


1 Received: 2 March 2017 | Revised: xxxx | Accepted: 8 April 2017

DOI: 10.1111/cbdd.13012

RESEARCH ARTICLE

WILEY 7 Documenting and harnessing the biological potential of molecules
in distributed drug discovery (D3) virtual catalogs10 Milata M. Abraham¹ | Ryan E. Denton¹ | Richard W. Harper¹ | William L. Scott¹12 2 | Martin J. O'Donnell¹ | Jacob D. Durrant² 15 ¹Department of Chemistry and Chemical
Biology, Indiana University Purdue
University Indianapolis, Indianapolis, IN,
USA18 ²Department of Biological
Sciences, University of Pittsburgh,
Pittsburgh, PA, USA

Correspondence

22 Jacob D. Durrant, Department of Biological
Sciences, University of Pittsburgh,
Pittsburgh, Pennsylvania, USA.25 Email: durrantj@pitt.edu26 Martin J. O'Donnell, Department of
Chemistry and Chemical Biology, Indiana
University Purdue University Indianapolis,
Indianapolis, Indiana, USA.29 Email: modonnel@iupui.edu

Funding information

31 National Science Foundation, Grant/
Award Number: DUE-1140602, MRI-
CHE-0619254 and MRI-DBI-0821661;
National Institutes of Health, Grant/
Award Number: RO1-GM28193; IUPUI
International Development Fund; Indiana
University Purdue University Indianapolis
(IUPUI) STEM Summer Scholars Institute

1 | INTRODUCTION

42 Virtual compound catalogs^[1–4]/libraries^[5–9] can be useful
43 sources of candidate molecules for drug discovery. However,
44 two major factors often limit their utility: (i) Member com-
45 pounds and analogs may be difficult to synthesize^[10] or (ii)
46 there is no inherent probability of biological activity. To the
47 first issue, we have created large virtual catalogs of *N*-acyl
48 α -amino acids and their methyl ester and primary amide de-
49 rivatives based on reproducible procedures and demonstrated
50 student syntheses of individual catalog members **1**^[2,4], **2**,^[3]
51 and **3**^[11] (presented here for the first time, Figure 1).52 The purpose of the current work is to address the second
53 issue by assessing the existing or potential biological activity

Virtual molecular catalogs have limited utility if member compounds are (i) difficult to synthesize or (ii) unlikely to have biological activity. The Distributed Drug Discovery (D3) program addresses the synthesis challenge by providing scientists with a free virtual D3 catalog of 73,024 easy-to-synthesize *N*-acyl unnatural α -amino acids, their methyl esters, and primary amides. The remaining challenge is to document and exploit the bioactivity potential of these compounds. In the current work, a search process is described that retrospectively identifies all virtual D3 compounds classified as bioactive hits in PubChem-cataloged experimental assays. The results provide insight into the broad range of drug-target classes amenable to inhibition and/or agonism by D3-accessible molecules. To encourage computer-aided drug discovery centered on these compounds, a publicly available virtual database of D3 molecules prepared for use with popular computer docking programs is also presented.

KEYWORDS

in silico chemoinformatics, molecular modeling, peptidomimetic, small molecule diversity, structure-based drug design, synthetic methods, unnatural amino acid derivatives, virtual screening

of these compounds through a chemoinformatics analysis, and providing an example of D3-centered computer-aided drug discovery. The pharmacological activity of virtual D3 molecules can be demonstrated in two ways: prospectively by synthesizing and testing catalog members in biological assays, or retrospectively by searching the literature and other databases for previously documented activity (Fig. SI-1).

The prospective approach was first pursued as an initial proof of concept (Fig. SI-1, left arrow). A small number of D3 compounds **1** were synthesized and submitted to the NIH Molecular Libraries Initiative for testing in established high-throughput screens,^[12] resulting in hits in disparate NIH bioassays. Further investigation of one of these hits (CID 971933, AID 132162) identified a published analog, also present in

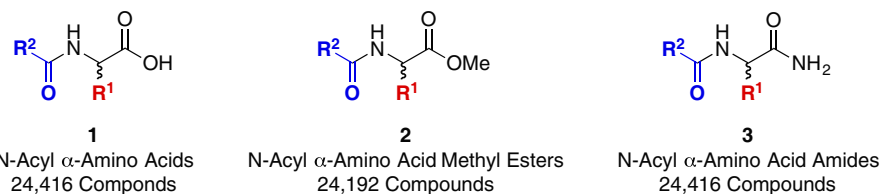


FIGURE 1 Generic structures of the 73,024 unnatural and natural *N*-acyl α -amino acid derivatives in the D3 virtual catalog

8 the D3 virtual catalog, with documented biological activity.
9 That analog had been key to the development of nateglinide,
10 a marketed drug to treat type 2 diabetes mellitus.^[13,14]

11 Inspired by this suggestive finding, a retrospective search
12 (Fig. SI-1, right arrow) was undertaken to identify other exam-
13 ples of virtual catalog members with documented biological
14 activity. In this article, we identify all enumerated D3 unnatu-
15 ral and natural *N*-acyl α -amino acids (**1**), *N*-acyl α -amino acid
16 methyl esters (**2**), and *N*-acyl α -amino acid amides (**3**) that are
17 listed among the active compounds in the PubChem database
18 of >1 million bioassays (AID, Assay ID). By unnatural, we
19 mean compounds with D (~~normally R~~) stereochemistry and/or
20 a non-proteinogenic side chain, versus natural compounds with
21 L (~~normally S~~) stereochemistry and/or a proteinogenic side
22 chain. Members of this list are termed “D3/PubChem actives.”
23 The hits from this search give credibility to the prospective
24 drug discovery potential of the D3 catalog and suggest areas
25 for future discovery research. Within the D3/PubChem actives
26 list, relevant PubChem screens that specifically target single
27 proteins are also categorized (Fig. SI-2). These protein targets
28 then offer the opportunity to construct structure-based compu-
29 tational models for predictive analysis of D3 virtual catalogs.

30 In addition to this cheminformatics analysis, a new
31 publicly available database of three-dimensional D3 struc-
32 tures for use in virtual screening projects, available free of
33 charge from http://durrantlab.com/liglib/iupui/d3_docking/
34 (accessed February 27, 2017), is described. To demonstrate
35 its utility, this database was used in a virtual screen against
36 peptidyl-prolyl cis-trans isomerase NIMA-interacting 1
37 (Pin1), a potential target for future cancer chemotherapies.^[15]
38 This analysis predicted the biological activity of known com-
39 pounds and suggested molecules with the potential for even
40 greater activity. The methodology presented should assist
41 others in the field who similarly wish to harness the power of
42 enumerated compound catalogs in drug discovery.

