions. Since these two databases are constructed from the same building blocks, and the major differences are in the chemistries used, we hypothesized that linking chemistries may favor certain pockets but not the others. We have now expanded the number of reliable chemistries that can be used for SAVI generation to more than 120. The number of commercially available synthesis building blocks has also increased. Enamine alone has now 680 million made-on-demand blocks (MADE) (<https://enamine.net/building-blocks/made-building-blocks>). Consequently, the next version of SAVI could have trillions of entries. The expansion of the accessible chemical space is a welcome trend that is likely to improve and accelerate drug discovery. However, enumeration of such databases and their use in their entirety is currently non-practical or outright impossible with the available computational resources. Although our computational capacities are likely to expand in the future, so are the accessible parts of the chemical universe. Enumeration of defined parts of chemical space (so-to-speak the "optimal chunks" of the space) is likely to continue to be widely used in the future as it allows to reduce screening efforts. Consequently, to enumerate the most appropriate part of the chemical space for a given target, it would be helpful to know not only which building blocks are better suited for the target pocket, but also which linking chemistries are more likely to generate high-scoring hits. To evaluate potential correlation between pocket properties and transforms used for library generation, we conducted docking of SAVI diversity set, which contains entries from all 53 transforms listed in Table 1, into 39 protein pockets (Table 2).

Results and Discussion

Target pockets (Table 2) have been selected from PDB to represent two types: small molecule (SM) pockets and protein-protein interaction (PPI) pockets. Majority of selected pockets bind well-characterized ligands that have advanced into the clinics. However, we also included several less studied but interesting and potentially impactful pockets that either are difficult to target or represent surfaces involved in protein-protein interactions because they are the types that scientific community is more likely to face in the future. To ensure that the structures were suitable for virtual screening, the ligands present in the chosen complexes have been redocked into the corresponding pockets, and only structures that showed favorable binding scores have been included in the analysis. For docking and virtual screens, we have used the ICM-Pro software (Molsoft). Although the software has been benchmarked before¹⁹⁻²¹, we have evaluated the correctness of docking poses for the pockets with known ligands (Table 2). Most of docked complexes had RMSD<1 when compared to the experimental structures. In two cases where it exceeded 1 (PDB:5i96 and 5vv0), all the differences were in the part of the molecules exposed to the solvent, while poses inside the pocket were determined with high accuracy. Testing binding properties for all identified hits was not possible in the context of this study. However, we were able to do this for eight targets, which are studied in the lab^{22, 23}.

Table 1. Docking hits rates for SAVI-2020 transforms applied in the generation of diversity set used for docking into 39 protein pockets. A hit was defined as a compound with a docking score below -32 as described in Methods.

ID	Name	Scheme	Hits rate, %	Number in the set	Number in SAVI
1031	Paal-Knorr Pyrolles synthesis		0.061	32785	65570

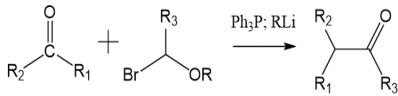
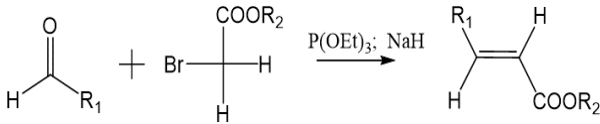
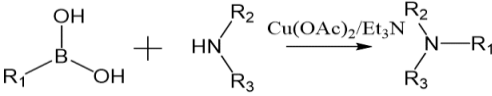
1039	Feist Synthesis of Pyrroles		0.127	1437	1437
1171	Hantzsch Thiazole Synthesis		0.157	9423	94336
1391	[2+ 2]-Cycloaddition of Allenes to Alkenes		0.000	20	20
1439	Pyrazoles from Beta Carbonyl Carboxylic Acid Derivatives		0.030	21137	42275
2201	Fused Arylpyridines via o-Aminocarboxyls		0.075	57453	582318
2218	Tetrazoles from Azide and Nitriles		0.107	4376	4376
2230	Phthalazin-1-ones from 2-Acylbenzoic Acids		0.352	22918	45836
2238	Fused Aryl(2,3-H/R)Pyridines (Pictet-Spengler)		0.021	90655	1827991
2267	Sonogashira Coupling		0.313	120752	24239698
2269	Kabbe Synthesis of 4-Chromanones		0.052	14642	146610

2630	Benzazepin-2-ones by Pictet-Spengler Reaction		0.094	10184	10184
2684	Benzo[b]furans from 2-Hydroxyphenyl Acetylenes		0.182	942	942
2875	Copper[I]-catalyzed azide-alkyne cycloaddition		0.119	59078	1208372
6003	Buchwald-Hartwig Ether Formation		0.396	86370	43731278
6004	Suzuki-Miyaura Cross-Coupling (Bromo)		0.465	56880	5803732
6005	Suzuki-Miyaura Cross-Coupling (Iodo)		0.369	39814	804723
6006	Suzuki-Miyaura Cross-Coupling (Chloro)		0.363	29010	2971512
6008	Suzuki-Miyaura Cross-Coupling with Alkene		0.238	24659	49318
6009	Suzuki-Miyaura Cross-Coupling of Alkenes		0.172	43535	876832

6013	Hiyama Aryl-Alkenyl Cross-Coupling		0.379	2966	2966
6014	Hiyama Non-Aromatic Cross-Coupling		0.206	8976	8976
6015	Hiyama Allyl Cross-Coupling		0.069	148	148
6016	Hiyama Carbonylative Cross-Coupling		0.433	12026	24052
6017	Hiyama Cross-Coupling with Arylhydrazine		0.125	1106	1106
6022	Liebeskind-Srogl Thioamide Coupling		0.201	9113	91767
6024	Liebeskind-Srogl Nitrile Formation		0.014	541	541
6025	Liebeskind-Srogl Heterocyclic Coupling		0.367	11642	116790
6026	Sulfonamide Schotten-Baumann		0.145	123951	124375067
6027	Sulfonamide Schotten-Baumann		0.137	66826	6803351

