ORIGINAL ARTICLE

Curation and cheminformatics analysis of a Ugi-reaction derived library (URDL) of synthetically tractable small molecules for virtual screening application

Mukesh Tandi[1] · Nancy Tripathi[1] · Animesh Gaur[1] · Balasubramanian Gopal[2] · Sandeep Sundriyal[1]

Abstract

Virtual screening (VS) is an important approach in drug discovery and relies on the availability of a virtual library of synthetically tractable molecules. Ugi reaction (UR) represents an important multi-component reaction (MCR) that reliably produces a peptidomimetic scaffold. Recent literature shows that a tactically assembled Ugi adduct can be subjected to further chemical modifications to yield a variety of rings and scaffolds, thus, renewing the interest in this old reaction. Given the reliability and efficiency of UR, we collated an UR derived library (URDL) of small molecules (total = 5773) for VS. The synthesis of the majority of URDL molecules may be carried out in 1–2 pots in a time and cost-effective manner. The detailed analysis of the average property and chemical space of URDL was also carried out using the open-source Datawarrior program. The comparison with FDA-approved oral drugs and inhibitors of protein–protein interactions (iPPIs) suggests URDL molecules are 'clean', drug-like, and conform to a structurally distinct space from the other two categories. The average physicochemical properties of compounds in the URDL library lie closer to iPPI molecules than oral drugs thus suggesting that the URDL resource can be applied to discover novel iPPI molecules. The URDL molecules consist of diverse ring systems, many of which have not been exploited yet for drug design. Thus, URDL represents a small virtual library of drug-like molecules with unexplored chemical space designed for VS. The structures of all molecules of URDL, oral drugs, and iPPI compounds are being made freely accessible as supplementary information for broader application.

Keywords Ugi reaction · Post-Ugi modification · Virtual screening · Similarity · Druglikeness · Datawarrior

Abbreviations

ArC	Carboaromatic rings
BBB	Blood–brain barrier
clogP	Calculated Partition coefficient
Da	Dalton
FDA	Food and drug administration
Fsp3	Fraction of sp3 carbons
HTS	High throughput screening
HBD	Hydrogen bond donor
HBA	Hydrogen bond acceptor
iPPI	Inhibitors of protein–protein interaction
IUPAC	International Union of Pure and Applied Chemistry
KNIME	Konstanz Information Miner
MCR	Multi-component reaction
MW	Molecular weight
PSA	Polar surface area
REAL	Readily accessible
RB	Rotatable bonds
SAR	Structure–activity relationship
SBVS	Structure-based virtual screening
SBAN	Sum of basic and aromatic nitrogen
TPSA	Topological surface area
t-SNE	T-distributed stochastic neighbour embedding
2D	Two dimension
3D	Three dimension
URDL	Ugi-reaction derived library
UR	Ugi reaction
Ugi3CR	Ugi 3-component reaction
USD	United States dollar
VS	Virtual screening

✉ Sandeep Sundriyal
sandeep.sundriyal@pilani.bits-pilani.ac.in

1 Department of Pharmacy, Birla Institute of Technology and Science Pilani, Pilani Campus, Rajasthan 333031, India

2 Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

Introduction

Modern drug discovery relies on the identification of small molecules capable of interacting with a biological target of interest (receptor or enzyme) to achieve a therapeutic action. In a typical 'target-based' approach, finding a 'hit' molecule for the desired target is one of the major bottlenecks in drug discovery. One of the widely used strategies is to conduct high throughput screening (HTS) [1] of a large or ultra-large molecule collection at a miniature scale using robotic tools. The obtained hits are then further validated and structurally optimized to obtain lead molecules with potent biological activity, optimum pharmacokinetics, and low toxicity potential. However, HTS requires enormous resources in terms of sophisticated equipment, time, and skilled manpower, adding to the overall cost of drug development [2]. One of the alternatives is to screen large libraries of molecules using in silico or molecular modelling techniques rather than in wet labs. The *in-silico* or virtual screening (VS) relies on computational models to represent biomolecular targets and small molecules [3, 4]. In the widely employed structure-based VS (SBVS) approach, interactions between the protein and ligands are modelled to predict the binding affinity and pose [5–9]. The inherent advantage of VS is that only the molecules that appear promising in these models need to be experimentally validated. Thus, several small molecule libraries have emerged over the years for their application in VS and drug discovery [10, 11]. The most notable among these are ZINC [12], ChEMBL [13, 14], DrugBank [15], and PubChem [16], which are publicly accessible to download and use. However, these databases have some common set of molecules and hence overlapping chemical space [17]. With the availability of various open-source computational tools, enumerating ultra-large virtual libraries has increasingly become effortless, and so is their application in drug design [7, 9, 18]. However, one of the biggest challenges after a hit is obtained through VS is the molecules' availability and synthetic tractability for experimental validation. Most commonly, the predicted hit compounds are purchased from commercial vendors who provide a variety of small molecules including target-focused and on-demand compounds. However, the price is often exorbitantly high and often in the range of 20–200 USD per milligram for in-stock compounds. The cost is even higher for tailor-made molecules synthesized on demand. Thus, dependence on the commercial supply of molecules is not always economically viable, especially in small academic labs. Additionally, even a small amount of impurity in the sourced samples due to inadequate quality control or generated during long shelf storage might result in false positives. Therefore, resynthesis of the hits to obtain high purity samples is important for any VS approach. However, the chemical synthesis of such commercial compounds may not be reported and might involve several steps consisting of challenging chemistry.

One solution is to curate virtual libraries of molecules obtainable from highly reliable reactions. One notable example is the commercially available REadily AccessibLe (REAL) database provided by Enamine company, reported to have ~ 80% synthesis success rate [19]. Recently, the LibINVENT tool reported by Patronov and coworkers [20] has also been made available to generate in silico libraries based on different reactions. However, the synthesis of molecules from these libraries may still involve multiple steps and unforeseen practical problems such as the unavailability of the starting materials and lower yields.

Multi-component reactions (MCRs) produce complex molecules with high pot economy [21] and hence, are potential candidates to enumerate virtual libraries [22–24]. The Ugi-reaction (UR) is one of the oldest and most widely studied MCRs that typically yields a peptidomimetic scaffold in a single pot reaction between an aldehyde, carboxylic acid, amine, and isonitrile components [25, 26]. UR also possess high atom economy since only one water molecule is produced as a by-product. In the last 2 decades, application of UR in the synthesis of novel scaffolds has increased, as indicated by the number of papers appearing in PubMed (Fig. 1). A typical strategy involves the formation of UR adducts from commercially available building blocks followed by a series post-Ugi modifications that may include intramolecular heteroatom alkylation/acylation, condensation, and rearrangement reactions [27–32]. In several cases UR and post-Ugi modifications can be done in one pot without the need of isolating the UR product, thus leading to facile access to diverse chemotypes for wider application [27, 33–38].

Additionally, with the commercial availability of a variety of starting materials for UR, it is convenient to generate analogues for structure–activity relationship (SAR) studies in a parallel fashion [39]. Despite these advantages recently reported UR-derived and post-Ugi-derived interesting chemotypes remain unexplored in medicinal chemistry (vide infra). The plausible reason might be that such reports mostly appear in organic chemistry literature, primarily focusing on optimising synthetic methodology and characterization. To tap this unexplored chemical space we planned to curate a UR derived library (URDL) of small molecules reported in the literature, intending to evaluate its chemical space [40] and make it accessible for VS. The property and chemical space of URDL is also compared with the FDA-approved oral drugs as a standard.

Fig. 1 PubMed search result using the term 'Ugi reaction' in title/abstract. The result shows the increase in UR literature in the recent years
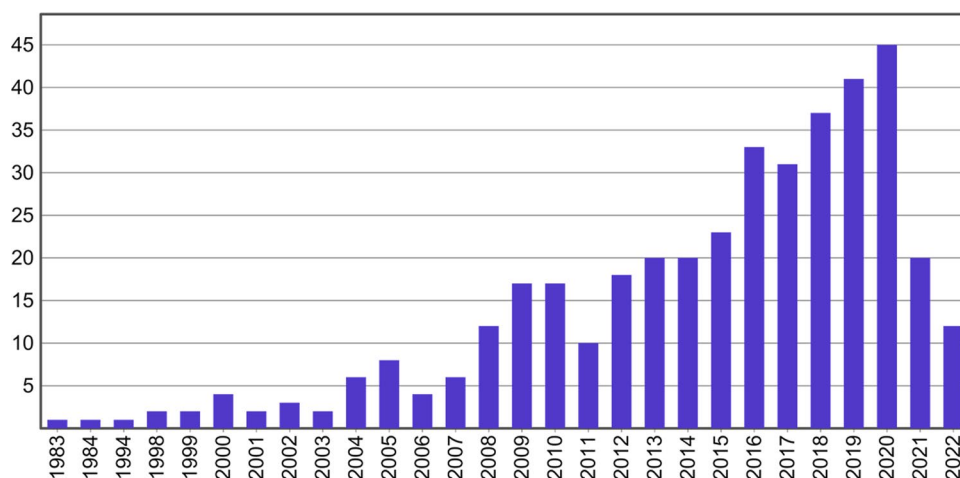
Results and discussion

Library curation and description

The PubMed search engine was used to collate literature discussing the synthesis of novel scaffolds/rings using either UR or post-Ugi modifications. Many of these molecules are reported to be synthesized conveniently in 1–4 synthetic steps which in several cases may be performed in 1–2 pots without the need of isolating/purifying the intermediates. In such cases, we included products of all the steps except the ones which are not isolated or not characterized to represent 'real' molecules. In addition, to signify the facileness of synthesis we have annotated

each compound with the number of pots required for its synthesis, rather than the number of synthetic steps.