45 2 | METHODS AND MATERIALS

46 2.1 | Enumeration protocol

48 Virtual catalogs of 73,024 unnatural or natural *N*-acyl α -
49 amino acids (**1**; 24,416 compounds), *N*-acyl α -amino acid
50 esters (**2**; 24,192 compounds), and *N*-acyl α -amino acid
51 amides (**3**; 24,416 compounds), called the “D3 *N*-Acyl α -
52 Amino Acid,” “D3 *N*-Acyl α -Amino Acid Ester,” and “D3
53 α -*N*-Acyl Amino Acid Amide” sets, respectively, were

enumerated from diversity sets of commercially available
electrophiles and carboxylic acids (100 members each)^[2]
using CombiChem for Excel^a or, more recently, ChemAxon,^b
Issues of stereochemistry were addressed in this enumeration
to yield compounds with two to eight stereoisomers (see Fig.
SI-4, and associated text for a detailed discussion).

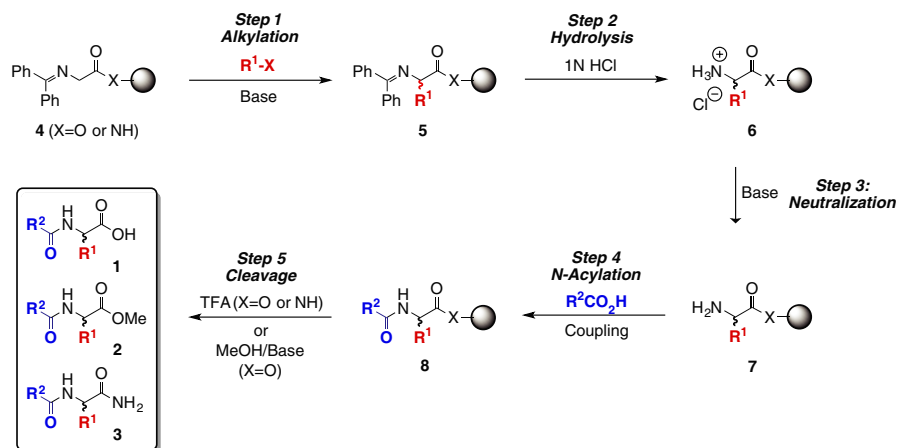
2.2 | PubChem/D3 searching protocol

A series of Boolean substructure searches was used to identify
all D3-like *N*-acyl α -amino acid derivatives (**4**, **5**, or **6**) and
N-acyl dipeptide derivatives (**7**, **8**, or **9**) present in PubChem
(i.e., “PC-CID” compounds in Figs. SI-3, SI-4, SI-5, and SI-
6). See the Supporting Information for full Boolean search
details.

2.3 | Calculating molecular properties

Schrodinger’s Maestro Suite^c was used to calculate the chemi-
cal properties of the compounds in the D3 *N*-Acyl α -Amino
Acid set, NCI Diversity Set III,^[16] and ChemBridge Diversity
CombiSet^[17,18] libraries, which contain 24,416, 1,597, and
30,000 compounds, respectively. One electrically neutral,
three-dimensional structure of each compound was created
using Schrodinger’s LigPrep module,^d as required to calcu-
late molecular properties. Molecular geometries were opti-
mized using the OPLS_2005 force field,^[19] hydrogen atoms
were added or removed as required to achieve electrical neu-
trality (e.g., carboxylates were protonated, protonated amines
were deprotonated, etc.), salts were removed, only the most
probable tautomer was considered, and one low-energy con-
formation was generated for each ring. The D3 input files ex-
plicitly specify the stereochemistry at each chiral center; for
compounds from other virtual catalogs that lack specified chi-
ralities, a single chiral molecule was created per input struc-
ture, chosen from among the many possible enantiomers that
could have been enumerated. Once these 3D structures had
been created, Schrodinger’s QikProp module^e was used to
calculate the molecular weight, predicted LogP, and number
of hydrogen-bond donors and acceptors for each compound.

Occasionally, the Schrodinger software was unable to
process a given molecule. In the end, molecular properties
were generated for 92%, 96%, and 99% of the D3 *N*-Acyl α -
Amino Acid Set, the NCI Diversity Set III, and ChemBridge
Diversity CombiSet libraries, respectively. The relevant fig-
ures and statistics were generated using these compounds.



SCHEME 1 Synthesis of D3 catalog members 1, 2, and 3

2.4 | Preparing a virtual catalog of three-dimensional D3 structures for docking

Schrodinger's LigPrep module was used to generate 3D structures of all D3 compounds for use in docking studies. Unlike the virtual library of 3D compounds generated to calculate molecular properties, the docking library should account for the multiple ionic, tautomeric, and conformational states potentially associated with each compound under physiological conditions. To this end, Schrodinger's Epik algorithm^[20,21] was used to determine all protonation states at pH values ranging from 5.0 to 9.0. All tautomers were enumerated, and one low-energy ring conformation was considered per compound. After salt molecules had been removed, the 3D geometry of each compound was optimized using the OPLS_2005 force field.^[19] The chiralities explicitly specified in the D3 database were retained. When alternate protonation and tautomeric states were considered, there were ultimately 27,220, 26,794, and 27,036 *N*-acyl α -amino acid (1), *N*-acyl α -amino acid ester (2), and *N*-acyl α -amino acid amide (3) derivatives, respectively. These structures were saved in the SDF format, which is compatible with computer docking programs like Schrodinger's Glide.^[22,23]

To facilitate docking with programs such as AUTODOCK^[24] and VINA,^[25] each of the structures was also saved in the PDBQT format. The SDF to PDB conversions were performed using OPEN BABEL.^[26] The PDB to PDBQT conversions, which involved computing Gasteiger partial charges for each atom,^[27] assigning AUTODOCK atom types, and merging non-polar hydrogen atoms with their parent heteroatoms, were performed using AUTODOCKTOOLS.^[24] Compound torsions were assigned using AutoDock's AutoTors utility to allow full molecular flexibility during docking.

2.5 | Benchmark virtual screen

AutoDock Vina was used to dock the entire 24,416-member D3 *N*-acyl α -amino acid set into the 3KAI Pin1 structure.^[28]

The ligand PDBQT files were taken from the database described in the previous paragraph. The protein was processed using Schrödinger's Protein Preparation Wizard^[29] to add hydrogen atoms, optimize hydrogen-bond networks, relax the structure, etc. The processed structure was then converted to the PDBQT format using AUTODOCKTOOLS.^[24] Compounds were docked into a cube (20 Å × 20 Å × 20 Å) centered on the Pin1 active site, using the Rocce cluster provided by the National Biomedical Computation Resource. Default Vina parameters were used.