	from Sulfonate				
6028	Sulfonamide Schotten-Baumann from Thiol	$\text{Ar-SH} \xrightarrow{\text{H}_2\text{O}_2/\text{SOCl}_2} \text{ArSO}_2\text{Cl} + \begin{array}{c} \text{R}_1 \\ \\ \text{NH} \\ \\ \text{R}_2 \end{array} \longrightarrow \text{Ar-SO}_2\begin{array}{c} \text{R}_1 \\ \\ \text{N} \\ \\ \text{R}_2 \end{array}$	0.135	91691	91704439
6029	Sulfonamide Schotten-Baumann from Aryl Bromide	$\text{Ar-Br} \xrightarrow{\text{RMgBr, SO}_2} [\text{ArSO}_2\text{Cl}] + \begin{array}{c} \text{R}_1 \\ \\ \text{NH} \\ \\ \text{R}_2 \end{array} \longrightarrow \text{Ar-SO}_2\begin{array}{c} \text{R}_1 \\ \\ \text{N} \\ \\ \text{R}_2 \end{array}$	0.178	105957	211944731
6031	Mitsunobu Reaction	$\begin{array}{c} \text{R}_1 \\ \\ \text{C-OH} \\ \\ \text{R}_2 \end{array} + \text{RCO}_2\text{H} \xrightarrow{\text{PPh}_3/\text{DEAD}} \begin{array}{c} \text{R}_1 \\ \\ \text{C-O} \\ \\ \text{R}_2 \end{array} \text{C(=O)R}$	0.119	155673	155748444
6032	Mitsunobu carbon-carbon bond formation	$\begin{array}{c} \text{CO}_2\text{R}_1 \\ \\ \text{H}_2\text{C} \\ \\ \text{CO}_2\text{R}_1 \end{array} + \text{ROH} \xrightarrow{\text{PPh}_3/\text{DEAD}} \begin{array}{c} \text{CO}_2\text{R}_1 \\ \\ \text{R} \\ \\ \text{CO}_2\text{R}_1 \end{array}$	0.017	18139	181524
6033	Mitsunobu SN2' Reaction	$\begin{array}{c} \text{R}_1 \\ \\ \text{C}=\text{CH}_2 \\ \\ \text{HO} \end{array} + \text{R}_3\text{CO}_2\text{H} \xrightarrow{\text{PPh}_3/\text{DEAD}/\text{Et}_3\text{N}} \begin{array}{c} \text{COOR}_2 \quad \text{R}_3 \\ \quad \quad \\ \text{C}=\text{C} \\ \\ \text{R}_1 \end{array}$	0.248	8368	83940
6034	Mitsunobu Imide Reaction	$\begin{array}{c} \text{O} \\ \\ \text{---NH} \\ \\ \text{O} \end{array} + \begin{array}{c} \text{R}_1 \\ \\ \text{R}_2\text{-C-OH} \\ \\ \text{R}_3 \end{array} \xrightarrow{\text{PPh}_3/\text{DEAD}} \begin{array}{c} \text{O} \\ \\ \text{---N} \\ \\ \text{O} \end{array} \begin{array}{c} \text{R}_1 \\ \\ \text{R}_2\text{-C} \\ \\ \text{R}_3 \end{array}$	0.068	132730	27177967
6035	Mitsunobu Aryl Ether Formation	$\text{Ar-OH} + \text{R-OH} \xrightarrow{\text{PPh}_3/\text{DEAD}} \text{R-O-Ar}$	0.086	84542	42306237
6036	Mitsunobu Sulfonamide Reaction	$\begin{array}{c} \text{O} \\ \\ \text{R}_2\text{-S-NH} \\ \\ \text{R}_1 \end{array} + \text{R}_3\text{-OH} \xrightarrow{\text{PPh}_3/\text{DEAD}} \begin{array}{c} \text{R}_3 \\ \\ \text{N-S} \\ \\ \text{R}_1 \end{array} \begin{array}{c} \text{O} \\ \\ \text{R}_2 \end{array}$	0.080	104307	10589664
6038	Ester or Amide or Thiolester Formation	$\text{R}_1\text{-O(S)H} + \begin{array}{c} \text{O} \\ \\ \text{R-C-OH} \end{array} \xrightarrow{\text{DCC}} \begin{array}{c} \text{O} \\ \\ \text{R-C-O} \\ \\ \text{(S)O-R}_1 \end{array}$	0.132	183070	366293581
6039	Williamson Ether Synthesis	$\text{R}_1\text{OH} + \text{R}_2\text{-Cl, Br, I} \longrightarrow \text{R}_1\text{-O-R}_2$	0.127	103046	103177836

6041	Buchwald-Hartwig Reaction - Amines		0.321	132160	264514821
6043	Buchwald-Hartwig Reaction - Sulfonamides		0.365	160097	32762479
7005	Benzimidazoles from o-Phenylenediamines		0.097	85452	1733461
7009	Acylsulfonamide from Sulfonamide and Carboxylic Acid		0.187	92318	46207962
7013	Benzimidazoles from o-Phenylenediamines and Aldehydes		0.165	77938	1575305
7014	Benzimidazoles from o-Phenylenediamines and Aldehydes		0.284	43989	888165
7015	Sulfonamide from sulfonic acid and amine		0.118	47678	4856868
7017	Sulfonamide alkylation with a cyclic ether		0.137	36416	3732596
7018	Sulfonamide acylation		0.045	29975	300300
7019	Wittig Reaction		0.077	142425	142522022

7020	Wittig via Methoxy-Ylide		0.070	11557	11557
7021	Horner-Wadsworth-Emmons Olefination		0.254	15922	31843
7022	Chan-Lam coupling		0.062	128600	26186137

We chose the SAVI diversity set containing 2,955,416 compounds for the exploration because of practical considerations. Docking the entire SAVI database into just one pocket would take more than 280 days when running 1000 parallel processes on the NIH supercomputer cluster. Docking of the diversity set into one pocket requires around 25,000 CPU Hours, which is doable on a computer cluster. Although docking of larger sets may allow for more sensitive detection of differences between different transforms, it would require prohibitively large computational resources when used for multiple pockets, which we aimed to evaluate for this study.