For instance, representative compound **1** (Fig. 2A) in URDL is synthesized in 2 steps; UR followed by post-Ugi Povarov-type reaction [41]. However, both steps can be carried out in a single pot without the need of isolation of the Ugi products. Thus, only the final molecules belonging to **1** substructure are included in the library and annotated with 1 pot synthesis. The Ugi-product intermediates which were neither isolated nor characterized, are not included in the library even though they were obviously synthesized during the 1 pot process. Likewise, Fig. 2B represents a cascade reaction where the synthesized Ugi-products are converted to molecule **2** in a single pot without isolation [42]. Thus,
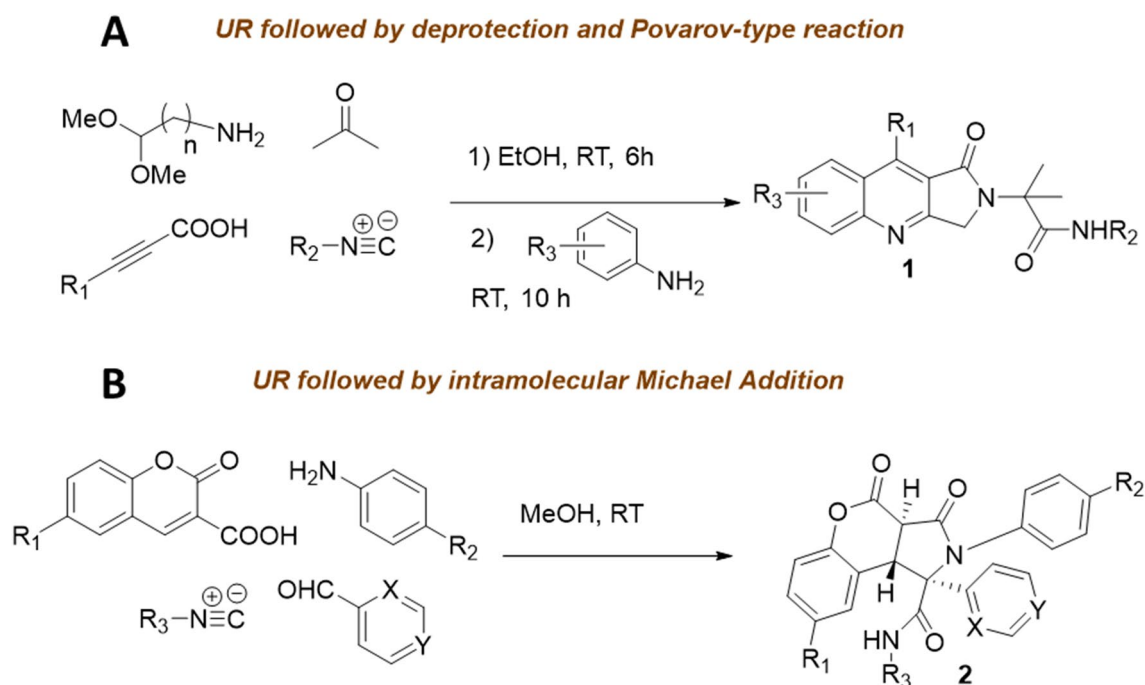

Fig. 2 Representative examples of products of UR and post-Ugi modification

only the molecules belonging to structure **2** were included in the library.

In certain reports, building blocks are synthesized and isolated to obtain the desired Ugi product. In such cases, the steps involved in the synthesis of a building block are added to the total pot count, but the resulting building block is not part of URDL. For example, isocyanide **3** is synthesized in 3 pots to obtain pyrrolidone derivatives **4** (Fig. 3A) [43]. Similarly, the structure of aldehyde **5** required for the synthesis of chromenepyrrole scaffold **7** (Fig. 3B) [44], was not included in the library although steps in its synthesis were counted towards the synthesis of molecules **6** and **7,** which are part of URDL. In rare cases where clarity regarding the source of

building blocks is not available, the later are assumed to be commercially available, and pots are numbered accordingly.

Many literature reports describe structurally close analogues with a minor variation in ring-substituents. In such situations, to avoid manual work and maintain diversity, we omitted certain close analogues with minor structural variations (e.g. methyl *vs* ethyl). We believe such omissions may not affect the VS results significantly, and a medicinal chemist would be able to design and access such missing analogues during SAR exploration. Since the intended application of URDL is in drug discovery, we retained only small molecules (MW $\leq$ 900 Da) lacking general reactivity. This curation process resulted in the Ugi reaction-derived library (URDL) consisting of about 5773
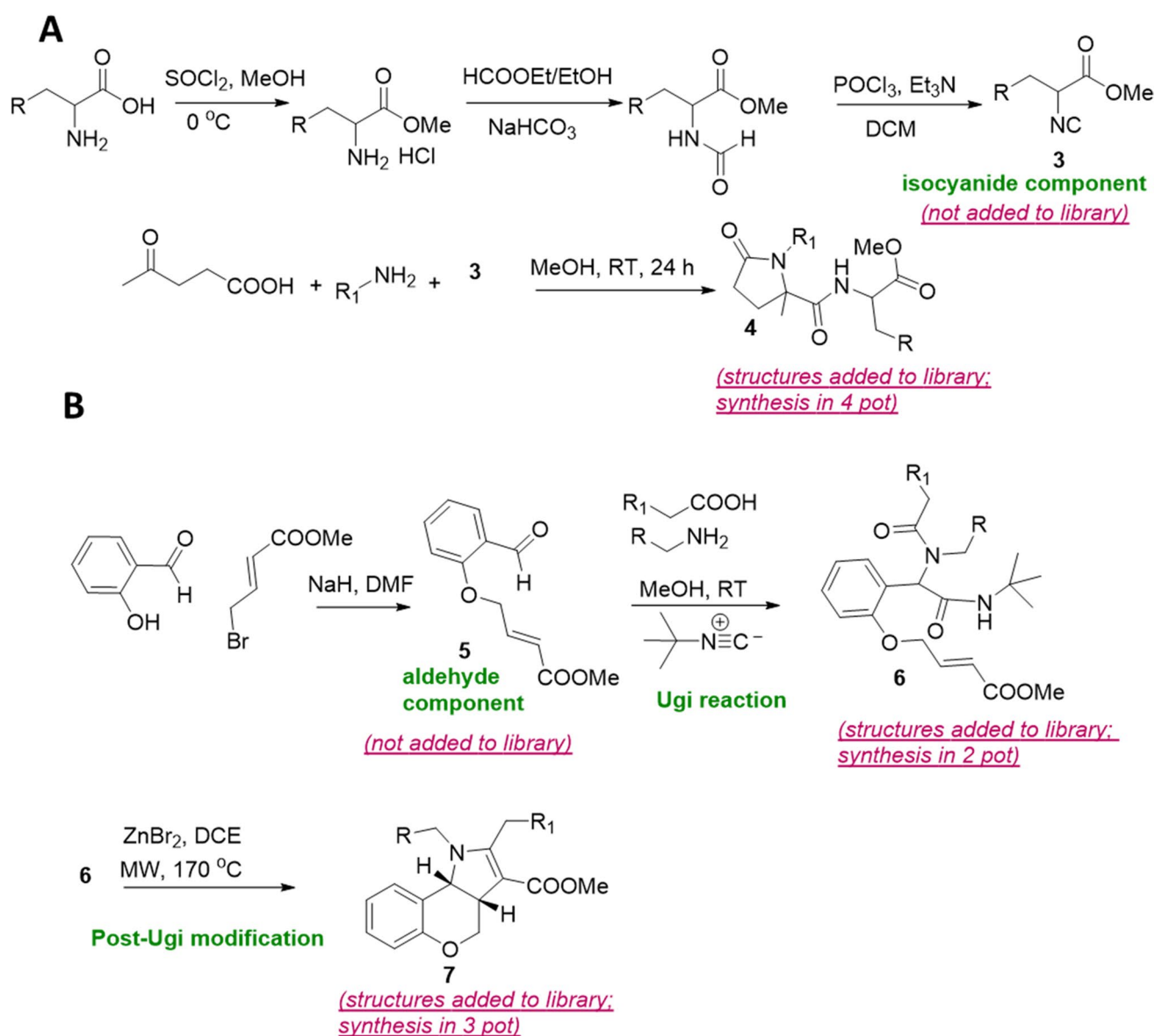

Fig. 3 Examples of URDL molecules where synthesis of UR building requires additional steps

molecules obtained from 274 references. About 92% of the molecules in URDL can be synthesized in either one or two pots (Fig. 4) thus, signifying the synthetic tractability of the library. Additionally, 85% of these molecules appeared in the last decade and hence represent the recent developments in this area.