2.6 | Combinatorial chemistry and the D3 initiative

In theory, the number of drug-like compounds exceeds 10⁶⁰.^[30] Any hope of exploring all of "drug space," whether by computational or experimental means, must be abandoned from the outset. Given that researchers are restricted to a limited subspace, it is prudent to focus on small molecules that can be readily synthesized using simple and robust synthetic procedures.^[1] These requirements can be met by designing drug discovery projects around readily accessible small molecules derived using carefully designed synthetic methodologies. Enumerated virtual combinatorial catalogs can potentially include far more compounds than even the largest preexisting physical collections used in traditional screening, and subsequent drug optimization is simpler because analogs of any initial hits can often be easily prepared using the same chemistry.

In this spirit, a subset of the chemistry employed in the D3 initiative focuses on chemical reactions that produce unnatural and natural *N*-acyl α -amino acids (1),^[2,4,31,32] *N*-acyl α -amino acid esters (2),^[3] and *N*-acyl α -amino acid amides (3),^[11] as well as their dipeptide counterparts. These compounds are derived from combinatorial scaffolds 1–3 (Scheme 1) with "sites of diversity" to which new chemical groups can be added (e.g., new α -side chains, *N*-terminal substituents, and/or various carboxylic acid derivatives: $-\text{CO}_2\text{H}$,

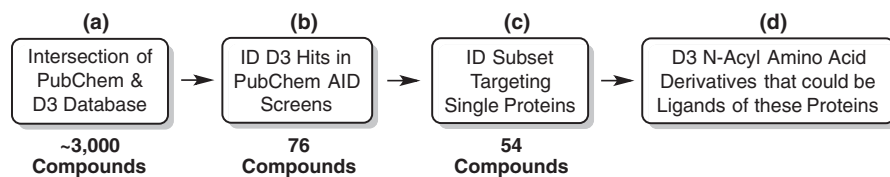


FIGURE 2 The cheminformatics analysis. Common D3 virtual and bioactive PubChem compounds

–CO₂Me, and –CONH₂). The simple solid-phase reactions employed in D3 can also be readily adapted to an educational setting with undergraduate or graduate students.^[1–4]

The synthetic chemistry for the five-step preparation of D3-catalog compounds **1**,^[2,4] **2**,^[3] and **3**^[11] is outlined below (Scheme 1). The starting benzophenone imine (Schiff base) of glycine bound to either Wang (X=O) or Rink (X=NH) resin (**4**) is deprotonated with base. The intermediate carbanion then reacts with an alkyl halide or Michael acceptor (both represented as the generic red electrophile **R¹X**) to form racemic **5**, which contains the first variable input, an **R¹** group. Hydrolysis to salt **6** is followed by neutralization to the free amine **7**. The second variable input, **R²** in blue, is then added with an *N*-acylation by coupling **7** with a carboxylic acid **R²CO₂H** to give **8** (when an *N*-acyl or other protected α -amino acid is used as the acylating agent **R²CO₂H**, and the D3 products are protected dipeptide derivatives). Finally, the resin-bound product is cleaved from the resin to yield an unnatural *N*-acyl α -amino acid **1** (using trifluoroacetic acid, TFA), an unnatural *N*-acyl methyl ester **2** (MeOH/base) from Wang resin, or an unnatural *N*-acyl primary amide **3** from Rink amide resin (TFA).

This protocol produces a racemic mixture of both (*S*)- and (*R*)-*N*-acyl α -amino acid derivatives, allowing both epimers to be simultaneously tested in any biological assay. When activity is detected, an optically enriched sample of the mixture can be produced by enantioselective solution- or solid-phase synthesis,^[31,32] or chiral chromatography.^[33,34]

A virtual catalog of 73,024 synthesizable compounds accessible by D3 solid-phase reaction sequences was enumerated from a diversity set of 100 and 100 commercially available electrophiles and carboxylic acids, respectively.^[2] This catalog contains 12,208 each of the (*S*)- and (*R*)-*N*-acyl α -amino acids **1**; 12,096 each of the (*S*)- and (*R*)-*N*-acyl α -amino acid methyl esters **2**; and 12,208 each of the (*S*)- and (*R*)-*N*-acyl α -amino acid primary amides **3**. See the Supporting Information (pp. SI-4 & SI-5) for a more complete explanation of the chemical enumeration and library sizes.

3 | RESULTS AND DISCUSSION

3.1 | α -Amino acid derivatives and biological activity

Many endogenous α -amino acids have known biological effects beyond the fundamental role they play as

protein-building blocks. As D3 compounds are based on the α -amino acid scaffold, it is reasonable to suppose that they may be enriched with bioactive compounds. A collection of 40 approved drugs containing the α -amino acid structural framework (nitrogen to alpha-carbon to carbonyl unit, highlighted in red), is shown in Table SI-1 (see Supporting Information for full structures, references, and other data). Of these compounds, 35% (14/40) are *N*-acyl α -amino acids, *N*-acyl dipeptides, or their ester or amide derivatives: alvimopan, bortezomib, ceftaroline, doripenem, folic acid, lacosamide, lenalidomide, methotrexate, pemetrexed, penicillin V, pralatrexate, raltitrexed, tacrolimus, and valsartan.

3.2 | Identifying known drug targets of D3 compounds

To verify the pharmacological utility of the D3 libraries, we explored whether or not any virtual D3 molecules had previously reported biological activity. Among the >51 million molecules with unique chemical structures (CIDs) deposited in the PubChem database (Fig. SI-2),^[35] approximately 3,000 compounds were also serendipitously present in D3 catalogs (Figure 2a). These molecules included 76 compounds (Figure 2b) that were active in 95 different PubChem bioassays (AIDs).

Many PubChem assays detect phenotypic responses or changes to entire biochemical pathways. These types of screens are useful for drug discovery, but structure-based drug design (“SBDD,” e.g., computer docking) requires a specific macromolecular target. By manually examining the PubChem assay descriptions of the 95 assays noted above, we identified 45 that are likely to involve binding to the orthosteric pockets^[36] of 32 distinct single-protein drug targets (Table SI-2). Given that D3 compounds are α -amino acid derivatives, the majority of these protein targets (25 of 32) had endogenous or natural substrates that were free α -amino acids, short peptides, specific protein residues, or α -amino acid metabolites. However, targets with nucleic acid and steroidal endogenous ligands are also amenable to D3 binding, suggesting even broader utility.