Average hits rates across 39 targets differed significantly between different transforms (Figure 1, Table 1). Several transforms had to be excluded from further analysis because they are represented by too few compounds in SAVI as well as in the diversity dataset. This underrepresentation occurs due to the low number of available of synthetic blocks that are needed for these transforms. These “starved” transforms included Feist synthesis of pyrroles (1039), [2+2]-cycloaddition of allenes to alkenes (1391), benzo[b]furans synthesis from 2-hydroxyphenyl acetylenes (2684), Hiyama allyl cross-coupling (6015), Hiyama cross-coupling with arylhydrazine (6017) and Liebeskind-Srogl nitrile formation (6024) (Table 1). Several transforms had sufficient representation in the database but could not be used for reliable evaluation because they produced too few hits across all tested targets and zero hits for many of them. The weak performance of

some of these transforms can be attributed to poor availability of one of the two building blocks needed. Although instances of the second type of blocks needed could be plentiful in the building block set and the number of generated compounds therefore relatively large, the overall diversity of the products is limited, which may be the reason for lower number of the hits in screens. Transforms that had to be excluded for this reason were Paal-Knorr pyrroles synthesis (1031), pyrazoles synthesis from beta carbonyl carboxylic acid derivatives (1439), fused arylpyridines via o-aminocarbonyls (2201), Kabbe synthesis of 4-chromanones (2269), Mitsunobu carbon-carbon bond formation (6032), and Wittig via methoxy-ylide (7020). The performance of these could be improved in the future by increasing the number of 1,4-diketones for 1031, 2-keto esters for 1439, o-acyl anilines for 2201, o-acyl phenols for 2269, esters of malonic acid for 6032 and 1-bromo ethers for 7020. Synthesis of fused aryl(2,3-H/R) pyridines by Pictet-Spengler reaction (2238), Mitsunobu imide reaction (6034) and sulfonamide acylation (7018) produced too few hits, and thus, may be less valuable for the current drug discovery efforts (Figure 1, Table 1). Remarkably, several chemistries produced subsets with very high hit rates. Suzuki-Miyaura cross-couplings (6004, 6005 and 6006) were among the most productive ones. Interestingly, Suzuki-Miyaura coupling is among the most frequently used chemistries in current medicinal chemistry²⁴. Our data shows that this chemistry deserves the attention it receives. However, the most frequently used reaction, amide bond formation²⁴ (transform 6038), was less productive with a hit rate that was about three times lower than that for Suzuki-Miyaura cross-couplings. Our data suggest those transforms that deserve additional efforts in expanding. For example, Hiyama carbonylative cross-coupling should be expanded by adding more aryl triethoxysilanes into the collection of the synthesis blocks. Expanding the collection of arylboronic acids would benefit not only Suzuki-Miyaura cross-coupling, but also the highly productive Liebeskind-Srogl heterocyclic coupling

(6025). It should be noted that the efficacy of a transform in producing potential hits can depend not only on the properties/geometry of the bond it generates but also on reaction selectivity and diversity of the building blocks available. Selectivity of the reaction allows to preserve the functional groups of the blocks that can be beneficial for protein binding while diversity increases the chances of finding a good fit for a particular pocket. Transforms 6004-6006 produce structurally similar di-aryl compounds through Suzuki-Miyaura cross-coupling. However, the hit rates for 6004, which uses bromo aryl blocks, is about 27% higher than for either 6005 or 6006 that use iodo and chloro aryls. The set of building blocks used in for SAVI-2020 has a 7.3 times higher number of bromo aromatic compounds than iodo-derivatives, thus allowing for higher diversity in the products of transform 6004 compared to 6005. Chloro-aromatic blocks are even more numerous than the bromo-derivatives. However, the reaction is less selective for chloro compounds, which results in a 44 times higher number of excluded products, effectively reducing the number of useful blocks for transform 6006. The diversity of the blocks that can potentially impact the hit rates is likely to change with time along additional synthetic efforts in building blocks generation. Thus, the hit rates can be improved for less productive transforms in the future.

Pocket properties analyzed included volume, area, radius, hydrophobicity, nonsphericity, aromaticity, buriedness, drug-like density (DLID²⁵), the numbers of hydrogen bonds donors, and the number of acceptors. The hydrogen bond forming potential of each pocket has been evaluated manually. The rest of the parameters were determined using the PocketFinder function of ICM-Pro. All these properties were correlated with the number of hits for each transform and entire SAVI diversity set.