Despite its small size, URDL has several advantages over commercially available libraries for VS application. For example, the URDL molecules are cherry-picked from high-impact organic chemistry literature with robust structural validation data of the synthesized molecules using spectroscopy and X-ray crystallography. The URDL molecules are synthetically tractable with high atom and pot economy and essentially from commercially available inexpensive building blocks. The conditions reported for synthesising these molecules are often mild, catalytic, and facile enough to be carried out by novice chemists. In many cases, Ugi adducts are known to precipitate from the reaction mixture, thus reducing the workup and purification steps. Moreover, the biological activity of most URDL molecules is not reported despite the presence of novel structural features in these molecules (vide infra). Thus, URDL has additional value in terms of unexplored chemical space for drug discovery. The availability of information on the number of pots required for the synthesis of URDL members may serve as one of the important criteria when shortlisting compounds for synthesis and experimental validation.
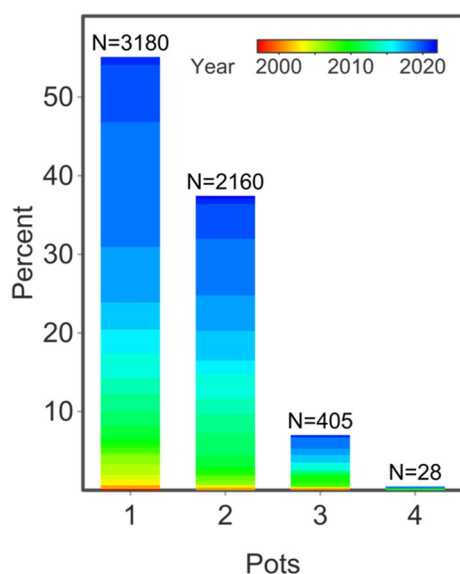

Fig. 4 Percentage of molecules vs the number of pots required for their synthesis. The bars are coloured according to the year in which molecule is reported depicting most of the reports to be recent

Physicochemical profiling

Physicochemical properties of molecules such as size, shape, polarity, and lipophilicity play important role in drug development. Indeed, the molecular descriptors such as molecular weight (MW), partition coefficient (clogP), number of H-bond donors/acceptors (HBD/HBA), topological surface area (TPSA), number of rotatable bonds (RB), fraction of sp3 carbons (Fsp3), are found to correlate with solubility, bioavailability, cell permeability, clinical success, and toxicity [45–52]. Thus, property-based criteria such as Lipinski's 'rule-of-five'[49] and Veber's rule [50] are widely used to estimate the 'druglikeness' of a molecule, albeit with known limitations [53, 54]. Certain molecular properties are also desirable for targeting a particular receptor or organ [55–58]. For instance, drugs crossing BBB are primarily restricted to a property space occupied by small, uncharged, and lipophilic molecules [59, 60]. The presence of primary amines and molecular globularity is reported to play a significant role in facilitating the entry of molecules in Gram-negative bacteria [61]. We have recently shown that the sum of basic and aromatic nitrogen (SBAN) is a key descriptor, among other properties, required for potent antimalarial activity [62]. Thus, understanding the physicochemical property space of a molecular library may assist in the identification of potential targets/diseases for its application.

For physicochemical characterization of URDL we calculated key properties of the molecules and compared it with the property-space of oral drugs as a reference. A recently compiled library of orally used small drugs (MW < 900 Da) by us [62], was updated with the oral drugs approved in the year 2021. Thus, the oral drug library consists of 1998 FDA-approved drug molecules with proven oral bioavailability. Since UR usually yields a dipeptide-like molecules that may serve as the inhibitor of protein–protein interactions (iPPI), we also compared URDL with the publicly available iPPI library [63] consisting of 3853 molecules. The t-test was used to determine the statistical significance among the three categories.

First, we compared the druglikeness of the three libraries using the widely used Lipinski's and Veber's rule. Not surprisingly, 91.4% of oral drugs cleared the criteria for Lipinski's rule while the percentage was lower (85.5%) for the URDL molecules (Table 1). Only, 74.3% molecules in the iPPI library passed the druglikeness criteria based on Lipinski's rule, which is expected since these molecules usually interact with the larger protein surface and thus, tend to have higher MW. Indeed, all 835 iPPI molecules non-compliant to Lipinski's rule possess MW equal or greater than 500 Da. In contrast, Veber's criteria which propose the cut-off of TPSA ≤ 140 Å^2 and RB ≤ 10, predicted URDL library to have the highest percentage (85.4%) of drug-like molecules, followed by oral drugs and iPPI (Table 1). Based

Table 1 The percentage of molecules meeting the criteria of druglikeness, and quality based on different cheminformatics rules

	Lipinski's rule (%)	Veber's rules (%)	PAINS failure (%)	Datawarrior 'nasty' functions (aromatic nitro included) (%)	Datawarrior 'nasty' functions (aromatic nitro excluded) (%)
Oral drugs	91.4	81.7	6.9	12.8	9.7
URDL	85.5	85.4	5.0	13.8	6.2
iPPI	74.3	82.1	18.8	11.8	7.2

on these two rules, it can be concluded that URDL molecules are closer to oral drugs in terms of druglikeness and are expected to have optimal oral bioavailability.

One of the criteria to judge the quality of screening compounds is to look for the reactive functional groups and structural motifs that may interfere with the biochemical assay readouts thus appearing as 'frequent hitters'. Such molecules were termed as pan assay interference compounds (PAINS) by Baell and Holloway [64] who proposed to exclude such compounds using a set of substructure filters. Similarly, Bruns and Watson from Eli Lilly labs also proposed a set of 275 rules to identify promiscuous compounds [65]. Thus, to obtain 'clean' molecules for screening, PAINS structural alerts and other rules are often used. However, these rules and alerts are not without limitations and should not be applied fastidiously without context [66–69]. For example, the PAINS filter should not be used in phenotypic screening or when looking for covalent inhibitors [66]. Nonetheless, there is a broad consensus that molecules possessing PAINS and 'nasty' functions should be flagged early and must be carefully validated before advancing them in the drug discovery pipeline.

We used a recently reported open-source Konstanz Information Miner (KNIME) workflow [70, 71] to identify molecules with PAINS feature in our libraries. Only 5% of URDL molecules have PAINS feature which is lower than the proportion of FDA-approved oral drugs (6.9%) that failed the test (Table 1). The iPPI library displayed 19% failure rate warranting cautious use and interpretation of PAINS alerts when applied to discovery of iPPIs. We also employed the open-source Datawarrior program [72] to identify 'nasty' or reactive functional groups defined by the medicinal chemists at Actelion [73]. The failure rate in terms of such problematic groups is comparable (~ 12–14%) in all three libraries. While molecules belonging to oral drugs and iPPI possess a variety of reactive functions, URDL molecules have limited types for such moieties (Supplementary Information Figure S1). An aromatic nitro group seems to be the most frequently occurring problematic function in all three libraries especially in case of URDL where ignoring this 'nasty' function reduces the failure rate by almost half (Table 1). The high occurrence of aromatic nitro group in URDL can be explained by the fact that most of these molecules are taken from organic chemistry literature discussing newly optimized reactions conditions. In such studies, authors are expected to demonstrate broad substrate scope and hence frequently use building blocks having electron-donating and electron-withdrawing (such as nitro) functional groups. Thus, URDL library molecules are drug-like and 'clean' when considering oral drugs as standard.

Next, we calculated and compared the average property space of the molecules in the three libraries using Datawarrior [72], an open-source cheminformatics program. The statistical details such as mean, median, p-values, quartiles, standard deviation, for the three libraries are provided in the Supplementary information (Table S1). Interestingly, the mean/median of most of the studied properties of URDL molecules are closer to iPPIs than oral drugs (Fig. 5). Among the three categories, the significantly higher values of MW, clogP, HBA/D, TPSA, RB, and aromatic rings for iPPIs are in line with the earlier reports [74, 75]. This observation is justified by the interactions of iPPIs over a large protein interface rather than smaller well-defined pockets. On average URDL molecules possess higher values for MW (Fig. 5A), clogP (Fig. 5B), HBA (Fig. 5C), and RB (Fig. 5F) than the oral drugs. In contrast, URDL molecules have significantly lower average values for HBD (Fig. 5D), TPSA (Fig. 5E), and Fsp3 (Fig. 5G) in comparison to oral drugs. The molecules belonging to URDL and iPPI, are also structurally more complex (Fig. 5H) as evaluated by the Datawarrior program. This complexity may result from more rings (7-membered or smaller) present in these molecules (Fig. 5I). However, it must be noted that molecular complexity may be calculated in several ways [76, 77]. For instance, oral drugs possess more chiral centres than URDL and iPPI molecules (Fig. 5J), which is another proposed measure of complexity together with Fsp3 [76]. The URDL and iPPI molecules have a higher number of carboaromatic rings (ArC) (Fig. 5K), suggesting these compounds are more disc-like than oral drugs. Similarly, orally used drugs have a lower number of heteroaromatic rings in their structure than the other two libraries (Fig. 5L).

Overall, this analysis indicates that URDL and oral drug libraries are comparable in terms of druglikeness as well as compound quality. However, in terms of physicochemical properties, the URDL molecules are closer to iPPIs.
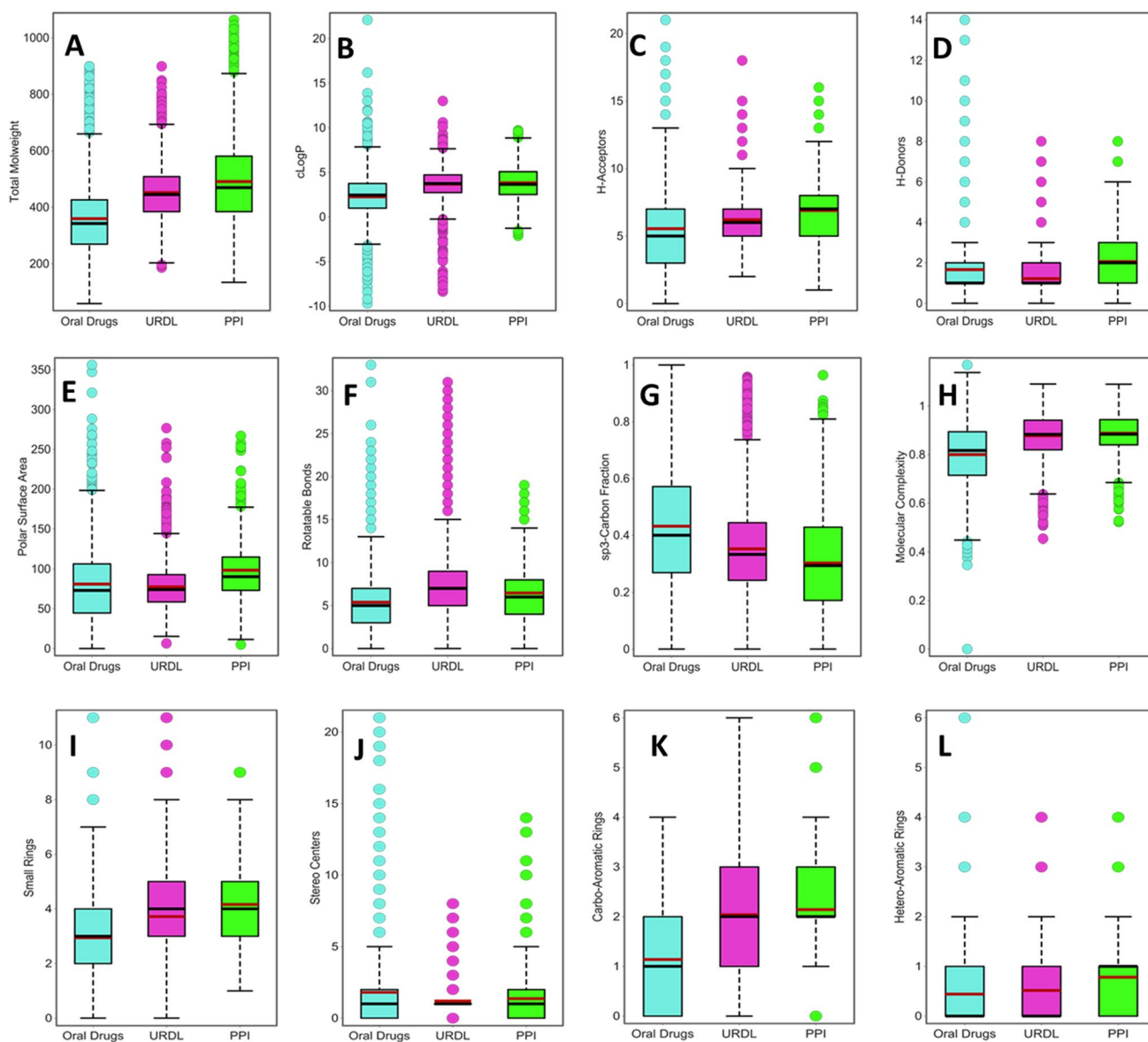
Fig. 5 Box plots displaying average properties of oral drugs (cyan), URDL (pink) and iPPI (green) molecules; **A** molecular weight, **B** calculated logP, **C** hydrogen bond acceptors, D) hydrogen bond donors, **E** topological polar surface area, **F** rotatable bonds, **G** fraction of sp3 carbons, **H** molecular complexity, **I** small rings, **J** stereocenters, **K** carbo-aromatic rings, **L** hetero-aromatic rings. The red and black lines within the boxplots represent the mean and median, respectively

For comparison and visualization of different sets of molecules, medicinal chemists often rely on different dimension reduction approaches [78]. The multi-dimensional information coded in structural descriptors or physicochemical properties can be reduced and projected in two (2D) or three dimensions (3D) for comprehensibility [79–81]. Thus, the molecules bearing structure or property-based similarity are expected to be placed nearer to each other in these projections.

To compare the chemical space of URDL with oral drugs and iPPI molecules, we calculated the SkelSphere descriptors implemented in the Datawarrior program. This descriptor encodes circular spheres of atoms and bonds into a hashed binary fingerprint of 1024 bits together with the stereochemistry and other structural details. The 1024 bits of structural information were then reduced using T-distributed stochastic neighbour embedding (t-SNE) [82, 83], a nonlinear dimensionality reduction technique. Consequently, a 3D projection of chemical space was obtained with the similar molecules being closer to each other than the dissimilar ones. The oral drugs and URDL molecules seem to form separate clusters with limited overlap, suggesting the structural features present in URDL molecules are distinct from the FDA-approved drugs (Fig. 6). On the other hand,

iPPI molecules form several smaller clusters, a few overlapping with either oral drugs or URDL, displaying a broader diversity in these molecules.

Overall, physicochemical profiling suggests that URDL molecules conform to a distinct chemical space in comparison to the existing oral drugs.
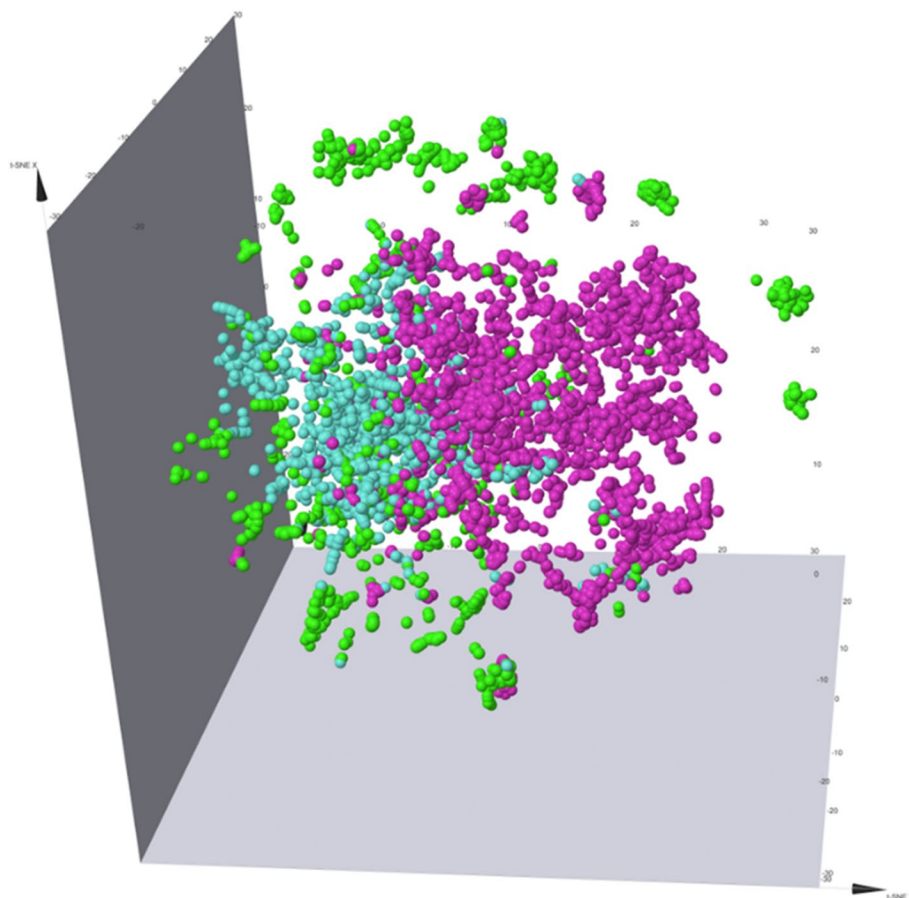
Scaffold and ring analysis

A recent analysis has revealed that FDA-approved drugs lack diverse rings [84]. On the other hand, 1 million ring systems (size 1–4, < 30 atoms) are possible theoretically, 98.6% of which do not exist in big databases like ZINC or ChEMBL [85]. The distinct chemical space in URDL compared to oral drugs (Fig. 6) indicates the presence of unique scaffolds and ring systems in URDL molecules which was also observed during the library curation. Indeed, several novel heterocycles can be generated using UR [27]. For comparison, the 'most central rings', the ring closest to the topological centre of the given molecule, were extracted from both URDL and oral drugs using Datawarrior. This analysis resulted in 417 and 316 'most central rings' from oral drugs and URDL, respectively. Among the top ten most frequently appearing rings,

benzene, piperidine, and pyrrolidine are common in both libraries (Fig. 7). One-third (103) of the rings in URDL display SkelSphere similarity of 80% or higher to the rings extracted from oral drugs rings, and only 62 rings (19.6%) are structurally identical. The t-SNE plot derived from the SkelSphere descriptor of 'most central rings' (Figs. 8 and 9) reveals several unique ring systems in URDL that are not represented in oral drugs. The majority of these diverse ring systems (Fig. 9) can be obtained in 2 pot reaction sequence from Ugi adducts indicating facile access to these rings. Additionally, these rings possess varying sizes, lipophilicity, H-bonding capacity (HBA/HBD), PSA, and globularity volume, suggesting these may be exploited to design ligands against different protein binding pockets. A substructure search using the rings displayed in Fig. 9 in the ChEMBL database resulted in no hits underscoring the structural novelty of URDL ring systems.

Together, this analysis confirms that the URDL library consists of molecules based on novel ring systems that are synthetically tractable and remain unexplored in drug design.

Fig. 6 t-SNE 3D plot for oral dugs, URDL and iPPI library molecules shown as cyan, pink and green spheres, respectively. The multi-dimensional information stored in the SkelSphere descriptor of the molecules was reduced using t-SNE algorithm of Datawarrior to display distinct clusters of similar molecules
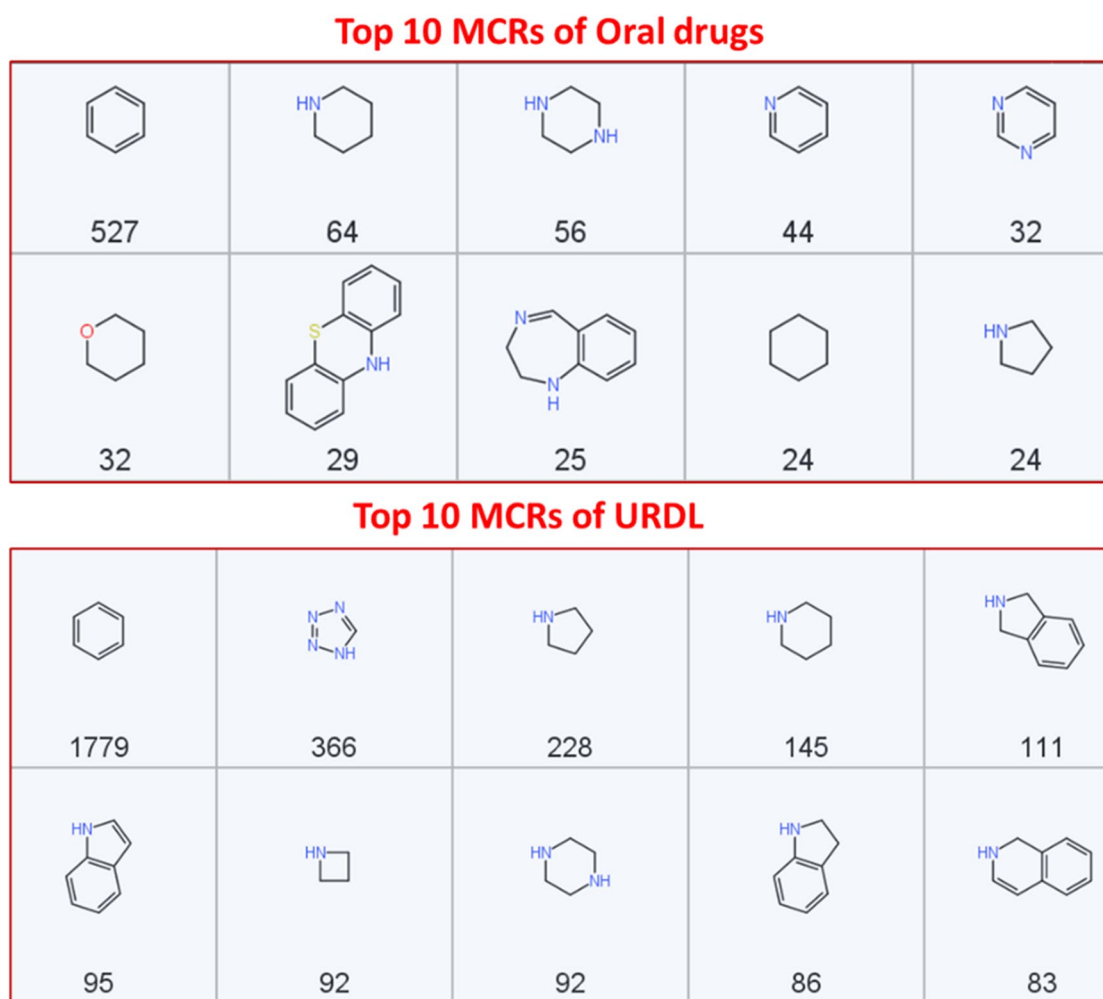
Top 10 MCRs of Oral drugs

527	64	56	44	32
32	29	25	24	24

Top 10 MCRs of URDL

1779	366	228	145	111
95	92	92	86	83

Fig. 7 Top 10 'most central rings' appearing in oral drugs and URDL

UR in the synthesis of drugs and their analogues

Some of the URDL scaffolds and molecules also showed structural overlap with oral drugs (Figs. 6 and 8) indicating the latter may be accessed in an efficient way using UR. To find the examples of oral drugs or their close analogues in URDL we performed a similarity search between URDL and oral drugs using SkelSphere descriptor. A total of 114 of the URDL molecules were found to be structurally identical or similar (≥ 0.75 SkelSphere similarity index) to the 46 unique oral drugs (Supplementary Information Figure S2 and S3). Among these, the two-pot gram-scale synthesis of praziquantel (**8**) is a well-known example displaying the efficiency of UR in drug synthesis (Fig. 10) [86]. A two-pot gram-scale enantioselective synthesis of R-lacosamide (**9**) has been achieved recently via Ugi3CR [87]. Similarly, a two-pot synthesis of an epimer of Tadalafil (**10**) is described using the UR-derived intermediate followed by cyclization [88]. In addition, close analogues of several other drugs,

such as Roxatidine [89], Vorinostat [89], Racecadotril [90], and Pinazepam [91], are present in URDL (Supplementary Information Figure S2 and S3). Given the efficient synthesis of URDL molecules, it would be interesting to synthesize structurally similar analogues of FDA-approved drugs and test them against the corresponding targets. Analogously, one can conduct a similarity analysis of any other molecule of interest against URDL to find synthetically tractable close analogues. Such efforts may also result in scaffold hopping [92, 93], an important strategy in drug design that may be useful to overcome intellectual property constraints.

Conclusions

In conclusion, we have curated a library of molecules derived from Ugi reaction, cherry-picked from the recently reported literature. The synthesis of the majority of the URDL molecules involves mild reaction conditions with

Fig. 8 t-SNE plot of 'most central rings' of oral drugs (cyan) and URDL (pink) using SkelSphere descriptor
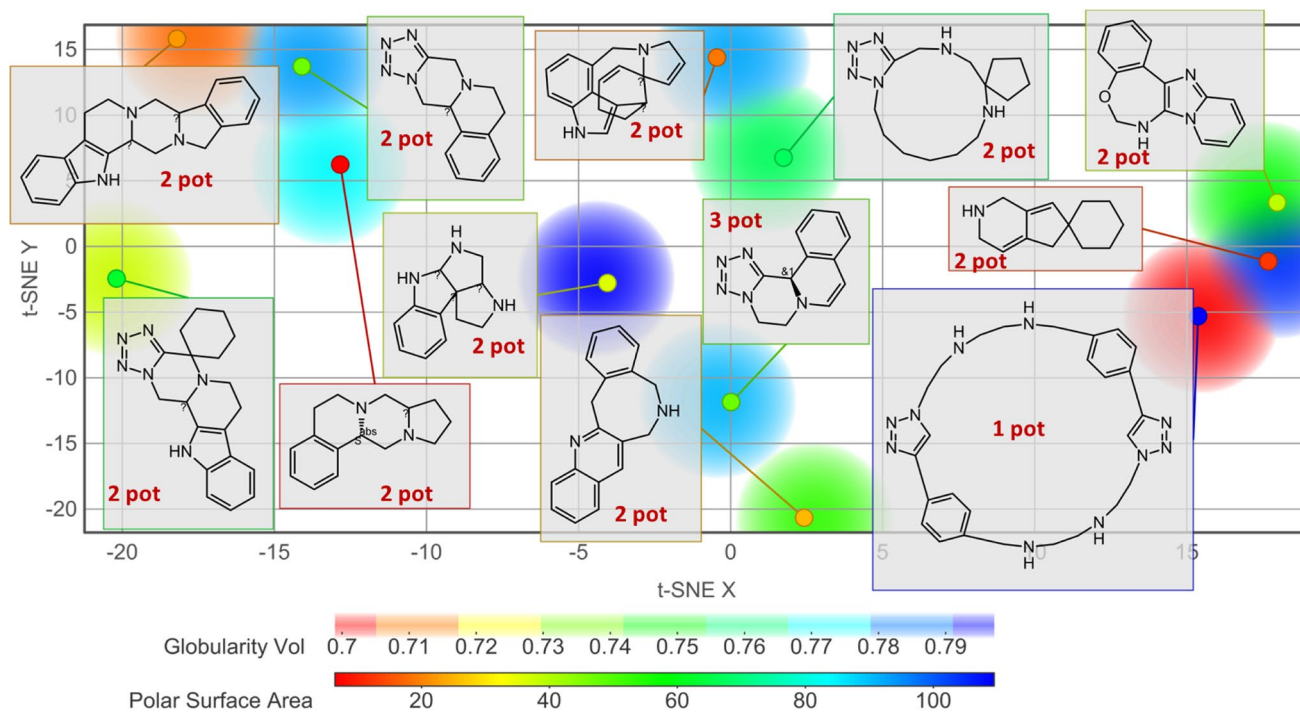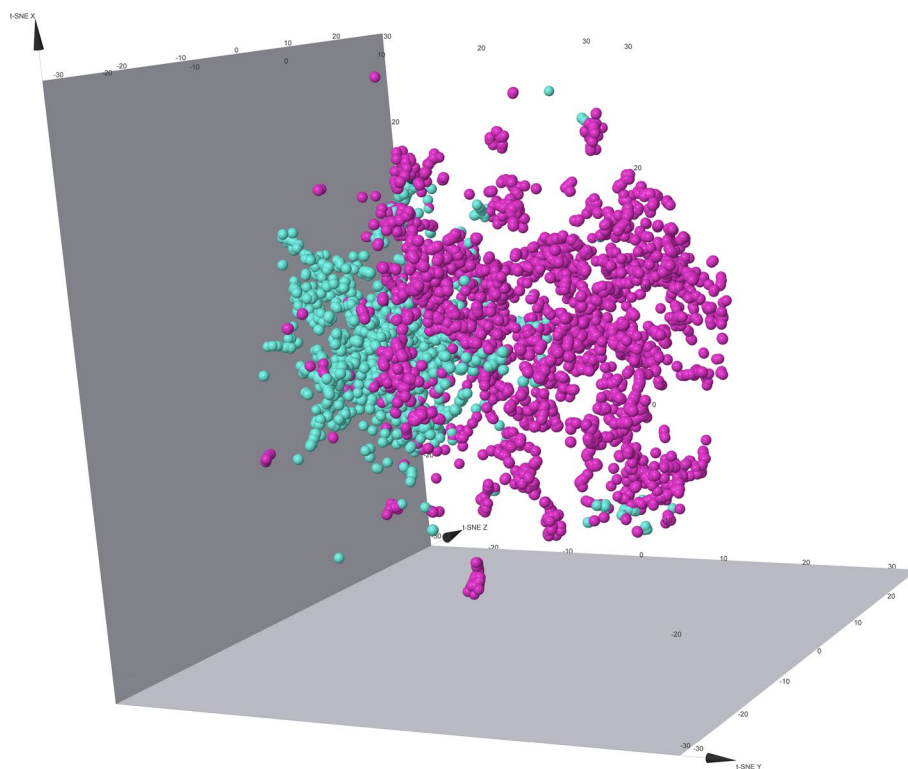

Fig. 9 A 2D t-SNE plot of the representative set of diverse URDL rings. The circle represents rings in the chemical space and colour is coded by TPSA. The background of the circles is coloured according to the globularity volume of the rings
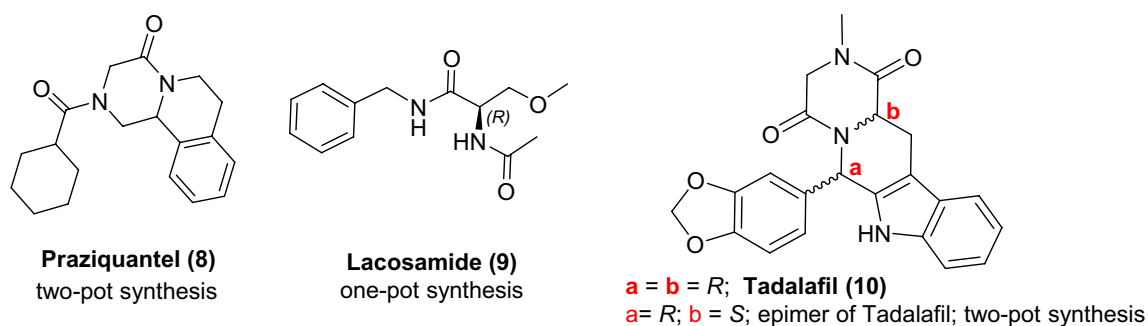
Fig. 10 Drug molecules and their analogues present in URDL, that can be efficiently synthesized using UR

high pot economy (1–2 pots) and reported methodology. Thus, large samples of the molecules can be obtained cost-effectively for the experimental validation of the in-silico screening results. The comparison with oral drugs shows that URDL consists of drug-like molecules but occupy a distinct chemical space in terms of structural descriptor. Additionally, URDL compounds show a lower frequency of PAINS alerts and reactive functional groups compared to oral drugs. In terms of property space, URDL molecules are closer to known inhibitors of PPIs. Many of the URDL molecules consist of novel ring systems absent in the currently approved drugs or in ChEMBL database. Several oral drugs and their close analogues are also included in URDL, suggesting that UR and post-Ugi modification may efficiently synthesise these molecules and result in scaffold hoping. Thus, the URDL molecules represent synthetically tractable unexplored chemical space fit for the purpose of VS and similarity searching. The URDL library is freely accessible as a part of the supplementary information of this manuscript.

Materials and methods

Library curation

The PubMed search was performed using term 'Ugi reaction' in Title/Abstract field to identify literature reports discussing the application of UR. The cross-references in the review articles were also used to identify relevant original papers. The structures of the molecules were drawn manually or interpreted from the IUPAC nomenclature provided in the original papers using Osiris Datawarrior (v. 5.5.0) [72]. The synthetic schemes were carefully studied to identify the number of pots involved in synthesising different molecules reported in the research article. The molecules with MW > 900 were filtered out. The URDL library was finally annotated with detailed references, DOIs and the number of pots required for the synthesis.

Cheminformatic analysis

The URDL library (total molecules = 5773) was appended with oral drugs (total molecules = 1998) [62] and iPPI database [63] (total molecules = 3853) annotated with dataset names (Oral drugs, iPPI, URDL). For the sake of reproducibility, all processing and cheminformatics analysis was performed in open-source programmes, KNIME analytics platform 4.5.2 [70] and Datawarrior 5.5.0 [72].

The KNIME workflow developed by Bren and coworkers (https://gitlab.com/Jukic/knime_medchem_filters/) [71] was used to identify PAINS alerts in all three libraries. The compounds comprising reactive functional groups were identified using the 'nasty functions' [73] feature of Datawarrior 5.5.0. The SkelSphere descriptor, physicochemical descriptors, and other properties shown in Fig. 5 were calculated using Datawarrior. The boxplots and the mean/median and p-values, were obtained using the 2D plot function of Datawarrior. The rings were extracted using the 'Most Central Ring' feature of Datawarrior. The t-SNE plots were generated with perplexity = 40, source dimensions = 50 and iterations = 1000. The similarity comparison between the oral drugs and URDL library was performed using the threshold of 0.75 SkelSphere-based similarity index. The ChEMBL database (v. 30) was searched within Datawarrior employing the 'superstructures of' option with various ring structures as queries.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest Authors declare no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

References

1. Mayr LM, Bojanic D (2009) Novel trends in high-throughput screening. Curr Opin Pharmacol 9:580–588. https://doi.org/10.1016/J.COPH.2009.08.004

2. Wouters OJ, McKee M, Luyten J (2020) Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA—J Am Med Assoc 323:844–853. https://doi.org/10.1001/jama.2020.1166

3. Talevi A (2018) Computer-aided drug design: an overview. Methods Mol Biol 1762:1–19. https://doi.org/10.1007/978-1-4939-7756-7_1

4. Boss C, Hazemann J, Kimmerlin T et al (2017) The screening compound collection: a key asset for drug discovery. Chimia (Aarau) 71:667–677. https://doi.org/10.2533/chimia.2017.667

5. Lionta E, Spyrou G, Vassilatis D, Cournia Z (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem 14:1923–1938. https://doi.org/10.2174/1568026614666140929124445

6. Lyne PD (2002) Structure-based virtual screening: an overview. Drug Discov Today 7:1047–1055. https://doi.org/10.1016/S1359-6446(02)02483-2

7. Lyu J, Wang S, Balius TE et al (2019) Ultra-large library docking for discovering new chemotypes. Nature 566:224–229. https://doi.org/10.1038/s41586-019-0917-9

8. Sundriyal S, Viswanad B, Ramarao P et al (2008) New PPARγ ligands based on barbituric acid: virtual screening, synthesis and receptor binding studies. Bioorganic Med Chem Lett 18:4959–4962. https://doi.org/10.1016/j.bmcl.2008.08.028

9. Gorgulla C, Boeszoermenyi A, Wang ZF et al (2020) An open-source drug discovery platform enables ultra-large virtual screens. Nature 580:663. https://doi.org/10.1038/S41586-020-2117-Z

10. Hoffmann T, Gastreich M (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today 24:1148–1156. https://doi.org/10.1016/J.DRUDIS.2019.02.013

11. Walters WP (2019) Virtual Chemical Libraries. J Med Chem 62:1116–1124

12. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182. https://doi.org/10.1021/CI049714

13. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100-1107. https://doi.org/10.1093/nar/gkr777

14. Bento AP, Gaulton A, Hersey A et al (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42:1083–1090. https://doi.org/10.1093/nar/gkt1031

15. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 46:D1074. https://doi.org/10.1093/NAR/GKX1037

16. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388–D1395. https://doi.org/10.1093/NAR/GKAA971

17. Hersey A, Chambers J, Bellis L et al (2015) Chemical databases: curation or integration by user-defined equivalence? Drug Discov Today Technol 14:17–24. https://doi.org/10.1016/j.ddtec.2015.01.005

18. Saldívar-González FI, Huerta-García CS, Medina-Franco JL (2020) Chemoinformatics-based enumeration of chemical libraries: a tutorial. J Cheminform. https://doi.org/10.1186/s13321-020-00466-z

19. Grygorenko OO, Radchenko DS, Dziuba I et al (2020) Generating multibillion chemical space of readily accessible screening compounds. Science. https://doi.org/10.1016/J.ISCI.2020.101681

20. Fialková V, Zhao J, Papadopoulos K et al (2021) LibINVENT: reaction-based generative scaffold decoration for in silico library design. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.1c00469

21. Hayashi Y (2016) Pot economy and one-pot synthesis. Chem Sci 7:866–880. https://doi.org/10.1039/c5sc02913a

22. Biggs-Houck JE, Younai A, Shaw JT (2010) Recent advances in multicomponent reactions for diversity-oriented synthesis. Curr Opin Chem Biol 14:371–382. https://doi.org/10.1016/j.cbpa.2010.03.003

23. Domling A, Wang W, Wang K (2012) Chemistry and biology of multicomponent reactions. Chem Rev 112:3083–3135. https://doi.org/10.1021/cr100233r

24. Elders N, Van Der Born D, Hendrickx LJD et al (2009) The efficient one-pot reaction of up to eight components by the union of multicomponent reactions. Angew Chem—Int Ed 48:5856–5859. https://doi.org/10.1002/anie.200902683

25. Ugi I (1962) The α-addition of immonium ions and anions to isonitriles accompanied by secondary reactions. Angew Chem Int Ed English 1:8–21. https://doi.org/10.1002/anie.196200081

26. Ugi I, Steinbrückner C (1960) Über ein neues Kondensations-Prinzip. Angew. Chemie 72:267–268. https://doi.org/10.1002/ange.19600720709

27. Bariwal J, Kaur R, Voskressensky LG, Van der Eycken EV (2018) Post-Ugi cyclization for the construction of diverse heterocyclic compounds: recent updates. Front Chem. https://doi.org/10.3389/fchem.2018.00557

28. Patil P, Khoury K, Herdtweck E, Domling A (2014) A universal isocyanide for diverse heterocycle syntheses. Org Lett 16:5736–5739. https://doi.org/10.1021/ol5024882

29. Tripolitsiotis NP, Thomaidi M, Neochoritis CG (2020) The Ugi three-component reaction; a valuable tool in modern organic synthesis. Eur J Org Chem 2020:6525–6554. https://doi.org/10.1002/ejoc.202001157

30. Gazzotti S, Rainoldi G, Silvani A (2019) Exploitation of the Ugi-Joullié reaction in drug discovery and development. Exp Opin Drug Discov 14:639–652. https://doi.org/10.1080/17460441.2019.1604676

31. Lei J, Meng JP, Tang DY et al (2018) Recent advances in the development of polycyclic skeletons via Ugi reaction cascades. Mol Divers 22:503–516. https://doi.org/10.1007/S11030-017-9811-2

32. Dömling A (2022) Innovations and Inventions: why was the Ugi reaction discovered only 37 years after the Passerini reaction? J Org Chem. https://doi.org/10.1021/acs.joc.2c00792

33. DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 47:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

34. Abdelraheem EMM, Khaksar S, Kurpiewska K et al (2018) Two-step macrocycle synthesis by classical Ugi reaction. J Org Chem 83:1441–1447. https://doi.org/10.1021/acs.joc.7b02984

35. Tao Y, Wang Z, Tao Y (2019) Polypeptoids synthesis based on Ugi reaction: advances and perspectives. Biopolymers 110:e23288. https://doi.org/10.1002/bip.23288

36. Yang B, Zhao Y, Wei Y et al (2015) The Ugi reaction in polymer chemistry: syntheses, applications and perspectives. Polym Chem. https://doi.org/10.1039/c5py01398d

37. Tandi M, Sundriyal S (2021) Recent trends in the design of antimicrobial agents using Ugi-multicomponent reaction. J Indian Chem Soc 98:100106. https://doi.org/10.1016/j.jics.2021.100106

38. Fouad MA, Abdel-Hamid H, Ayoup MS (2020) Two decades of recent advances of Ugi reactions: synthetic and pharmaceutical applications. RSC Adv 10:42644–42681. https://doi.org/10.1039/d0ra07501a

39. Musonda CC, Taylor D, Lehman J et al (2004) Application of multi-component reactions to antimalarial drug discovery. Part 1: parallel synthesis and antiplasmodial activity of new 4-aminoquinoline Ugi adducts. Bioorganic Med Chem Lett 14:3901–3905. https://doi.org/10.1016/j.bmcl.2004.05.063

40. Reymond JL, Awale M (2012) Exploring chemical space for drug discovery using the chemical universe database. ACS Chem Neurosci 3:649–657. https://doi.org/10.1021/cn3000422

41. Ghoshal A, Yugandhar D, Nanubolu JB, Srivastava AK (2017) An Efficient one-pot synthesis of densely functionalized fused-quinolines via sequential Ugi4CC and acid-mediated povarov-type reaction. ACS Comb Sci 19:600–608. https://doi.org/10.1021/acscombsci.7b00095

42. Che C, Li S, Jiang X et al (2010) One-pot syntheses of chromeno[3,4- c]pyrrole-3,4-diones via Ugi-4CR and intramolecular Michael addition. Org Lett 12:4682–4685. https://doi.org/10.1021/ol1020477

43. Khalesi M, Halimehjani AZ, Martens J et al (2019) Synthesis of a novel category of pseudo-peptides using an Ugi three-component reaction of levulinic acid as bifunctional substrate, amines, and amino acid-based isocyanides. Beilstein J Org Chem 1582(15):852–857. https://doi.org/10.3762/BJOC.15.82

44. Srinivasulu V, Sieburth SMN, El-Awady R et al (2018) Post-Ugi cascade transformations for accessing diverse chromenopyrrole collections. Org Lett 20:836–839. https://doi.org/10.1021/acs.orglett.7b03986

45. Ritchie TJ, Macdonald SJF (2009) The impact of aromatic ring count on compound developability - are too many aromatic rings a liability in drug design? Drug Discov Today 14:1011–1020. https://doi.org/10.1016/j.drudis.2009.07.014

46. Ritchie TJ, MacDonald SJF, Peace S et al (2012) The developability of heteroaromatic and heteroaliphatic rings—do some have a better pedigree as potential drug molecules than others? Medchemcomm 3:1062–1069. https://doi.org/10.1039/c2md20111a

47. Ritchie TJ, MacDonald SJF, Young RJ, Pickett SD (2011) The impact of aromatic ring count on compound developability: further insights by examining carbo- and hetero-aromatic and -aliphatic ring types. Drug Discov Today 16:164–171. https://doi.org/10.1016/j.drudis.2010.11.014

48. Lovering F, Bikker J, Humblet C (2009) Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem 52:6752–6756. https://doi.org/10.1021/jm901241e

49. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 23:3–25. https://doi.org/10.1016/s0169-409x(96)00423-1

50. Veber DF, Johnson SR, Cheng HY et al (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45:2615–2623. https://doi.org/10.1021/jm020017n

51. Young RJ, Green DVS, Luscombe CN, Hill AP (2011) Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. Drug Discov Today 16:822–830. https://doi.org/10.1016/j.drudis.2011.06.001

52. Tinworth CP, Young RJ (2020) Facts, patterns, and principles in drug discovery: appraising the rule of 5 with measured physicochemical data. J Med Chem 63:10091–10108. https://doi.org/10.1021/acs.jmedchem.9b01596

53. Shultz MD (2019) Two decades under the influence of the rule of five and the changing properties of approved oral drugs. J Med Chem 62:1701–1714. https://doi.org/10.1021/acs.jmedchem.8b00686

54. Petit J, Meurice N, Kaiser C, Maggiora G (2012) Softening the rule of five—where to draw the line? Bioorganic Med Chem 20:5343–5351. https://doi.org/10.1016/j.bmc.2011.11.064

55. Vieth M, Sutherland JJ (2006) Dependence of molecular properties on proteomic family for marketed oral drugs. J Med Chem 49:3451–3453. https://doi.org/10.1021/jm0603825

56. Young RJ, Leeson PD (2018) Mapping the efficiency and physicochemical trajectories of successful optimizations. J Med Chem 61:6421–6467. https://doi.org/10.1021/acs.jmedchem.8b00180

57. Macielag MJ (2012) Chemical properties of antimicrobials and their uniqueness. In: Dougherty TJ, Pucci MJ (eds) Antibiotic discovery and development. Springer, New York Dordrecht Heidelberg London, pp 793–820

58. Leeson PD, Davis AM (2004) Time-related differences in the physical property profiles of oral drugs. J Med Chem 47:6338–6348. https://doi.org/10.1021/jm049717d

59. Wager TT, Chandrasekaran RY, Hou X et al (2010) Defining desirable central nervous system drug space through the alignment of molecular properties, in vitro ADME, and safety attributes. ACS Chem Neurosci 1:420–434. https://doi.org/10.1021/cn100007x

60. Doan KMM, Humphreys JE, Webster LO et al (2002) Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. J Pharmacol Exp Ther 303:1029–1037. https://doi.org/10.1124/jpet.102.039255

61. Richter MF, Drown BS, Riley AP et al (2017) Predictive compound accumulation rules yield a broad-spectrum antibiotic. Nature 545:299–304. https://doi.org/10.1038/nature22308

62. Bhanot A, Sundriyal S (2021) Physicochemical profiling and comparison of research antiplasmodials and advanced stage antimalarials with oral drugs. ACS Omega 6:6424–6437. https://doi.org/10.1021/acsomega.1c00104

63. Labbé CM, Kuenemann MA, Zarzycka B et al (2016) iPPI-DB: an online database of modulators of protein-protein interactions. Nucleic Acids Res 44:D542–D547. https://doi.org/10.1093/NAR/GKV982

64. Baell JB, Holloway GA (2010) New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem 53:2719–2740. https://doi.org/10.1021/jm901137j

65. Bruns RF, Watson IA (2012) Rules for identifying potentially reactive or promiscuous compounds. J Med Chem 55:9763–9772. https://doi.org/10.1021/jm301008n

66. Senger MR, Fraga CAM, Dantas RF, Silva FP (2016) Filtering promiscuous compounds in early drug discovery: is it a good idea? Drug Discov Today 21:868–872. https://doi.org/10.1016/j.drudis.2016.02.004

67. Capuzzi SJ, Muratov EN, Tropsha A (2017) Phantom PAINS: problems with the utility of alerts for P an- assay in terference compound S. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.6b00465

68. Jasial S, Hu Y, Bajorath J (2017) How frequently are pan-assay interference compounds active? Large-scale analysis of screening data reveals diverse activity profiles, low global hit frequency, and many consistently inactive compounds. J Med Chem. https://doi.org/10.1021/acs.jmedchem.7b00154

69. Baell JB, Nissink JWM (2018) Seven year itch: pan-assay interference compounds (PAINS) in 2017—utility and limitations. ACS Chem Biol 13:36–44. https://doi.org/10.1021/acschembio.7b00903

70. Nicolas Dill F, Gabriel TR et al (2007) KNIME: the Konstanz information miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R (eds) Studies in classification, data analysis, and knowledge organization (GfKL 2007). Springer, Berlin, Heidelberg

71. Gomez-Puertas P, Kralj S, Jukič M, Bren U (2022) Comparative analyses of medicinal chemistry and cheminformatics filters with accessible implementation in Konstanz information miner (KNIME). Int J Mol Sci 23:572723–5727. https://doi.org/10.3390/IJMS23105727

72. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55:460–473. https://doi.org/10.1021/ci500588j

73. Datawarrior User Forum. https://openmolecules.org/forum/index.php?t=msg&th=627&start=0&. Accessed 21 Nov 2022

74. Sperandio O, Reynès CH, Camproux AC, Villoutreix BO (2010) Rationalizing the chemical space of protein-protein interaction inhibitors. Drug Discov Today 15:220–229. https://doi.org/10.1016/j.drudis.2009.11.007

75. Kuenemann MA, Labbé CM, Cerdan AH (2016) Sperandio O (2016) Imbalance in chemical space: how to facilitate the identification of protein-protein interaction inhibitors. Sci Reps 61(6):1–17. https://doi.org/10.1038/srep23815

76. Méndez-Lucio O, Medina-Franco JL (2017) The many roles of molecular complexity in drug discovery. Drug Discov Today 22:120–126. https://doi.org/10.1016/j.drudis.2016.08.009

77. Leach AR, Hann MM (2011) Molecular complexity and fragment-based drug discovery: ten years on. Curr Opin Chem Biol 15:489–496

78. Osolodkin DI, Radchenko EV, Orlov AA et al (2015) Progress in visual representations of chemical space. Exp Opin Drug Discov 10:959–973. https://doi.org/10.1517/17460441.2015.1060216

79. Awale M, Van Deursen R, Reymond JL (2013) MQN-mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. J Chem Inf Model 53:509–518. https://doi.org/10.1021/CI300513M/ASSET/IMAGES/CI-2012-00513M_M003.GIF

80. Nguyen KT, Blum LC, Van Deursen R, Reymond J (2009) Classification of organic molecules by molecular quantum numbers. ChemMedChem. https://doi.org/10.1002/cmdc.200900317

81. Ruddigkeit L, Van Deursen R, Blum LC, Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 52:2864–2875. https://doi.org/10.1021/ci300415d

82. Van Der Maaten L (2015) Accelerating t-SNE using tree-based algorithms. J Mach Learn Res 15:3221–3245

83. Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2625

84. Shearer J, Castro JL, Lawson ADG et al (2022) Rings in clinical trials and drugs: present and future. J Med Chem. https://doi.org/10.1021/acs.jmedchem.2c00473

85. Visini R, Arús-Pous J, Awale M, Reymond JL (2017) Virtual exploration of the ring systems chemical universe. J Chem Inf Model 57:2707–2718. https://doi.org/10.1021/acs.jcim.7b00457

86. Cao H, Liu H, Dömling A (2010) Efficient multicomponent reaction synthesis of the schistosomiasis drug praziquantel. Chem—A Eur J 16:12296–12298. https://doi.org/10.1002/chem.201002046

87. Zhang J, Wang YY, Sun H et al (2020) Enantioselective three-component Ugi reaction catalyzed by chiral phosphoric acid. Sci China Chem 63:47–54. https://doi.org/10.1007/s11426-019-9606-2

88. Jida M, Ballet S (2018) An efficient one-pot synthesis of chiral N-protected 3-substituted (Diketo)piperazines via Ugi-4CR/De-boc/cyclization process. ChemistrySelect 3:1027–1031. https://doi.org/10.1002/slct.201702943

89. Bhela IP, Serafini M, Del Grosso E et al (2021) Tritylamine as an ammonia surrogate in the ugi reaction provides access to unprecedented 5-sulfamido oxazoles using burgess-type reagents. Org Lett 23:3610–3614. https://doi.org/10.1021/acs.orglett.1c01002

90. Borase BB, Godbole HM, Singh GP et al (2019) Application of Ugi three component reaction for the synthesis of quinapril hydrochloride. Synth Commun 50:48–55. https://doi.org/10.1080/00397911.2019.1682168

91. Huang Y, Khoury K, Chanas T, Dömling A (2012) Multicomponent synthesis of diverse 1,4-benzodiazepine scaffolds. Org Lett 14:5916–5919. https://doi.org/10.1021/ol302837h

92. Sun H, Tawa G, Wallqvist A (2012) Classification of scaffold-hopping approaches. Drug Discov Today 17:310–324. https://doi.org/10.1016/j.drudis.2011.10.024

93. Zhao H (2007) Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. Drug Discov Today 12:149–155. https://doi.org/10.1016/j.drudis.2006.12.003