Given that so many of these drug targets bind to natural α -amino acids, it is not surprising that there are many *N*-acyl α -amino acids with natural side chains among the D3/PubChem actives. The strength of the D3 protocol lies

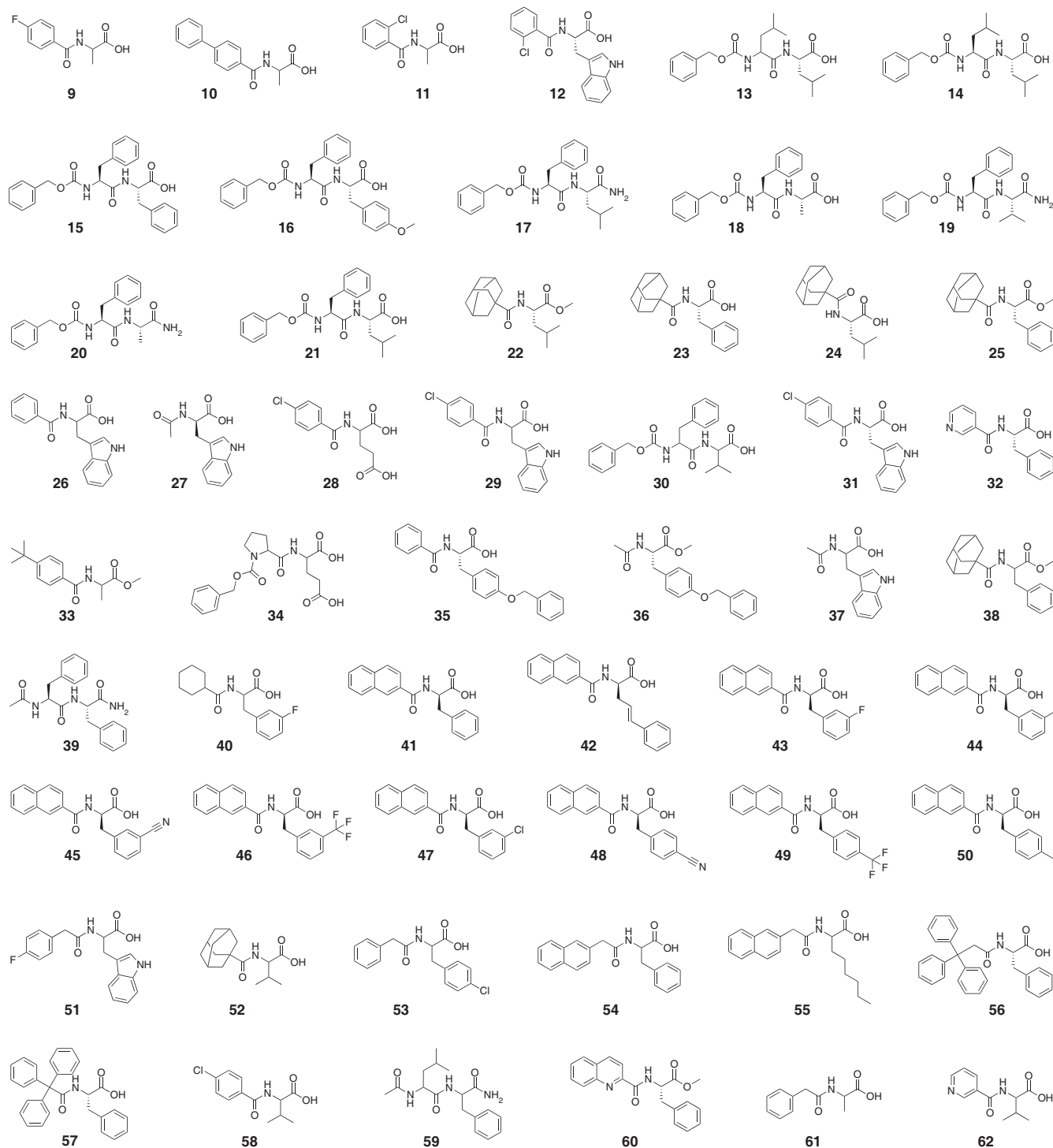


FIGURE 3 Fifty-four D3/PubChem hits in SBDD-amenable assays (i.e., assays judged likely to involve binding to orthosteric pockets of single-protein targets). See Table SI-3 for further details

in its ability to produce combinatorially diverse sets of *both* unnatural and natural α -amino acid derivatives. The 45 SBDD-amenable PubChem assays contained 54 unique D3 active compounds (Figures 2c, 3; Table SI-3). Insight into the structural features governing the activity of these 54 compounds is provided in the section entitled “Comparing the D3 virtual catalog members with D3/PubChem actives,” below.

3.3 | A computer-aided drug discovery example: Pin1 inhibitors

To further demonstrate the utility of the D3 virtual database, attention was next focused on PubChem SBDD-amenable assays that include active compounds also present in D3 catalogs. One of these was a PubChem assay against Pin1

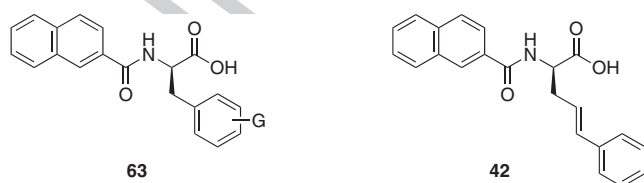
1 reported in 2010 (AID: 474989, Table SI-2, entry 14). Also
 2 known as “peptidyl-prolyl cis-trans isomerase (PPIase)
 3 NIMA-interacting 1,” Pin1 catalyzes the cis-trans isomeri-
 4 zation of prolyl amide bonds in select substrate proteins.
 5 Because it is thought to play a role in cancer pathogenesis,
 6 Pin1 has been studied as a potential drug target for novel can-
 7 cer chemotherapies. The PubChem assay AID 474989 de-
 8 tected specific binding (leading to inhibition) to the primary
 9 Pin1 active site^[15] and reported compounds that have a K_i
 10 less than 50 μM .

11 Thirteen of the 32 molecules tested for Pin1 activity in
 12 2010^[15] had been previously disclosed in the 2008 virtual D3
 13 *N*-acyl α -amino acid set.^[2] Of these thirteen compounds, ten
 14 were reported as active in PubChem ($K_i < 50 \mu\text{M}$), and three
 15 with lower affinities were reported as inactive. Twelve of the
 16 13 D3 compounds were monosubstituted *N*-acyl phenylalan-
 17 nines (**63**, G = H, 2- or 3-F, 2- or 3- or 4-CF₃, 2- or 3- or 4-
 18 CN, 3- or 4-Me, 3-Cl), and one was a phenylalanine vinyllog
 19 **42** (Figure 4).