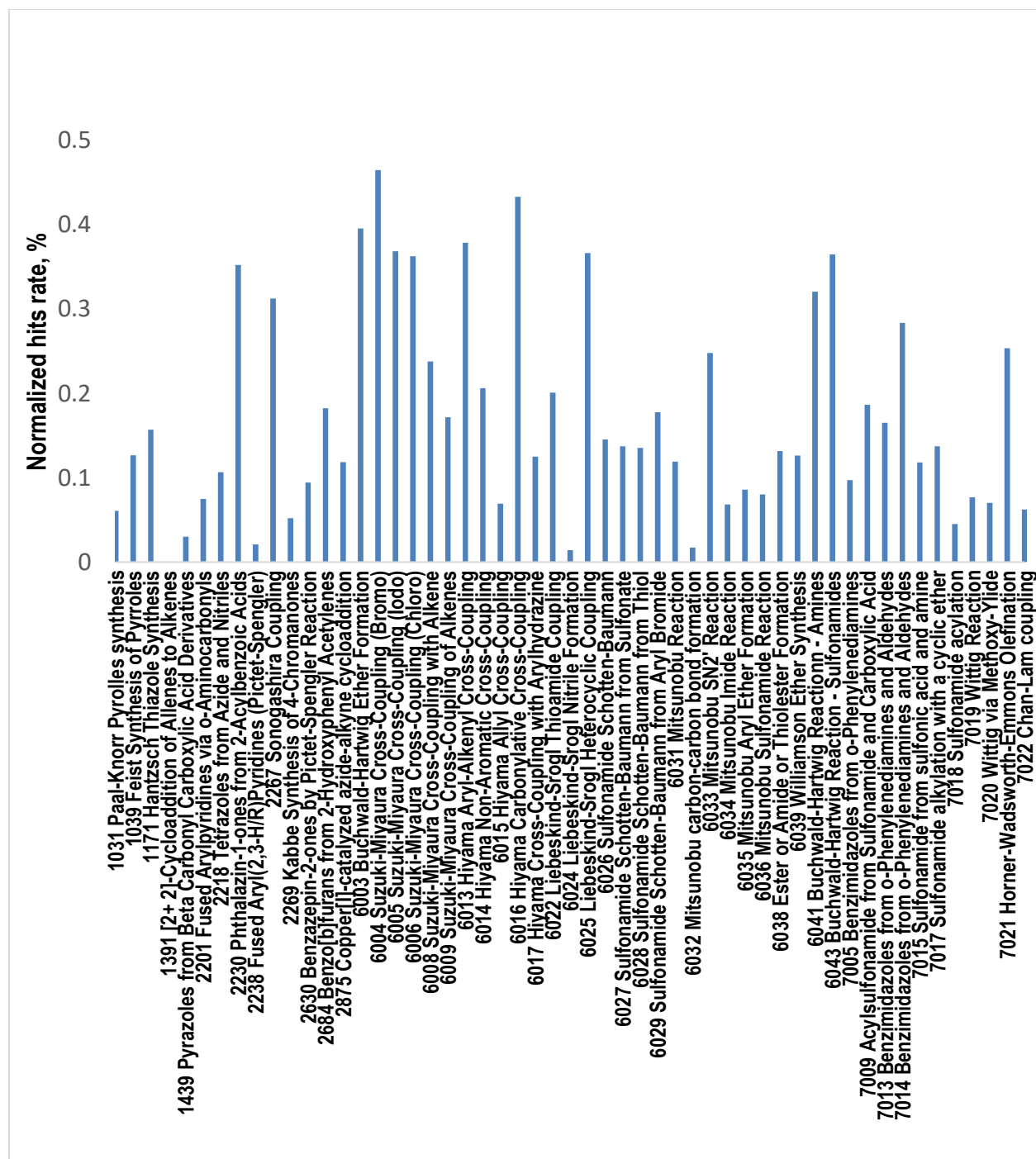


Figure 1. Hit rates for 53 transforms used for SAVI generation. The hits were identified by docking 2,955,416 compounds of SAVI diversity set into 39 well characterized protein pockets. To compensate for differences in representation of a particular transform in the diversity set, total number of hits has been normalized by dividing by the number of compounds produced by the transform in the screening library.

Table 2. Protein targets used in the study had two pocket types: small molecules druggable pockets (SM) and protein-protein interaction interfaces (PPI). Addition of a capital letter to the PDB ID (such as "A", "B") denotes the protein subunit/chain used.

PDB	Target	Pocket Type	Volume, Å ³	DLID	Number of hits**	RMSD, Å***
1sj0	ESR1	SM	598	1.23	9818	0.4869
3kl6	FA10	SM	424	0.16	5244	0.0027
5dwr	PIM1	SM	764	0.63	1109	0.2978
3ruk	CP17A	SM	683	1.73	774	0.0000
6gt3	A2A	SM	771	1.56	14387	0.4042
3odu	CXCR4	SM	619.1	0.8	84	0.9068
4mbs	CCR5	SM	523.1	0.15	111	0.3636
3lpb	JAK2	SM	1064	1.22	1911	0.6140
2owb	PLK1	SM	859	1.06	66418	0.6446
7khk	KIT	SM	469.3	0.55	11147	0.6820
5i96	IDHP	SM	451.9	1.45	430	1.1244
5ef8	HDAC6	SM	325	0.64	279	0.9170
4tvj	PARP2	SM	792.8	0.57	7541	0.5388
5fhz	ALDH1A3	SM	1060	1.41	11031	0.8559
2oj9	IGF1R	SM	640.2	0.31	892	0.5842
4xe0	PK3CD	SM	296.8	0.38	145	0.0001
3d4q	BRAF	SM	741.4	0.73	13715	0.0000
5vv0	NOS1	SM	707.5	0.64	236	1.1333
6tz7	calcineurin	SM	925.8	0.5	13889	0.7967
5kj2	p300	SM	599.9	0.76	204	0.1866
4ivd	JAK1	SM	1210	0.99	6063	0.5382
5gmh	TLR7	SM	596.8	0.8	3414	0.3937
4ixd	ITGAL	SM	413	-0.1	93	0.6677
1qw6	NOS1	SM	315	0.04	4	0.3305
4ziaB	STAT3 ND	PPI	561	-0.5	1295	
SARS-CoV2 Spike						
6m0jA	Protein	PPI	474.7	0.38	349	
4lvt	BCL-2	PPI	572.6	0.27	1971	0.4819
5lof	MCL1	PPI	307.9	0.76	40	0.8268
5v52	TIGIT	PPI	108	-0.7	326	
5wlb	K-Ras	PPI	339.9	0.18	11213	
5wha	K-Ras	PPI	450	0.58	6102	
6dhb	TIM-3	PPI	297.1	-0.5	163	
5v1y ²²	Rpn13	PPI	328	0.13	4284	
4lwv	MDM2	PPI	289	0.19	78	0.4179
4mr4	BRD4	PPI	316.7	0.19	2163	0.0000

4lxd	BCL-2	PPI	132.1	-0.6	375	0.4201
6h6q	XIAP	PPI	291.3	-0.2	1	0.4291
6o5i	MEN1	PPI	949	0.33	229	0.3858
7p5e	KEAP1	PPI	1007	0.68	850	0.3728

*DLID: Drug-like density²⁵

**Number of hits obtained by virtual screening of 2,955,416 compounds of SAVI diversity set.

***Ligands present in the structures of the complexes were docked into the corresponding protein pocket and the docking pose was compared to the experimental structure.

The binding score produced by docking for every molecule is influenced by many factors. That is why we did not expect strong dependencies for any single parameter, but rather tendencies. For the whole database, the number of hits showed a statistically significant positive correlation (with p -value < 0.05) with properties related to pocket size: volume, radius, and surface (Figure 2). Most of the pockets with high numbers of hits had volumes between 300 and 1000 Å³, and hit rates were significantly lower both below and above this range. Similarly, the graphs suggest that the most productive values are between 4 and 6.2 Å for the radius and between 300 and 900 Å² for the pocket surface area. This can be explained by the size distribution of the entries in the database entries as it contains small numbers of molecules with $MW < 200$ and > 550 ¹⁶. The degree of hydrophobicity of the pocket did not yield any definite trends. Surprisingly, aromaticity appeared to have negative correlation, although aromatic interactions have been suggested to contribute to ligand-protein binding^{26, 27}. However, the correlation was not statistically significant.

Nonsphericity and buriedness demonstrated positive correlation with the number of hits (Figure 2) but it was statistically insignificant for both parameters. The number of hydrogen bond acceptors (HBA) in the pocket did not show any significant correlation. In contrast, the number of hydrogen bond donors (HBD) appeared to have significant positive correlation with the number of docking hits (Figure 3). The observed dependencies on HBD could be caused by prefiltering of the database

building blocks for “drug-like” properties. Hydrogen bond acceptors of potential drugs are widely believed to be less detrimental than hydrogen-bond donors for solubility, cell permeability and bioavailability²⁸. Lipinski’s rule of 5 is more restrictive to hydrogen bond donors than to hydrogen bond acceptors allowing no more than 5 of HBDs and no more than 10 HBAs²⁹. Consequently, the database will have more HBA-rich compounds that prefer HBD-rich pockets.