20 It is notable that these Pin1 inhibitors have the (R)- or
 21 D-stereochemistry not normally present in proteinogenic α -
 22 amino acids and their derivatives. This highlights a strength
 23 of the D3 approach. D3-based chemistry enables the synthe-
 24 sis and testing of both α -amino acid stereochemistries (usu-
 25 ally in the racemic form, and then separately after racemic
 26 activity is observed). In contrast, medicinal chemistry is often
 27 biased in favor of α -amino acids (and their derivatives) with
 28 the naturally occurring stereochemistry, perhaps due to avail-
 29 ability of starting materials. Without access to the unnatural
 30 stereoisomer, important drugs or drug leads could be missed.
 31 Examples of drugs that are unnatural α -amino acids or their
 32 derivatives include nateglinide,^[13] lacosamide,^[37] penicilla-
 33 mine,^[38] penicillin V,^[39] and ximelagatran^[40] (see Table SI-1
 34 and the Supporting Information text for more examples).

35 To further demonstrate the utility of the D3 catalogs, D3-
 36 compound binding to Pin1 was evaluated in an in silico bench-
 37 mark screen. A crystal structure of Pin1 (PDB ID: 3KAI) was
 38 used for docking.^[28] To validate this target, the co-crystallized
 39 compound **64** (Figure 5) was removed from the 3KAI pocket,
 40 and a known inhibitor and member of the D3 virtual catalog
 41 (**42**, Figures 3 and 4, Table SI-3) was docked using AutoDock
 42 Vina.^[25]

43 The resulting docked pose of bound **42** (Figure 6) closely
 44 approximated the crystallographic pose captured in the 3JYJ
 45



52 **FIGURE 4** Known Pin1 inhibitors^[15] present in the D3 *N*-acyl
 53 α -amino acid set^[2]

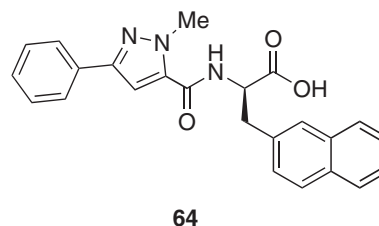


FIGURE 5 The co-crystallized compound removed from the 3KAI Pin1 structure prior to docking

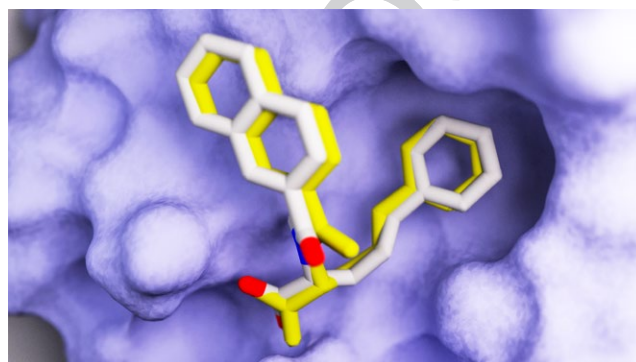


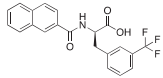
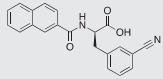
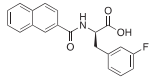
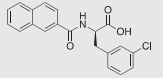
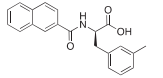
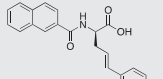
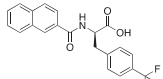
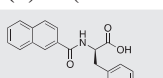
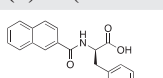
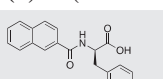
FIGURE 6 A comparison of the docked and crystallographic poses of **42** (CID 45100499, a known Pin1 inhibitor and also a member of the D3 virtual database). The docked and crystallographic poses are shown in white and yellow, respectively. The compound was docked into the 3KAI structure.^[28] The actual crystallographic pose was taken from the 3JYJ structure^[15]

Pin1 structure.^[15] Given that AutoDock Vina was able to reproduce the experimental pose, the entire 24,416-member D3 *N*-acyl α -amino acid set (including the identified Pin1 inhibitors) was next docked into 3KAI. As the co-crystallized compound **64** was not present in the D3 database, the results of the virtual screen were not biased in favor of any D3 compound.

This benchmark virtual screen was particularly adept at predicting the activity of the ten known Pin1 inhibitors present in the 24,000-member D3 virtual catalog of *N*-acyl α -amino acids (**1**, Table 1). Six of the ten compounds listed as active in PubChem were contained in the top 5% (roughly 1,200) of Vina-ranked D3 compounds. A receiver-operating characteristic (ROC) curve was next generated, with the known inhibitors and remaining compounds serving as the true positives and negatives, respectively. The area under this curve was 0.93, suggesting that, given a randomly selected pair of compounds comprised of one known inhibitor and one uncharacterized compound, there is a 93% chance that the known inhibitor has the better Vina score.^[25] Random ranking would put this probability at only 50%.

Given the performance of this virtual screen, some top-ranking but untested D3 compounds are also likely Pin1

TABLE 1 The Vina scores and percentile rankings of the ten positive controls (i.e., known inhibitors) included in the benchmark Pin1 screen

ID	Vina score	Percentile rank
 (R)-52 (DDD-000012382)	-7.3	1.5
 (R)-51 (DDD-000012386)	-7.3	1.5
 (R)-49 (DDD-000012361)	-7.2	2.4
 (R)-53 (DDD-000012365)	-7.2	2.4
 (R)-50 (DDD-000012379)	-7.2	2.4
 (R)-48 (DDD-000012394)	-7.2	2.4
 (R)-55 (DDD-000012383)	-6.9	7.5
 (R)-47 (DDD-000012350)	-6.8	10.5
 (R)-56 (DDD-000012380)	-6.8	10.5
 (R)-54 (DDD-000012387)	-6.6	18.8

inhibitors. Although it is beyond the scope of this paper, future directions include experimentally synthesizing and testing a number of these top-ranked compounds to identify small molecules with improved activity.

3.4 | Comparing the D3 virtual catalog members with D3/PubChem actives

We next compared the structural characteristics of the 73,024-member D3 virtual catalog with the 54-member D3/

PubChem active set (Figure 7). These two sets both contain all three D3 structural classes [*N*-acyl α -amino acids (**1**) and their methyl ester (**2**) and amide (**3**) derivatives], but in differing proportions. In the larger D3 set (Figure 7, left column), the numbers of compounds in each class are nearly equal because the D3 molecules were derived combinatorially from the same set of 100 electrophile (R^1X) and 100 carboxylic acid (R^2CO_2H) inputs, differing only in the method of resin cleavage (Scheme 1). The varying numbers of **2** are a result of the different cleavage condition used, which caused different side chain side reactions. For example, if R^1 in compounds **1–3** is CH_2CO_2tBu , cleavage of **1** or **3** with TFA converts the protected *t*-butyl ester to the carboxylic acid ($CH_2CO_2tBu \rightarrow CH_2CO_2H$). In contrast, cleavage of **2** with Et_3N leaves the side chain unchanged ($CH_2CO_2tBu \rightarrow CH_2CO_2tBu$).

As might be expected, the proportions of structural classes in the D3 virtual catalog set are quite different from the D3/PubChem active set (Figure 7, right column). A lower percentage of a particular structural class in the active set could imply that it is not fertile territory for drug discovery and is overrepresented in the D3 virtual catalog. On the other hand, perhaps that structural class is simply underrepresented among the PubChem molecules made and tested. Without going through a complete analysis of the structural classes present in PubChem, both tested and untested, we are confident that this latter hypothesis is the case, especially when comparing the structural categories of “proteinogenic/non-proteinogenic side chains” and “natural/unnatural side chain stereochemistry.” The introduction, via a racemic alkylation reaction, of side chains usually not found in proteinogenic amino acids is at the core of D3 chemistry. This process gives rise to many unnatural side chain residues with equal proportions of the natural (e.g. L- or S-) and unnatural D- or R- stereoisomers. To account for the presence, potential activity, and computational accessibility of all enantiomers, each is explicitly enumerated in the D3 virtual catalog. In cases where D3 compounds contain two or three stereogenic centers, all four or eight stereoisomers, respectively, are included in the catalog.

The preponderance of the natural stereochemistry and side chains in the PubChem active set is not surprising given that amino acid-based compounds submitted to PubChem for testing are likely biased in favor of derivatives of the more readily available natural (L)- α -amino acid stereoisomer. What is noteworthy is the significant number of D3/PubChem active compounds (20%) with the unnatural (D)-stereochemistry. These would not have been present in the D3/PubChem intersection if the D3 project had considered only enantiomerically pure molecules with the (L)-stereochemistry. The observed activity of some racemic PubChem actives (39%) may also be explained in part by the D3 methodology, which generates the unnatural and natural stereoisomers in equal

73K-Member D3 Virtual Catalog of N-Acyl Amino Acids and Derivatives	54 D3/PubChem Active N-Acyl Amino Acids and Derivatives
N-Acyl Amino Acids & Derivatives (73,024 Total)	N-Acyl Amino Acids & Derivatives (54 Total)
24,416 N-Acyl Amino Acids (1) (33%)	43 N-Acyl Amino Acids (1) (80%)
24,192 N-Acyl Amino Acid Me Esters (2) (33%)	6 N-Acyl Amino Acid Me Esters (2) (11%)
24,416 N-Acyl Amino Acid Amides (3) (33%)	5 N-Acyl Amino Acid Amides (3) (9%)
R¹Xs Yielding Side Chain Types (100 Total)	Side Chain Types (54 Total)
13 Proteinogenic Side Chains (13%)	42 Proteinogenic (78%)
Ala (1), Asp (3), Glu (2), Ile (1), Leu (1), Phe (1), Trp (1), Tyr (1), Val (1)	Ala (7), Asp (0), Glu (2), Ile (0), Leu (6), Phe (12), Trp (7), Tyr (3), Val (5)
87 Non-proteinogenic Side Chains (87%)	12 Non-proteinogenic Side Chains (22%)
40 Substituted PhCH ₂ X (46%)	10 Unnatural Ring-Substituted Phe (83%)
5 HetCH ₂ X (6%)	2 Unnatural Alkyls (17%)
5 Polycyclic ArCH ₂ X (6%)	Stereochemistry (54 Total)
8 Allylic or Propargylic Halides (9%)	22 Natural (L) Side Chains (41%)
15 Alkyl Halides (17%)	11 Unnatural (D) Side Chains (20%)
14 Michael Acceptors (16%)	21 Racemic (D,L) Side Chains (39%)
N-Acyl Amino Acids vs. Dipeptides (100 Total)	N-Acyl Amino Acids vs. Dipeptides (54 Total)
83 N-Acyl Amino Acids (83%)	41 N-Acyl Amino Acids (76%)
17 N-Protected Dipeptides (17%)	13 N-Protected Dipeptides (24%)

FIGURE 7 Comparing the 73,024 D3 compounds and the 54 D3/PubChem actives

proportions. In any case, the presence of these “unnatural” amino acid derivatives in both the D3 virtual catalog and PubChem active set demonstrates the value of the D3 virtual catalog as a resource for identifying biologically active compounds from an underrepresented structural class.

Finally, when an *N*-acyl or other protected α -amino acid is used as the acylating agent R^2CO_2H , the D3 products are protected dipeptide derivatives (see Fig. SI-5 and associated text). In total, such dipeptide derivatives comprise 17% of the 73K D3 virtual catalog and 24% of the 54-member D3/PubChem set (Figures 3 and 7), pointing to the ready availability of such derivatives via the D3 methodology.

3.5 | The D3 *N*-acyl α -amino acid virtual catalog

The Pin1 virtual screen of the above 24,416-member “D3 *N*-Acyl α -Amino Acid set” illustrates how the D3 chemical library can be particularly useful for drug discovery projects. Computed chemical descriptors suggest that the compounds in this catalog are chemically similar to those of other libraries commonly used in virtual and experimental screens (Figure 8). For comparison, consider the NCI Diversity Set III (NCI III) and the ChemBridge Diversity CombiSet (ChemBridge). The average calculated molecular properties of the D3 *N*-acyl α -amino acid set are intermediate between those of these two popular libraries, with the exception of the hydrogen-bond-donor count, where the ChemBridge set is anomalous (Table 2, Figure 7). The overwhelming majority of the catalog compounds (97.2%, 98.6%, and 99.8% of the D3 *N*-Acyl α -Amino Acid, NCI III, and ChemBridge compounds, respectively) satisfy Lipinski’s “Rule of Five” for druglikeness.^[41]

To encourage broad use of the D3 virtual catalog, structures of the D3 *N*-acyl α -amino acids, *N*-acyl α -amino acid esters, and *N*-acyl α -amino acid amides have been prepared in formats that are compatible with several computer

docking programs, including AutoDock,^[24] Vina,^[25] and Schrodinger’s Glide.^[22,23] Various possible tautomeric, protonation, and ring conformational states of each compound were carefully considered. These virtual catalogs may be downloaded free of charge from http://durrantlab.com/liglib/iupui/d3_docking/ (accessed February 27, 2017).

Many researchers may wish to prepare D3 structures according to their own protocols. The SMILES string of each D3 compound has also been published on both the D3 (http://durrantlab.com/liglib/iupui/d3_docking/, accessed February 27, 2017) and Collaborative Drug Discovery (CDD) (<https://www.collaborativedrug.com>, accessed February 27, 2017) websites.^[42,43] CDD also allows researchers to search and download subsets of the entire D3 virtual catalog.

4 | CONCLUSIONS

Two complementary strengths of D3 virtual catalogs are their bioactivity potential and synthetic accessibility. By demonstrating the retrospective and prospective biological utility of these catalogs, this manuscript encourages computational chemists to probe D3 molecules with their own tools and hypotheses. Compounds enumerated in the virtual catalog are based on well-documented D3 procedures, so synthesizing and testing computationally selected molecules are entirely feasible. The D3 catalogs thus enable collaborations between computationalists, synthetic chemists, and biologists.

In summary, to demonstrate the utility of the D3 virtual catalog, we identified a number of D3 molecules with known biological activity that were present in the PubChem database. These results will help guide future drug discovery efforts. Using computational techniques such as docking, students, and researchers can identify analogs of known modulators that are likely to bind to the same target. These compounds can then be synthesized and distributed to collaborators,

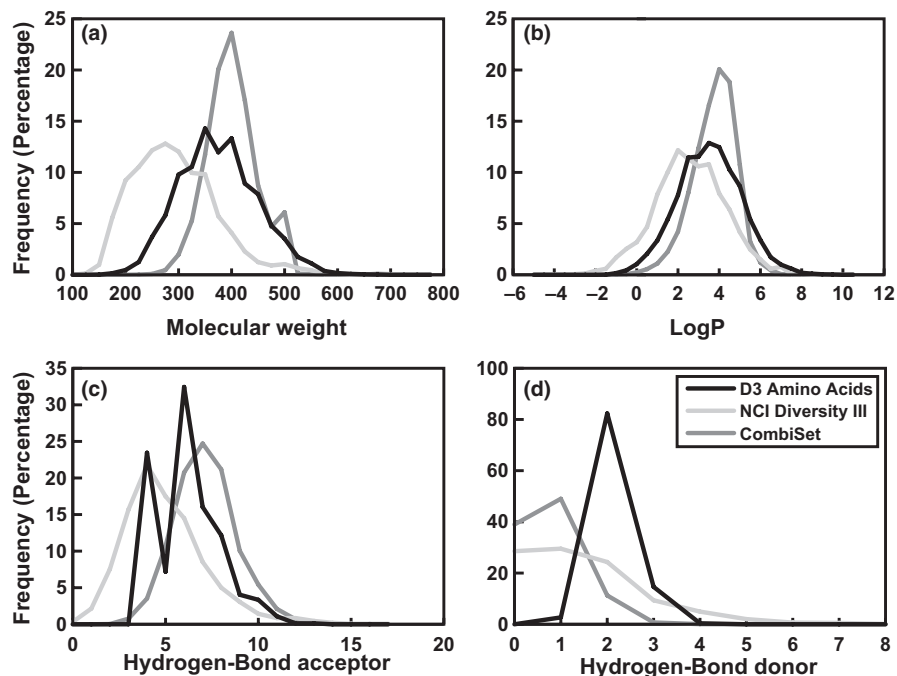


FIGURE 8 Histograms showing the molecular weights, calculated LogP values, and hydrogen-bond acceptor/donor counts for the D3 *N*-Acyl α -Amino Acid set, the NCI Diversity set III, and the ChemBridge Diversity CombiSet catalogs

TABLE 2 The average molecular properties of compounds in the D3 *N*-Acyl α -Amino Acid Set, the NCI Diversity Set III, and the ChemBridge Diversity CombiSet

	D3	NCI III	ChemBridge
Molecular weight (Daltons)	361.8 (73.4)	280.5 (80.6)	388.6 (47.3)
logP	3.3 (1.5)	2.3 (1.7)	3.5 (1.1)
Hydrogen-bond acceptors	5.8 (1.6)	4.7 (2.2)	6.8 (1.6)
Hydrogen-bond donors	1.5 (0.4)	1.4 (1.3)	0.7 (0.7)

Standard deviations are shown between parentheses.

who evaluate the pharmacological, enzymatic, or signaling effect to confirm or refute the computational predictions. Others who wish to identify additional D3 drug targets via virtual screening can download a carefully prepared catalog of D3 compounds from http://durrantlab.com/liglib/iupui/d3_docking/.

ACKNOWLEDGEMENTS

The National Science Foundation (NSF/DUE-1140602, NSF/MRI-CHE-0619254, and NSF/MRI-DBI-0821661), National Institutes of Health (RO1-GM28193), IUPUI International Development Fund, Indiana University Purdue University Indianapolis (IUPUI) STEM Summer Scholars Institute, and Analytical Technologies at Eli Lilly and Company are acknowledged for their generous support. We thank Collaborative Drug Discovery Inc. for generously hosting D3

molecular data on their servers, ChemAxon for granting access to their chemical-enumeration software, Guillermo Morales for helpful discussions, and the NCI/DTP Open Chemical Repository (<http://dtp.cancer.gov>, accessed February 27, 2017) for providing compound structures for download.

CONFLICT OF INTEREST

The authors declare that they have no financial or commercial conflicts of interest.

ENDNOTES

- ^a *CombiChem for Excel* from CambridgeSoft, PirkinElmer, 2015.
- ^b *ChemAxon*, ChemAxon Kft., 2015.
- ^c *Maestro*, Schrödinger, LLC, 2015.
- ^d *LigPrep*, Schrödinger, LLC, 2015.
- ^e *QikProp*, Schrödinger, LLC, 2015.

REFERENCES

- [1] W. L. Scott, M. J. O'Donnell, *J. Comb. Chem.* **2009**, *11*, 3.
- [2] W. L. Scott, J. Alsina, C. O. Audu, E. Babaev, L. Cook, J. L. Dage, *et al.*, *J. Comb. Chem.* **2009**, *11*, 14.
- [3] W. L. Scott, C. O. Audu, J. L. Dage, L. A. Goodwin, J. G. Martynow, L. K. Platt, *et al.*, *J. Comb. Chem.* **2009**, *11*, 34.
- [4] W. L. Scott, R. E. Denton, K. A. Marrs, J. D. Durrant, J. G. Samaritoni, M. M. Abraham, *et al.*, *J. Chem. Educ.* **2015**, *92*, 819.
- [5] A. R. Leach, M. M. Hann, *Drug Discov. Today*. **2000**, *5*, 326.
- [6] J. M. Barnard, G. M. Downs, A. von Scholley-Pfab, R. D. Brown, *J. Mol. Graph. Model.* **2000**, *18*, 452.
- [7] V. S. Lobanov, D. K. Agrafiotis, *J. Mol. Graph. Model.* **2001**, *19*, 571.

- [8] V. S. Lobanov, D. K. Agrafiotis, *Comb. Chem. High Throughput Screen.* **2002**, *5*, 167.
- [9] D. V. S. Green, S. D. Pickett, *Mini Rev. Med. Chem.* **1067**, 2004, 4.
- [10] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, *et al.*, *J. Med. Chem.* **2014**, *57*, 4977.
- [11] J. G. Samaritoni, A. T. Copes, D. K. Crews, C. Glos, A. L. Thompson, C. Wilson, *et al.*, *J. Org. Chem.* **2014**, *79*, 3140.
- [12] C. P. Austin, L. S. Brady, T. R. Insel, F. S. Collins, *Science* **2004**, *306*, 1138.
- [13] H. Shinkai, M. Nishikawa, Y. Sato, K. Toi, I. Kumashiro, Y. Seto, *et al.*, *J. Med. Chem.* **1989**, *32*, 1436.
- [14] D. C. Swinney, J. Anthony, *Nat. Rev. Drug Discov.* **2011**, *10*, 507.
- [15] L. M. Dong, J. Marakovits, X. Hou, C. Guo, S. Greasley, E. Dagostino, *et al.*, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2210.
- [16] S. L. Holbeck, *Eur. J. Cancer* **2004**, *40*, 785.
- [17] P. V. Desai, A. Patny, Y. Sabnis, B. Tekwani, J. Gut, P. Rosenthal, *et al.*, *J. Med. Chem.* **2004**, *47*, 6609.
- [18] L. Zhang, D. Fourches, A. Sedykh, H. Zhu, A. Golbraikh, S. Ekins, *et al.*, *J. Chem. Inf. Model.* **2013**, *53*, 475.
- [19] J. L. Banks, H. S. Beard, Y. Cao, A. E. Cho, W. Damm, R. Farid, *et al.*, *J. Comput. Chem.* **2005**, *26*, 1752.
- [20] J. R. Greenwood, D. Calkins, A. P. Sullivan, J. C. Shelley, *J. Comput.-Aided Mol. Des.* **2010**, *24*, 591.
- [21] J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin, M. Uchimaya, *J. Comput. Aid. Mol. Des.* **2007**, *21*, 681.
- [22] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, *et al.*, *J. Med. Chem.* **2004**, *47*, 1739.
- [23] T. A. Halgren, R. B. Murphy, R. A. Friesner, H. S. Beard, L. L. Frye, W. T. Pollard, *et al.*, *J. Med. Chem.* **2004**, *47*, 1750.
- [24] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, *et al.*, *J. Comput. Chem.* **2009**, *30*, 2785.
- [25] O. Trott, A. J. Olson, *J. Comput. Chem.* **2009**, *31*, 455.
- [26] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, *J. Cheminf.* **2011**, *3*, 33.
- [27] J. Gasteiger, M. Marsili, *Tetrahedron* **1980**, *36*, 3219.
- [28] A. J. Potter, S. Ray, L. Gueritz, C. L. Nunns, C. J. Bryant, S. F. Scrace, *et al.*, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 586.
- [29] G. M. Sastry, M. Adzhigirey, T. Day, R. Annabhimoju, W. Sherman, *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221.
- [30] M. Ohlmeyer, M. M. Zhou, *Mt Sinai J. Med.* **2010**, *77*, 350.
- [31] M. J. O'Donnell, *Aldrichim. Acta.* **2001**, *34*, 3.
- [32] M. J. O'Donnell, *Acc. Chem. Res.* **2004**, *37*, 506.
- [33] N. M. Maier, P. Franco, W. Lindner, *J. Chromatogr. A* **2001**, *906*, 3.
- [34] J. H. Kennedy, M. D. Belvo, V. S. Sharp, J. D. Williams, *J. Chromatogr. A* **2004**, *1046*, 55.
- [35] T. J. Cheng, Y. M. Pan, M. Hao, Y. Wang, S. H. Bryant, *Drug Discov. Today* **2014**, *19*, 1751.
- [36] F. De Smet, A. Christopoulos, P. Carmeliet, *Nat. Biotechnol.* **2014**, *32*, 1113.
- [37] S. de Biase, G. L. Gigli, M. Valente, G. Merlino, *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 459.
- [38] S. Wadhwa, R. J. Mumper, *Cancer Lett.* **2013**, *337*, 8.
- [39] K. Brown, *Penicillin Man: Alexander Fleming and the Antibiotic Revolution*, Sutton Publishing Limited, London **2005**.
- [40] L. Testa, R. Bhindi, P. Agostoni, A. Abbate, G. G. L. B. Zoccai, W. J. van Gaal, *Expert Opin. Drug Saf.* **2007**, *6*, 397.
- [41] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.
- [42] R. R. Gupta, E. M. Gifford, T. Liston, C. L. Waller, M. Hohman, B. A. Bunin, *et al.*, *Drug Metab. Dispos.* **2010**, *38*, 2083.
- [43] S. Ekins, B. A. Bunin, *Computational Approaches and Collaborative Drug Discovery for Trypanosomal Diseases*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany **2013**.

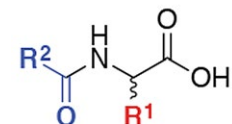
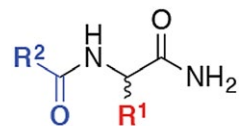
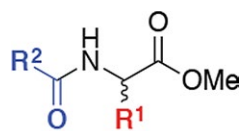
SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Abraham MM, Denton RE, Harper RW, Scott WL, O'Donnell MJ, Durrant JD. Documenting and harnessing the biological potential of molecules in distributed drug discovery (D3) virtual catalogs. *Chem Biol Drug Des.* 2017;00:1–10. <https://doi.org/10.1111/cbdd.13012>

Graphical Abstract

The contents of this page will be used as part of the graphical abstract of html only. It will not be published as part of main.



The Distributed Drug Discovery (D3) program provides scientists with a free virtual catalog of 73,024 easy-to-synthesize *N*-acyl unnatural α -amino acids, their methyl esters, and primary amides. In the current work, we identify all virtual D3 compounds classified as bioactive hits in PubChem-cataloged experimental assays. The results (i) provide insight into the broad range of drug-target classes amenable to modulation by D3-accessible molecules and (ii) suggest future avenues for virtual screening/drug discovery projects.