To compare the degrees of dependencies for different transforms, we used correlation coefficients (Tables 3 and S1). Correlations with pockets’ properties differ for different transforms (Table 3) and frequently have opposite signs. The relatively small number of pockets screened does not allow one to make statistically justified conclusions for many correlations as p-values fall short, sometimes just slightly short of 0.05. The data shows that those differences do exist, and additional future screens will permit to establish comprehensive correlations. Nevertheless, several dependencies could be established. Pocket volume and area showed positive correlations with the hit rate for all transforms with transforms 6003, 7013 and 7014 showing the strongest correlations. Although the number of hits increased with an increase of pocket buriedness and nonsphericity for the majority of transforms, only transform 7005 had statistically significant correlation with buriedness. Transform 7005 makes benzimidazoles from o-phenylenediamines and carboxylic acids. Interestingly, two other transforms that produce benzimidazoles, 7013 and 7014 (Table 1) also show relatively high correlation with buriedness suggesting that benzimidazoles could be particularly suited for pockets well shielded from the solvent.

Aromaticity had negative or very low positive, but insignificant correlation for all transforms except for 7005, which had a strong negative correlation with $r=-0.35$ (Table 3) suggesting that benzimidazoles should be avoided for targeting pockets with many aromatic residues. Although the number of hydrogen bond acceptors in the pocket did not show any definite correlation for the

whole diversity set, it demonstrated strong positive correlation for transforms 2630 (benzazepin-2-ones by Pictet-Spengler reaction), 2875 (copper[I]-catalyzed azide-alkyne cycloaddition) and 6009 (Suzuki-Miyaura cross-coupling of alkenes). Transforms 2630 and 2875 produce heterocycles with hydrogen-donating properties that can explain this trend. For transform 6009, the reason could be the properties of the blocks that it utilizes as the newly formed C=C double bond does not have any hydrogens to donate. The number of hydrogen bond donors appeared to have positive correlation with the number of hits for all transforms except for 6036 (Mitsunobu sulfonamide reaction), which had an insignificant negative correlation. The strongest correlations were found for transforms 6016 (Hiyama carbonylative cross-coupling), 6035 (Mitsunobu aryl ether formation), 2267 (Sonogashira Coupling), 7021 (Horner-Wadsworth-Emmons olefination), 6041 (Buchwald-Hartwig reaction of amines), and 7009 (acylsulfonamides from sulfonamides and carboxylic acids). Hydrogen bonds are strong contributors to the binding energy. Thus, hydrogen-forming capacity of the pocket can be expected to have a positive effect on the number of hits. However, as discussed before, prefiltering of the building blocks for “drug-like” properties, which excludes hydrogen-bond donor-rich compounds to avoid cell permeability and bioavailability issues, limits the number of HBD-rich compounds making the observed dependences less pronounced. The hydrogen bond forming capacity of a transform can be impacted by reaction selectivity. For example, both transform 7013 and 7014 generate benzimidazoles from aromatic o-diamines and aldehydes. However, 7014 uses boric acid to produce a reactive intermediate while 7013 uses molecular iodine under basic conditions. Consequently, the sets of restrictions for the starting blocks are different. As a result, 7013 generated almost twice as many compounds as 7014, but has a significantly lower overall hit rate (Table 1). Nevertheless, the correlations with pocket properties are similar for these two transforms. All observed trends can assist in generation of

optimally targeted virtual libraries. Our data suggest that with expanding number of synthetically accessible building blocks, the efforts in enumeration of virtual libraries will benefit from focusing on cross coupling reactions such as Sonogashira, Suzuki-Miyaura, Hiyama and Liebeskind-Srogl coupling as they produce the highest numbers of hits. Interestingly, the DLID (drug-like density) descriptor had a positive correlation with the number of hits, but the correlation was statistically insignificant for the entire diversity set and all transforms except 7005, 7013 and 7014. These three transforms produce benzimidazoles. The results suggest that low druggability score, in its traditional definition²⁵, although being a useful parameter, shouldn't discourage one from attempting virtual screens for a particular pocket as exceptions to the rule are not uncommon.

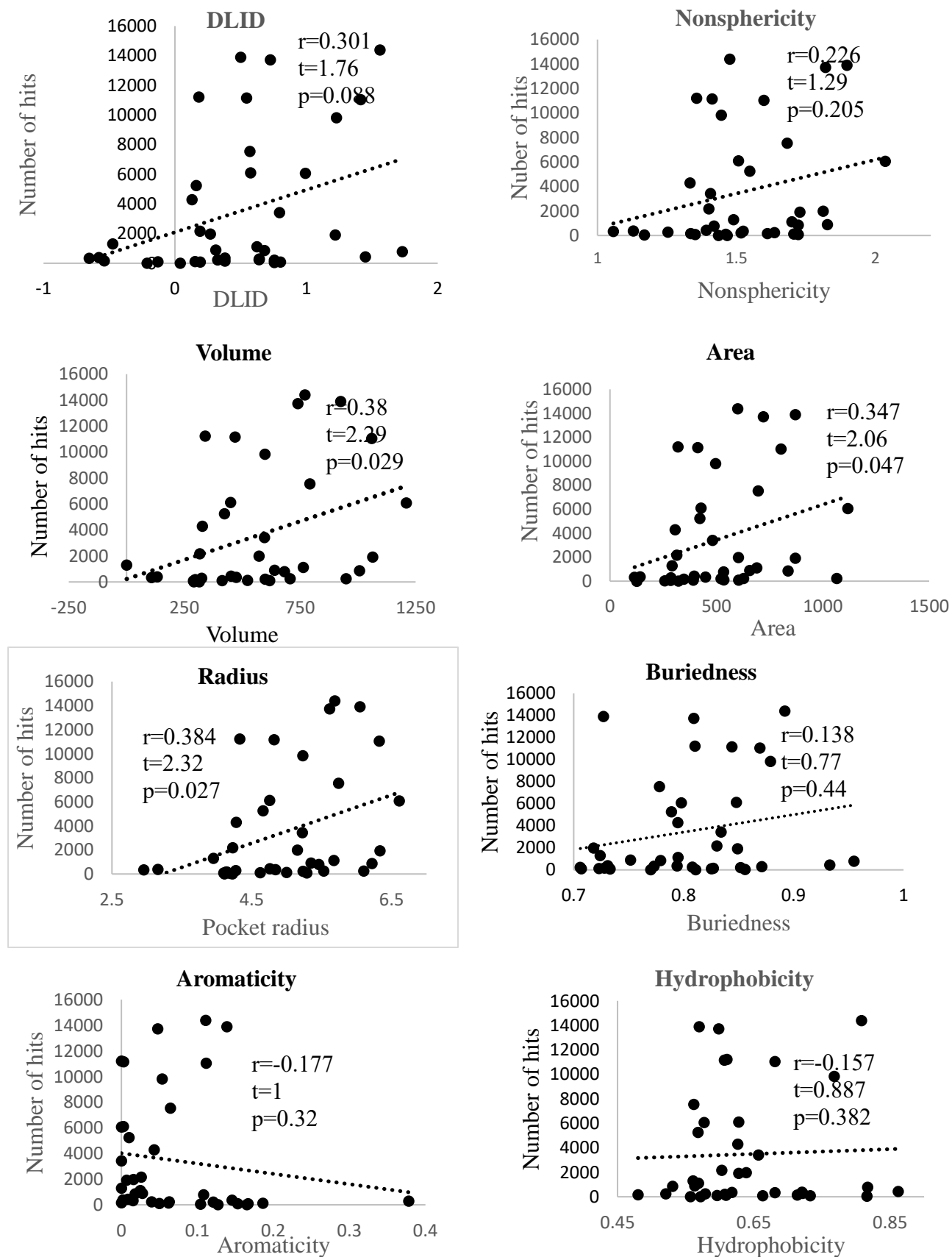


Figure 2. Total number of hits generated by virtual docking of SAVI diversity set into protein

pockets with different properties. The parameters for each property have been determined using PocketFinder function of ICM-Pro software (Molsoft). Punctate lines represent linear trends with corresponding correlation coefficient (r), Student's t -distribution and p -values shown.

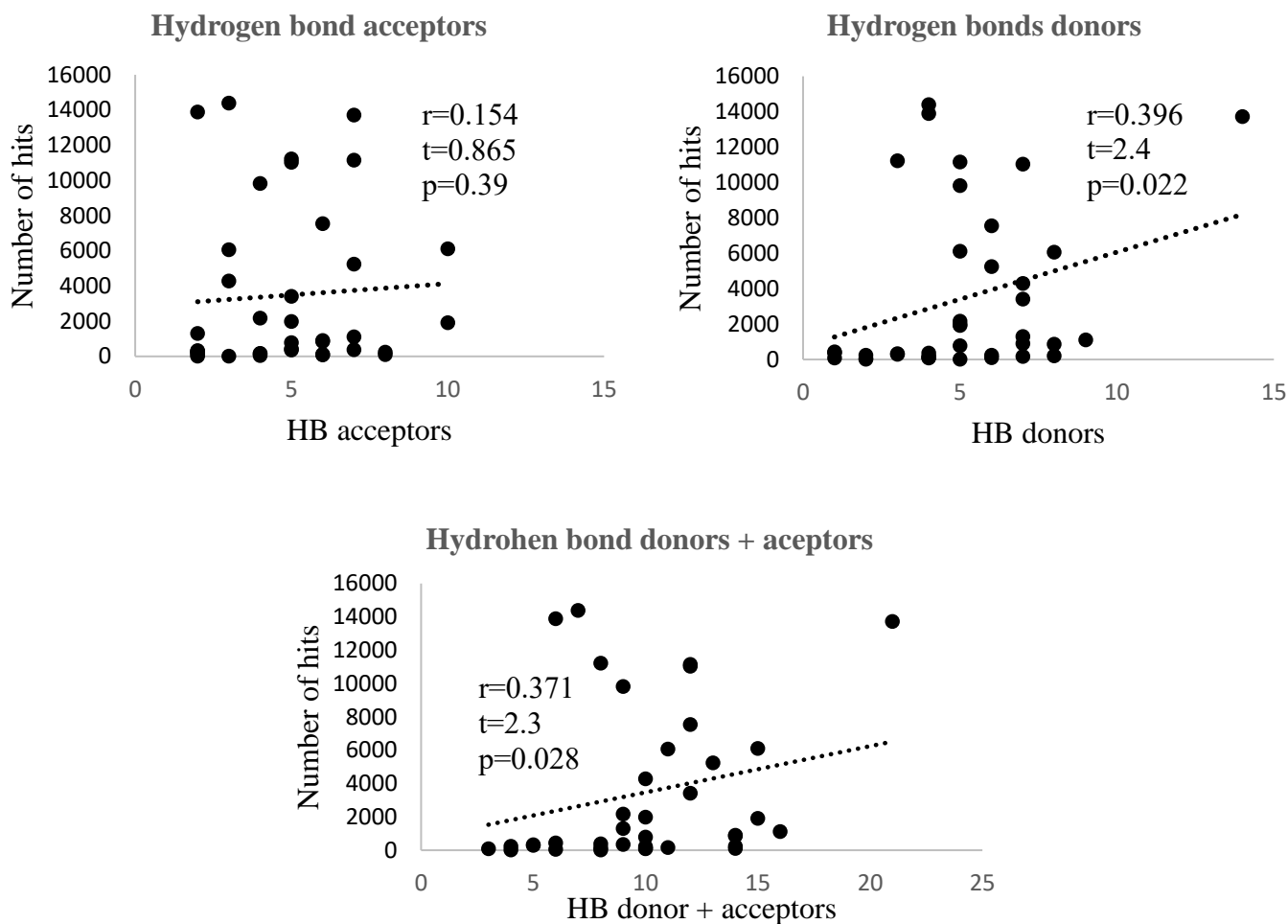


Figure 3. The impact of hydrogen bonds-forming capacity of the pockets on the total number of hits for entire SAVI diversity set. The number of hydrogen bond donors had a positive statistically significant correlation with the number of docking hits.

Table 3. Pearson’s coefficients for correlations between protein pocket properties and the number of docking hits in SAVI diversity set. Statistically significant correlations with p-values below 0.05 are highlighted in yellow. Table S1 contains the full set of data with t- and p-values included.

ID	Volume	Radius	Area	DLID	Buriedness	Nonsphericity	Aromaticity	HB donors	HB acceptors	Hydrophobicity
2230	0.334	0.24	0.33	0.105	-0.16	0.27	0.06	0.06	-0.21	-0.15
2267	0.289	0.3	0.255	0.289	0.23	0.157	-0.23	0.545	0.265	-0.12
2630	0.149	0.184	0.164	0.196	0.22	0.08	-0.27	0.365	0.366	-0.14
2875	0.149	0.184	0.164	0.196	0.22	0.08	-0.27	0.365	0.366	-0.14
6003	0.427	0.416	0.235	0.298	0.08	0.233	-0.07	0.345	0.062	-0.151
6004	0.349	0.341	0.287	0.322	0.239	0.16	-0.21	0.419	0.199	-0.097
6005	0.338	0.35	0.18	0.32	0.198	0.16	-0.21	0.43	0.22	-0.1
6006	0.334	0.33	0.29	0.3	0.18	0.17	-1.8	0.37	0.16	-0.1
6008	0.21	0.22	0.174	0.238	0.21	0.1	-0.2	0.46	0.3	-0.09
6009	0.11	0.14	0.09	0.172	0.21	0.05	-0.23	0.27	0.42	-0.1
6016	0.26	-0.02	0.21	0.27	0.15	0.1	-0.11	0.61	0.21	-0.12
6026	0.39	0.38	0.35	0.246	0.039	0.227	0.044	0.255	-0.09	-0.118
6027	0.359	0.354	0.365	0.178	-0.085	0.278	-0.02	0.258	-0.149	-0.166
6028	0.283	0.285	0.281	0.128	-0.082	0.219	0.023	0.144	-0.103	-0.159
6029	0.336	0.34	0.313	0.21	0.013	0.209	-0.05	0.322	-0.017	-0.164
6031	0.116	0.139	0.103	0.171	0.23	0.069	-0.28	0.331	0.326	-0.134
6034	0.27	0.106	0.232	0.22	0.056	0.129	-0.13	0.397	0.04	-0.09
6035	0.289	0.171	0.27	0.241	0.101	0.2	-0.17	0.56	0.194	-0.13
6036	0.174	0.161	0.171	0.044	-0.13	0.13	-0.07	-0.04	-0.092	-0.128
6038	0.222	0.241	0.221	0.175	0.107	0.169	-0.22	0.281	0.232	-0.186
6039	0.248	0.262	0.2	0.252	0.2	0.1	-0.23	0.354	0.266	-0.13
6041	0.31	0.322	0.288	0.269	0.179	0.191	-0.25	0.523	0.194	-0.171
6043	0.32	0.315	0.275	0.229	0.056	0.14	0.096	0.235	0.06	-0.101
7005	0.393	0.302	0.359	0.455	0.35	0.217	-0.34	0.258	0.283	0.048
7009	0.34	0.225	0.318	0.258	0.065	0.225	-0.11	0.482	0.11	-0.126
7013	0.446	0.425	0.375	0.404	0.227	0.21	-0.05	0.396	0.094	-0.038
7014	0.416	0.389	0.331	0.408	0.273	0.153	-0.06	0.345	0.11	-2E-04
7015	0.341	0.333	0.269	0.174	-0.034	0.272	-0.09	0.279	-0.08	-0.179
7017	0.383	0.373	0.334	0.295	0.114	0.189	-0.13	0.375	0.045	-0.092
7019	0.232	0.228	0.252	0.249	-0.01	0.216	-0.14	0.412	0.041	-0.144
7021	0.323	0.332	0.263	0.336	0.243	0.148	-0.13	0.535	0.276	0.08
7022	0.303	0.129	0.279	0.232	0.064	0.186	-0.12	0.395	0.071	-0.122

Experimental Methods.

Databases.

The SAVI diversity set of 2,955,416 compounds was generated from entire SAVI-2020 database, which contain 1,748,464,003 compounds using mini-batch k-means clustering performed with RDKit (<https://www.rdkit.org/>) and scikit-learn (<https://scikit-learn.org/stable/>). The Tanimoto coefficient for any two compounds in the set was <0.6. The entire SAVI database and diversity sets are available for downloading from the SAVI download page:

https://cactus.nci.nih.gov/download/savi_download/.

Database docking.

Docking screens were conducted using the ICM-Pro software (Molsoft L.L.C., San Diego, CA) by running 590 parallel processes (5000 compounds per job) on 590 cores of the National Institutes of Health (NIH) Biowulf cluster supercomputer. Each core contained 2 CPUs. The PocketFinder software (Molsoft) was used for the identification of the pockets. Screens were run in large-scale parallel way as so-called "swarm" jobs. The cutoff score was set to -32 for all docking runs. Hits were extracted as Excel files. Every compound in the SAVI database has an identifier (SAVI ID) with its last four digits indicating the transform number. These numbers were used for counting the hits produced by every transform. Correlation coefficients, Student's t-distribution and p-values were determined using the Data Analysis function of Excel (Microsoft).

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ACKNOWLEDGMENT

This work utilized the computational resources of the NIH HPC Biowulf cluster

(<http://hpc.nih.gov>). We thank Megan L. Peach for generation of the SAVI diversity sets. This research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, CCR, CIL.

REFERENCES

1. Nazarova, A. L.; Katritch, V., It all clicks together: In silico drug discovery becoming mainstream. *Clin Transl Med* **2022**, *12* (4), e766.
2. Beroza, P.; Crawford, J. J.; Ganichkin, O.; Gendele, L.; Harris, S. F.; Klein, R.; Miu, A.; Steinbacher, S.; Klingler, F. M.; Lemmen, C., Chemical space docking enables large-scale structure-based virtual screening to discover ROCK1 kinase inhibitors. *Nat Commun* **2022**, *13* (1), 6447.
3. Muller, J.; Klein, R.; Tarkhanova, O.; Gryniukova, A.; Borysko, P.; Merkl, S.; Ruf, M.; Neumann, A.; Gastreich, M.; Moroz, Y. S.; Klebe, G.; Glinca, S., Magnet for the Needle in Haystack: "Crystal Structure First" Fragment Hits Unlock Active Chemical Matter Using Targeted Exploration of Vast Chemical Spaces. *J Med Chem* **2022**, *65* (23), 15663-15678.
4. Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S., Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23* (11), 101681.
5. Lyu, J.; Wang, S.; Balias, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Alga, E.; Tolmacheva, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J., Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566* (7743), 224-229.
6. Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X. P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V., Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601* (7893), 452-459.
7. Gahbauer, S.; Correy, G. J.; Schuller, M.; Ferla, M. P.; Doruk, Y. U.; Rachman, M.; Wu, T.; Diolaiti, M.; Wang, S.; Neitz, R. J.; Fearon, D.; Radchenko, D. S.; Moroz, Y. S.; Irwin, J. J.; Renslo, A. R.; Taylor, J. C.; Gestwicki, J. E.; von Delft, F.; Ashworth, A.; Ahel, I.; Shoichet, B. K.; Fraser, J. S., Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proc Natl Acad Sci U S A* **2023**, *120* (2), e2212931120.
8. Perebyinis, M.; Rognan, D., Overlap of On-demand Ultra-large Combinatorial Spaces with On-the-shelf Drug-like Libraries. *Mol Inform* **2023**, *42* (1), e2200163.
9. Danel, T.; Leski, J.; Podlowska, S.; Podolak, I. T., Docking-based generative approaches in the search for new drug candidates. *Drug Discov Today* **2023**, *28* (2), 103439.
10. Lyu, J.; Irwin, J. J.; Shoichet, B. K., Modeling the expansion of virtual screening libraries. *Nat Chem Biol* **2023**.

11. Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M., Exploration of Ultralarge Compound Collections for Drug Discovery. *J Chem Inf Model* **2022**, *62* (9), 2021-2034.
12. Andrianov, G. V.; Gabriel Ong, W. J.; Serebriiskii, I.; Karanicolas, J., Efficient Hit-to-Lead Searching of Kinase Inhibitor Chemical Space via Computational Fragment Merging. *J Chem Inf Model* **2021**, *61* (12), 5967-5987.
13. Zhou, H.; Cao, H.; Skolnick, J., FRAGSITE: A Fragment-Based Approach for Virtual Ligand Screening. *J Chem Inf Model* **2021**, *61* (4), 2074-2089.
14. Meyenburg, C.; Dolfus, U.; Briem, H.; Rarey, M., Galileo: Three-dimensional searching in large combinatorial fragment spaces on the example of pharmacophores. *J Comput Aided Mol Des* **2023**, *37* (1), 1-16.
15. Galyan, S. M.; Ewald, C. Y.; Jalencas, X.; Masrani, S.; Meral, S.; Mestres, J., Fragment-based virtual screening identifies a first-in-class preclinical drug candidate for Huntington's disease. *Sci Rep* **2022**, *12* (1), 19642.
16. Patel, H.; Ihlenfeldt, W. D.; Judson, P. N.; Moroz, Y. S.; Pevzner, Y.; Peach, M. L.; Delannee, V.; Tarasova, N. I.; Nicklaus, M. C., SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci Data* **2020**, *7* (1), 384.
17. Judson, P. N.; Ihlenfeldt, W. D.; Patel, H.; Delannee, V.; Tarasova, N.; Nicklaus, M. C., Adapting CHMTRN (CHeMistry TRaNslator) for a New Use. *J Chem Inf Model* **2020**, *60* (7), 3336-3341.
18. Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S., Computation and management of chemical properties in CACTVS: An extensible networked approach toward modularity and compatibility. *Journal of Chemical Information and Computer Sciences* **1994**, *34* (1), 109-116.
19. Lam, P. C.; Abagyan, R.; Totrov, M., Hybrid receptor structure/ligand-based docking and activity prediction in ICM: development and evaluation in D3R Grand Challenge 3. *J Comput Aided Mol Des* **2019**, *33* (1), 35-46.
20. Lam, P. C.; Abagyan, R.; Totrov, M., Macrocyclic modeling in ICM: benchmarking and evaluation in D3R Grand Challenge 4. *J Comput Aided Mol Des* **2019**, *33* (12), 1057-1069.
21. Scarpino, A.; Ferenczy, G. G.; Keseru, G. M., Comparative Evaluation of Covalent Docking Tools. *J Chem Inf Model* **2018**, *58* (7), 1441-1458.
22. Lu, X.; Sabbasani, V. R.; Osei-Amponsa, V.; Evans, C. N.; King, J. C.; Tarasov, S. G.; Dyba, M.; Das, S.; Chan, K. C.; Schwieters, C. D.; Choudhari, S.; Fromont, C.; Zhao, Y.; Tran, B.; Chen, X.; Matsuo, H.; Andresson, T.; Chari, R.; Swenson, R. E.; Tarasova, N. I.; Walters, K. J., Structure-guided bifunctional molecules hit a DEUBAD-lacking hRpn13 species upregulated in multiple myeloma. *Nat Commun* **2021**, *12* (1), 7318.
23. Andrade Bonilla P, H. C., Stefanisko K, Tarasov S, Sinha S, Nicklaus M, Tarasova NI, Virtual screening of ultra-large chemical libraries identifies cell-permeable small-molecule inhibitors of a "non-druggable" target, STAT3 N-terminal domain *ChemRxiv* **2013**.
24. Brown, D. G.; Bostrom, J., Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J Med Chem* **2016**, *59* (10), 4443-58.
25. Sheridan, R. P.; Maiorov, V. N.; Holloway, M. K.; Cornell, W. D.; Gao, Y. D., Drug-like density: a method of quantifying the "bindability" of a protein target based on a very large set of pockets and drug-like ligands from the Protein Data Bank. *J Chem Inf Model* **2010**, *50* (11), 2029-40.
26. Brylinski, M., Aromatic interactions at the ligand-protein interface: Implications for the development of docking scoring functions. *Chem Biol Drug Des* **2018**, *91* (2), 380-390.

27. Li, S.; Xu, Y.; Shen, Q.; Liu, X.; Lu, J.; Chen, Y.; Lu, T.; Luo, C.; Luo, X.; Zheng, M.; Jiang, H., Non-covalent interactions with aromatic rings: current understanding and implications for rational drug design. *Curr Pharm Des* **2013**, *19* (36), 6522-33.
28. Kenny, P. W., Hydrogen-Bond Donors in Drug Design. *J Med Chem* **2022**, *65* (21), 14261-14275.
29. Lipinski, C. A., Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* **2004**, *1* (4), 337-41.

TOC:

