ACS Combinatorial Science

Review

pubs.acs.org/acscombsci

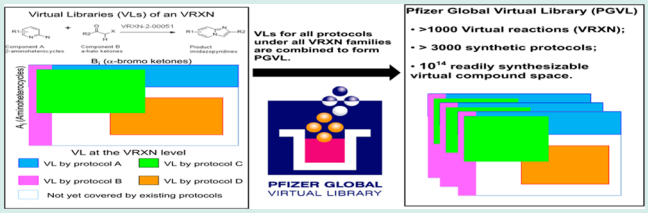
Pfizer Global Virtual Library (PGVL): A Chemistry Design Tool Powered by Experimentally Validated Parallel Synthesis Information

Qiyue Hu,*,† Zhengwei Peng, Scott C. Sutton, Jim Na, Jaroslav Kostrowicki, Bo Yang, Thomas Thacher, Xianjun Kong, Sarathy Mattaparti, Joe Zhongxiang Zhou, Javier Gonzalez, Michele Ramirez-Weinhouse, and Atsuo Kuki

Pfizer Global Research and Development, La Jolla Laboratories, 10770 Science Center Drive, San Diego, California 92121, United States

Supporting Information

ABSTRACT: An unprecedented amount of parallel synthesis information was accumulated within Pfizer over the past 12 years. This information was captured by an informatics tool known as PGVL (Pfizer Global Virtual Library). PGVL was used for many aspects of drug discovery including automated reactant mining and reaction product formation to build a synthetically feasible virtual compound collection. In this report, PGVL is discussed in detail. The chemistry information within PGVL has been used to extract synthesis and design information using an intuitive desktop Graphic User Interface, PGVL Hub. Several real-case examples of PGVL are also presented.



KEYWORDS: *drug discovery, cheminformatics, molecular design, parallel synthesis, combinatorial library, synthesis protocol, knowledge system, Pfizer Global Virtual Library (PGVL), reactant, product, enumeration, canonical parallel synthetic protocols (CPSP), VL reaction registrars, reaction/protocol developers, automated reactant mining objects (ARM Object), virtual reaction objects (VRXN Object), , virtual library (VL)*

INTRODUCTION

Parallel synthesis has been adopted as one of many drug discovery tools by design chemists.[1] As the amount of parallel synthesis and building block information expanded over the years, the need for a tool that enabled easy search and retrieval of that information for the end user has increased. Smart applications, like GLARE[2] and PGVL Hub[3] have made hit to lead efforts an easier and more systematic endeavor for design chemists by organizing searchable information such as reaction type, synthesis instructions, and building block availability all in one desktop package. The strategy behind PGVL was to invest in four pillars of information. The first pillar was the systematic investment to develop experimentally validated parallel enabled synthetic protocols. Although parallel enabled protocol development, at the level of detail described herein, required a major investment of resources, the end result was envisioned to have a good return on investment over the long run. However, to fully leverage that investment, a tool was needed to capture and organize the information and to store it in a searchable database. This need led to the second pillar of information which involved the systematic capture of chemistry knowledge in machine-readable format for automated reactant mining and product formation using fast cheminformatics technology. The third pillar was to deliver an enterprise wide library design desktop tool. And, the fourth pillar was the commitment of substantial resources to enable enrichment of the screening collection with over a million compounds of parallel synthesis origin, using PGVL as the informatics hub. During the past decade, Pfizer engaged in a major initiative to increase its corporate screening compound collection through both internal production and external collaborations (mainly with ArQule, ChemBridge, ChemRx/DPI, and Tripos).[4] This initiative led to the addition of two million compounds with a parallel synthesis origin and spawned over 1000 optimized parallel synthesis protocols. This initiative represented an unprecedented effort even among large pharmaceutical companies. Although some technical details of PGVL have been discussed in detail elsewhere,[3] this paper is designed to give a more general perspective for discovery chemists, including a retropsective analysis of several drug discovery project application examples.

BACKGROUND

The predecessor of PGVL was a system called LiBrain based on MDL technology,[5] which was developed at a biotech company called Alanex between 1996 and 1999. In 2000, Tripos scientists reported the utility of a virtual compound library constructed from commercially available reactants and seven parallel synthesis reactions published in the literature.[6] They demonstrated that this virtual space contained drug-like

Received: August 26, 2012
Revised: September 29, 2012
Published: September 30, 2012

   dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579–589

molecules searchable by their Topomer shape similarity search. However, the reactants used in their study were not associated with any explicit synthetic protocol. Our experience over the years has proven that association of reactants with reactant scope-and-limitation filters, derived from a written synthetic protocol and embedded in a computer algorithm, is a key component to defining a synthetically feasible virtual library. In 2005, Nikitin and co-workers also constructed a large and diverse space of virtual compounds with potential applications in drug discovery.[7] This collection was built based on reaction schemes from approximately two hundred literature papers and four hundred individual combinatorial libraries. The building blocks were either from the same literature papers or chosen from commercially available reactants using drug-like and reaction suitability filters. The Nikitin work represented a significantly broader coverage of diversity in chemistry and the virtual compounds were more likely to be synthetically accessible since they were based on experimental precedence. One of the best examples of the association of synthetic feasible virtual space with tight integration of synthesis protocols and design is found in the AIDD system published by Manly.[8] Because of the limited number of synthetic protocols captured in AIDD, the virtual compound space of 150 million compounds was, understandably, smaller than those from Andrews et al[6] and Nikitin at al.[7] Lessel and co-workers reported on similar software called BI-CLAIM[9] in 2009, which used 300 000 reactants leading to about 5×10^{11} products. Even though the number of explicit synthetic protocols was not disclosed in the BI-CLAIM work, the products were claimed to have an association with a parallel synthesis protocol. In addition, BI-CLAIM was considered a dynamic and growing system, similar to what is described here for PGVL. Indeed, the practice of building and maintaining synthetically feasible virtual product space via systematic knowledge capture is quite common in the drug discovery industry, albeit not routinely published on. In this paper, we discuss the construction of a virtual product space using synthetically feasible reactant combinations based on scope and limitations information that was experimentally derived. The reader is walked through (a) the general design of the system, (b) detailed illustrations on reaction and reactant information recorded, (c) product enumeration instructions, and (d) a few use cases with a description of their impact at Pfizer.

CONCEPTS AND TERMS DEFINED

We start with carefully defining the core terminology used in this report. Similar terms might have been used slightly differently in literature while other terms may not be familiar at all. Figure 1 shows a reaction involving the cyclocondensation of aminoheterocycles (component A) with alpha-halo ketones (component B).

The reaction starting materials which are used to construct part of the product are called reactants (often called monomers internally at Pfizer). Reactants typically belong to one class of

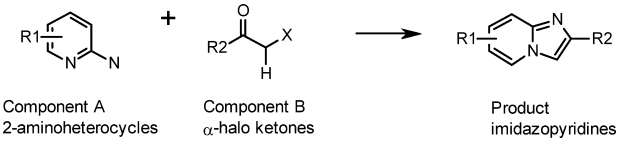


Figure 1. Imidazopyridine ring formation using amino-heterocycles and alpha-halo ketones.

organic functional groups and are described using the word component with a letter designation. For example, a two component reaction would have a design consisting of one or more component A reactants and one or more component B reactants. Another key term is the reaction scheme, which is defined as a reaction drawing that unambiguously defines the regio- and stereochemical outcome of a given synthetic transformation, in general terms, using R groups to show optional substitution. A virtual library is a collection of possible products and is usually defined by a set of synthetic protocols and reactants. So, while PGVL represents the entire Pfizer virtual library, it is more common to discuss a virtual library based on a specific reaction scheme and a reactant set which is compatible with the experimental synthesis protocol intended to be used. The term scope and limitations is used to define reactants compatibilities with a given synthetic protocol. Most parallel synthesis protocols are developed to cover a wide range of reactant features but there are often steric or electronic features that prohibit successful use of specific reactants. In the PGVL data system, the hierarchy of information starts with a term called a virtual reaction (abbreviated VRXN). This virtual reaction is defined by a reaction scheme and synthesis protocols are grouped under the VRXN umbrella into a family under the same VRXN ID. Figure 2 provides a schematic explanation of this hierarchy.

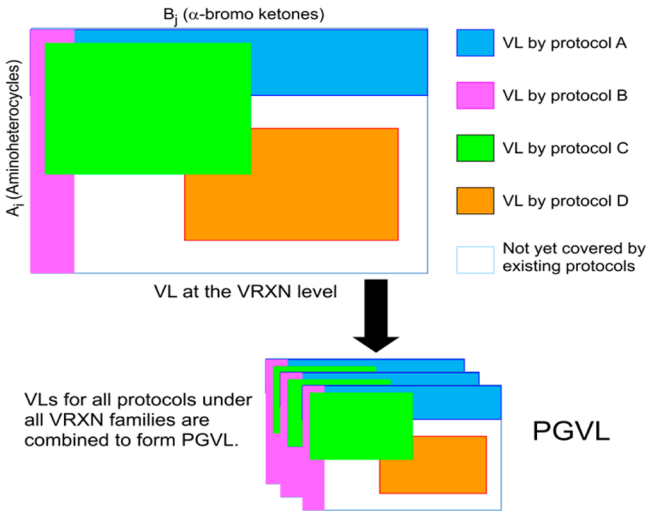


Figure 2. Illustration of the scope and limitation for a given synthetic protocol, its VRXN family, and the whole PGVL.

One can appreciate the fact that one reactant may be suitable for one synthetic protocol but not another, even though the two protocols may share the same reaction scheme (VRXN). Virtual compounds in the white bands are not accessible through existing synthetic protocols and are not considered part of PGVL. Nevertheless, a medicinal chemist may still design in that space, provided that he/she is willing to test and modify reaction conditions to ensure a reasonable success rate for that region of chemical space. Alternatively, singleton synthesis using extra care and a skilled chemist at the bench may enable access to this region.

PART 1: REACTANT FILTERING

As described in the background section, PGVL is organized around enabled or accessible chemistry space. There are a number of reasons to define virtual compound collections in

dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579–589

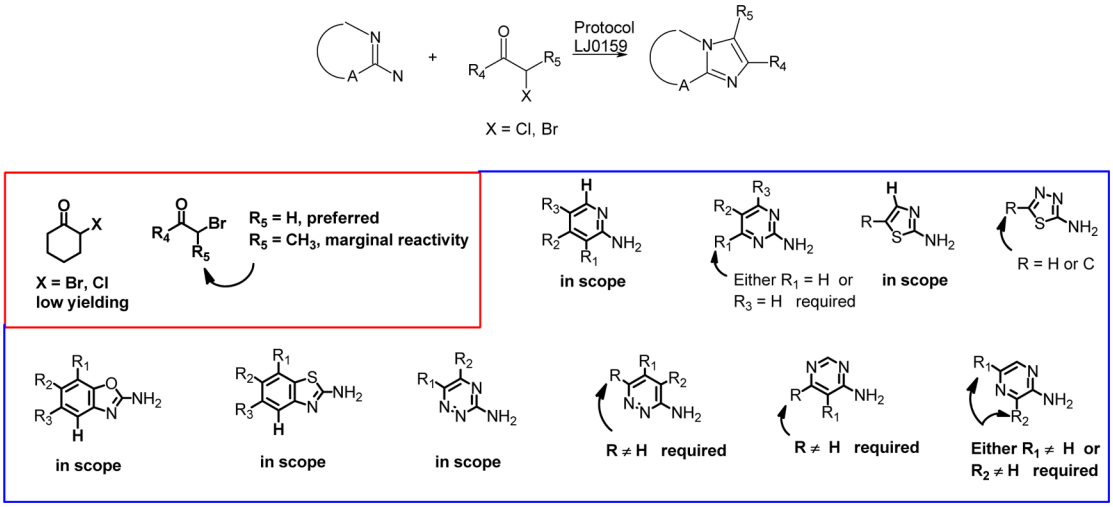


Figure 3. Scope and limitations information as described in protocol LJ0159. Suitable reactants for imidazopyridine-forming reaction by amino-heterocycles and alpha-bromo ketone. Various cores and explicit hydrogens are used to define the score and limitation of the reactants for LJ0159.

this way. One very practical reason is that the organization of information into synthetically accessible chemistry space accelerates follow-up of hits during the hit to lead phase of a program. The rapid retrieval of synthesis information along with scope and limitations can guide successful design and synthesis without spending time and resources on synthetic enablement. In fact, the synthetic enablement part was done up-front and resulted in a protocol document explicitly containing scope and limitations information which gets translated into a machine readable format. For example, the scope and limitations information from synthetic protocol LJ0159 is shown in Figure 3 as it appears in the protocol document. Translation of this into a machine readable format involves using explicit hydrogen atoms in automatic sub-structure searches. The reactants passing these filters have the best chance of synthetic success and are retrieved for the designer to select. Use of this information in design, although not strictly enforced, translates into improved synthetic yields during synthesis and purification. This information is especially useful for designers who may not have an extensive knowledge of organic synthesis and reactivity. Translation of this knowledge into a machine readable format is a topic worthy of further elaboration.

Translating the chemistry scope and limitations information from the synthetic protocol into a machine readable language presents an interesting challenge. Most cheminformatics systems (including PGVL) capture chemistry rules that are encoded in some chemistry-savvy query format. The most popular encoding formats are (a) MDL CTAB[10] and (b) daylight SMILES/SMARTS.[11] Each encoding format has its own set of strengths and weaknesses. In PGVL, the scope and limitations and product enumerations are all encoded as MDL CTAB strings for exact and substructure queries with several useful SciTegic Pipeline Pilot[12] extensions.[13] This choice was made because of the chemists' familiarity with the ISIS/Draw and ISIS/Base drawing rules. In addition, the SciTegic substructure query extensions were effectively used for smart queries, such as those which specify atomic and bond hybridization status (aliphatic, aromatic, sp1, sp2, and sp3). Those extensions were easily constructed using ISIS/Draw. The result was a coordinated design with a flexible and powerful backend engine to drive batch reactant mining and product structure formation. PGVL was a vast improvement to our early LiBrain system, which was built on top of an old MDL backend for query generation and reactant mining.

A simple amine alkylation reaction serves as an initial example to illustrate the need to understand the scope and limitations of each reaction component as well as for the reaction itself, and to translate that into a format, which can be stored within a database and retrieved for later use. The amine component could be specified as primary amines and secondary amines with the exclusion of anilines. Further filters could be set up to exclude amines that have a quaternary carbon center adjacent to the NH. The halide component could be limited to benzylic halides and activated alkyl halides as specified by having a sp^2 or sp^3 carbon alpha to the reactive center. Finally, any competing electrophilic or nucleophilic groups could be excluded from all the reactants to form a clean set of reactants to start the library design. While component exclusions remove component-specific undesirable functional groups, PGVL also removes functional groups that are undesirable for all components. These are called general interference functional groups. These general interference queries typically remove functional groups that are strong electrophiles such as epoxides and acid chlorides (unless they are the actual reaction components), as well as strong nucleophiles, such as primary and unhindered secondary amines. For the registration of a new reaction, a common set of interferences is used as a starting point which is then modified specifically for the given synthetic protocol. Together, the general interference rules and the reactant filters, codify the logic steps used for reactant mining.

The task of mining for desirable reactants from a corporate or vendor reactant database is carried out in the following way. For each reactant, a set of queries tests to see if that reactant is compatible with a given synthesis protocol. The reactant is kept if it satisfies all the criteria imposed and rejected if it fails one of the queries in the set. The obvious goal for reactant mining is to include all suitable reactants and reject those that are not suitable. Since corporate and vendor reactant databases are quite large, and contain a large number of nonsuitable reactants, automated exclusion of unreactive or incompatible reactants is highly desirable. This frees up the design chemist to focus on a cleaner initial list of reactants and concentrate on library design with less emphasis on reactants incompatibility. As shown in

dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579−589

Figure 4a and 4b, the protocol-level reactant mining instructions based on scope and limitations information are

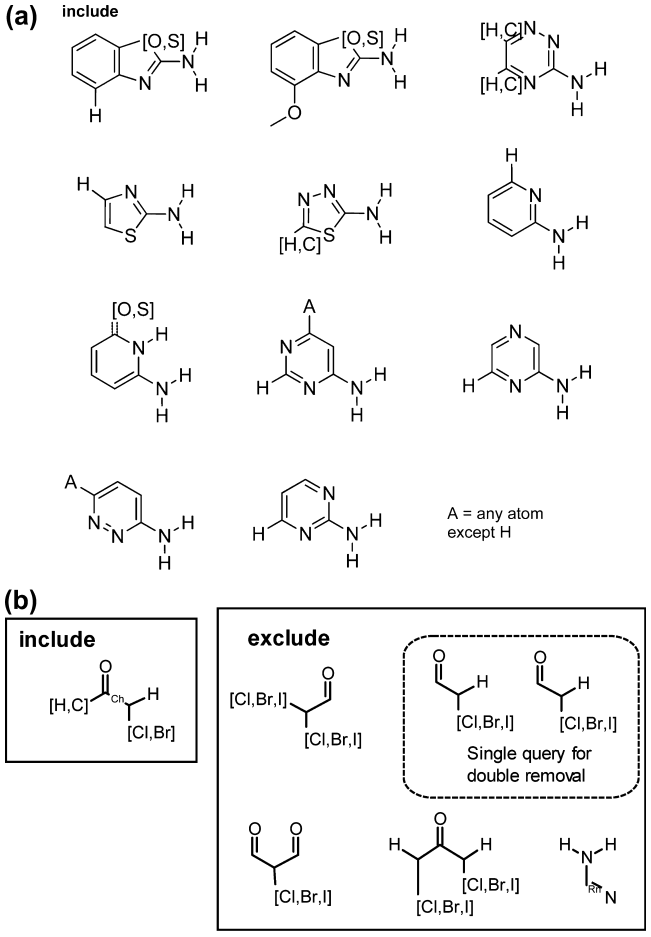


Figure 4. (a) 2-Aminoheterocycles to include as suitable reactants. (b) α-haloketones to include and exclude.

captured within a set of queries called the Automated Reactant Mining Object (ARM Object). The use of MDL CTAB strings, as described earlier, make up the substructure or exact queries.

The most fundamental data element inside the ARM Object is a substructure or exact query expressed as a MDL CTAB string created using ISIS/Draw.[14] The detailed data structure of ARM Object is shown in Supporting Information Figure 1. The ARM Object uses both inclusion and exclusion criteria in the form of substructure or exact queries to retain or reject a reactant. A reactant is first checked to see if it contains any functional group that interferes with the reaction. If it survives this test, it will then be tested to see if it contains any structural feature that would exclude it from being an acceptable reaction component. If it passes this second test, it will be mapped by the query structures in the variation-level inclusion and exclusion lists to ensure that it contains the valid reactive functional groups for the reaction. If the reactant passes all the tests, it is retained as a suitable reactant for the design.

At the protocol level, Figure 3 illustrates the scope and limitations for the reaction of imidazopyridine formation using 2-aminoheterocycles and α-haloketones. From this information, a chemist must input the chemical rules that capture and enforce a filtering algorithm that reflects Figure 3 but is translated into the queries of **4a** and **4b**, which are machine readable. For example, 2-aminoheterocycles having substitution at the position specified to require an explicit hydrogen atom shown in Figure 3 would be rejected by placing an explicit hydrogen in the substructure for the inclusion query. The chemists responsible for writing these queries are called reaction registrars. These registrars use a registration tool to construct these query filters with the help of ISIS/Draw. Once these filters are set up, reactant mining occurs automatically and is updated as new reactants are added to the corporate database. A screen shot of the registration tool is provided in Supporting Information Figure 2. There are, of course, situations when the designer may want to include reactants that do not pass these filters. By design, PGVL is flexible and can easily include user-selected reactants as directed by the design chemist. In fact, the designer can easily over-ride many of the automatic reactant mining features described in this paper if he or she desires. This over-ride feature also incorporates the ability to design using reactions for which there are no explicit parallel synthesis protocols.

PART 2: FILTERING REACTANTS AND ENUMERATION OF PRODUCTS

Enumerating large virtual libraries requires the use of automated procedures. The automatic generation of product structures requires that the rules of transforming reactant structures into products be precisely conveyed by the enumeration instructions. These instructions often adopt a format of a reaction scheme where reactive groups in a reactant are depicted using explicit functional group symbols whereas the variable elements of the reactants are symbolized by R_1, R_2, ..., R_n notations. Although quite natural for trained chemists, this method of conveying chemistry logic can be ambiguous to a computer, especially given the complexity of chemistry involving the rearrangement of bonds. To facilitate the knowledge transformation from chemistry logic to computer program, we created a special data structure called, VRXN Object, which encoded the same two-step, clip-and-assemble approach popular in many cheminformatics packages for PGVL product formation[15] (Figure 5).

In addition to a product core structure with numbered R-group attachments, the VRXN Object contains similar information as that inside an ARM Object to identify the appropriate reactive functional groups before the clipping and labeling with numbered *z*-labels occurs (detailed data structure is shown in Supporting Information Figure 3). Each suitable reactant is expected to contain the appropriate reactive functional group (reactive site or reactive element). Since a linker reactant in a three-component reaction has two reactive functional groups, the VRXN Object also contains one more layer of hierarchy to deal with this complexity. Therefore one reaction component can have one or more "reactive elements". For each reactive element, the variations (such as primary amine, secondary amine, and aniline etc.) are then ordered based on expected reactivity, parallel to that in the corresponding ARM object.

Most useful reactions allow more than one particular functional group as the reactive group. The variation in functional groups may be trivial. For example, in addition to chloride as a leaving group, one can use other halides and sulfonates as leaving groups for S_N2 chemistry. For a single atom variation one can use an atom list instead of an explicit atom symbol in the reaction scheme, such as R_1-CH_2-[Cl,Br,I]. For more complicated variations, it is necessary to

dx.doi.org/10.1021/co300096q | *ACS Comb. Sci.* 2012, 14, 579−589

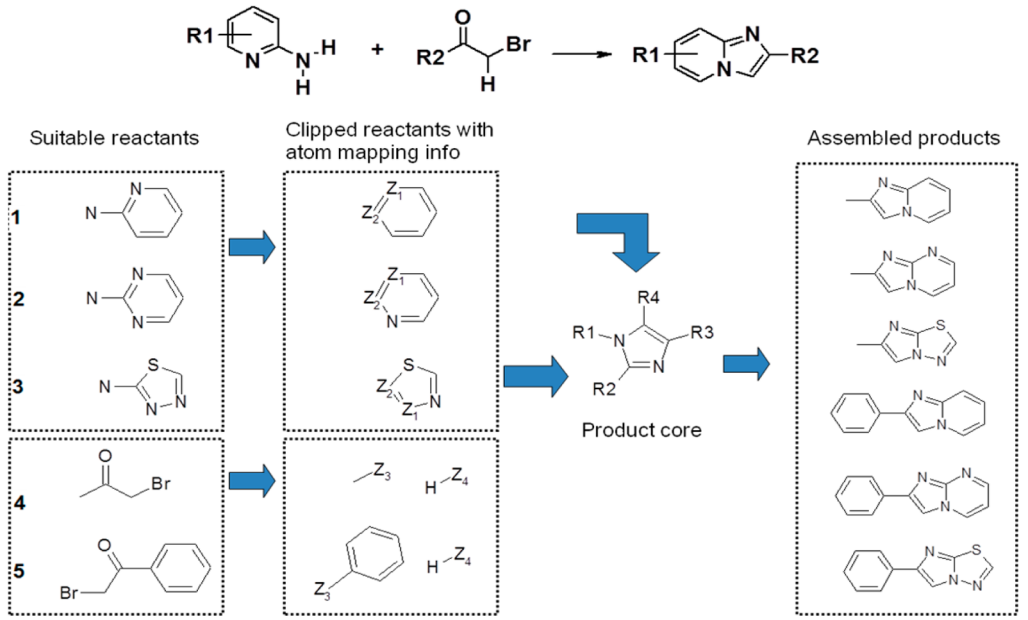


Figure 5. Product enumeration using reactant clipping and product assembly. The imidazopyridine-forming reaction using aminoheterocycles and alpha-bromo ketones is shown as an example. The first step involves clipping of reactant molecules at their respective reactive centers and label the clipping site with numbered *z*-labels (z_1, z_2, etc.). The second step involves assembly of products based on a product core structure with numbered R-group attachments (R_1, R_2, etc.) and forming new bonds at the appropriate z-labeled sites.

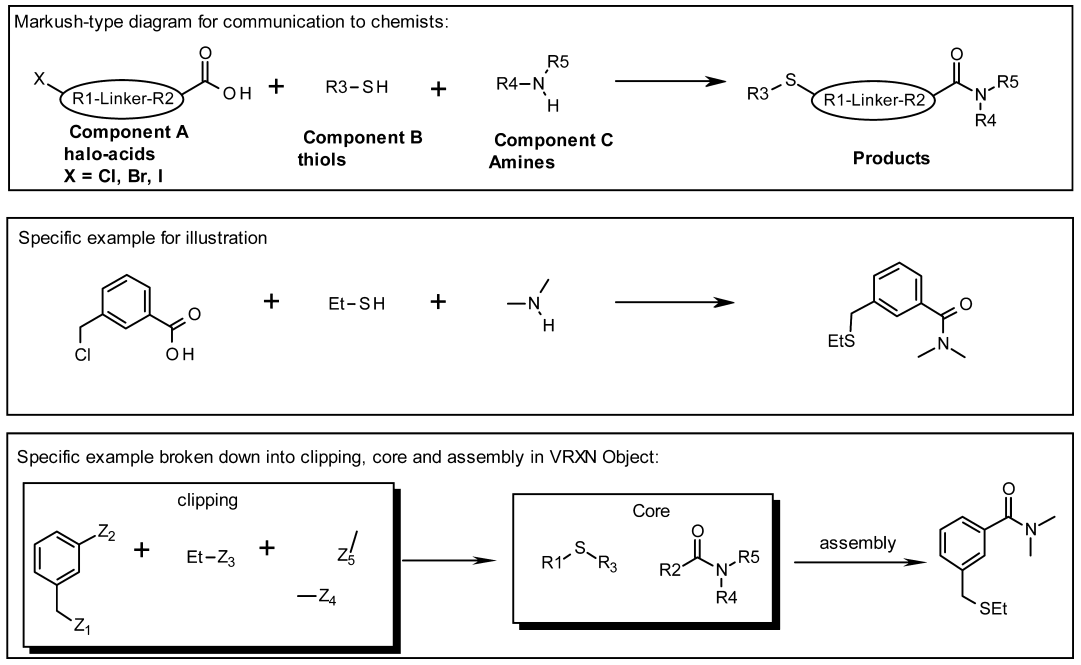


Figure 6. Progression of information from protocol to reaction example to machine readable format.

draw the reactive functional group as an additional structure while providing the corresponding clipping instructions. The system is set up to easily handle situations where multiple reactive functional groups are tolerated in a single reaction. An example of this is the Suzuki reaction where boronic acids and boronic esters are reacted with Ar–X (X = Cl, Br, I, OTf). To accommodate the most general case, PGVL views the reaction as a collection of independent clipping transformations associated with a single core. Once the reactants are clipped at the reactive functional groups, they are assembled into final products using the product core structure and following the mapping rule ($Z_1 \rightarrow R_1$, $Z_2 \rightarrow R_2$, and so on).

The terms component A, component B, and so on are important because the order of the components is closely tied to the sequence of the enumeration instructions. This sequence of events is especially important when building up enumeration instructions for reactions containing bifunctional reactants. For example, Figure 6 depicts the reaction of a set of bifunctional halo acids (component A) with thiols (component B), followed by amide bond formation with amines (component C). The order of events happening in the lab and in the computer match

dx.doi.org/10.1021/co300096q | *ACS Comb. Sci.* 2012, 14, 579–589

up and to illustrate this further a specific example is used. The first reaction to occur is the S_N2 displacement of the halide of the alpha-halo acid (component A) with the sulfur of the thiol (component B). The corresponding clipping instructions are shown below where the halide reactivity element is clipped to Z_1 and the SH reactivity element is clipped to Z_3 as instructed by the reaction core instructions. Next, an amide bond formation occurs. Again, as directed by the reaction core, the COOH reactivity element is clipped to Z_2 and the amine reactivity element is clipped to Z_4, Z_5 (sp3 nitrogen has 3 points of substitution). These components are then assembled according to the reaction core in which R_1 and R_3 are connected via a sulfur atom while R_2, R_4 and R_5 are connected via an amide linkage. Finally, because Z1 and Z2 are originally connected in the reactant of component A, all the clipped Rs are brought together in the final product automatically at the assembly step. Occasionally, PGVL will prompt either the reaction registrar or the user when more than one reactivity element is present in a reactant. In these cases, it requires a chemist, not a computer, to decide which site will react preferentially.

In most cases, these ambiguities can be resolved by more extensive sequence of structural queries, which will correctly pick the desired site for clipping.

A final consideration for product enumeration deals with the use of protecting groups present in the reactants that are removed as a last step transformation. Protecting group removal at the end of a synthetic sequence is a common event that the PGVL system handles through a feature termed "last step transformation". A menu of common protecting groups with a radio button is used to turn on this feature. When turned on, the final enumeration step removes all of the indicated protecting groups from all penultimate products by using the special PLP extension.[13] An example of the menu and the clipping scheme for removal of the Boc group is shown in Supporting Information Figure 4. This feature is particularly useful for peptide synthesis when multiple protecting group types are removed at the end of a synthetic sequence.

PART 3: MAINTAINING AND GROWING PGVL IN A DYNAMIC ENVIRONMENT

Parallel synthesis protocols are continuously developed and validated experimentally by Pfizer and CRO (Contract Research Organization) chemists. This information is captured by a small team of reaction registrars in a searchable data format with the help of an internally developed tool called the VRXN reaction editor (A screenshot of the user interface is shown in Supporting Information Figure 5). Reaction registrars are typically experienced organic chemists willing to spend part of their time encoding parallel synthesis information into PGVL. Once reaction filters are set up, they are not static and can be modified as new chemistry is developed or as information about existing chemistry changes. Because the corporate and commercial collections of reactants are also not static. The acceptable reactants for a given transformation grows as periodic reactant mining of all protocols and reactions is updated to include new reactants added to the corporate collection plus commercially available reactant databases, such as Available Chemicals Directory (ACD) collection.[16] The system is also capable of maintaining unpublished protocols which are visible only to reaction registrars and administrators. This allows for capture of parallel synthesis information where validation was attempted but the chemistry was unsuccessful or still being worked out. Once validation of the chemistry is complete, the reaction registrar publishes the new reaction so that it will appear to the general PGVL user community. This dynamic nature of PGVL is a factor that differentiates itself from other published virtual compound space of static or quasistatic nature.

PART 4. SCALE AND SCOPE OF PGVL

The virtual library sizes within PGVL are summarized in Table 1. A total of 1244 unique transformations, called VRXNs, are captured. The enumeration of all possible chemistry space when using the Pfizer available inventory encompasses 10^{14} virtual compounds. When using the commercial (ACD) inventory, the possible chemistry space encompasses 10^{18} virtual compounds.

Table 1. Snapshot of PGVL's Scale (as of January 2011)

starting material	no. VRXNs	no. of basis products	no. of virtual products
reagents from in-house inventory			
2-component RXN	436	2 453 094	4.034×10^9
3-component RXN	725	5 907 543	1.102×10^{13}
4-component RXN	83	786 037	2.993×10^{14}
total	1244	9 146 674	3.103×10^{14}
reagents from ACD of Accelrys			
2-component RXN	436	23 308 473	4.905×10^{11}
3-component RXN	725	46 975 889	9.216×10^{15}
4-component RXN	83	5 504 761	1.269×10^{18}
total	1244	75 789 123	1.278×10^{18}

Two reactant databases were used to populate the numbers in Table 1. The Pfizer in-house inventory of reactants represented one database, while the other database was the Accelrys commercially available listing known as the Available Chemicals Directory or ACD.[16] Not surprisingly, 3-component reactions make up over half of the unique transformations (VRXNs) because this class of reactions was a mainstay during the period of file enrichment at Pfizer. This is because 3-component reactions typically lead to optimal balance of lower MW and greater diversity. Because of the combinatorial nature of the calculations done in Table 1, 4-component reactions dominate the size of the virtual libraries. Although 4-component libraries typically led to products of higher average MW when produced in a combinatorial array, the practice of cherry-picking from a combinatorial array increases the utility of these 4-component transformations. In fact, libraries at Pfizer are typically not produced in combinatorial arrays any longer. Cherry-picked libraries are frequently designed in which compounds outside of some specified calculated space are excluded from the design and execution. PGVL in combination with robotic liquid handlers make cherry-picked libraries a routine practice. The third column in Table 1 is the breakdown numbers of virtual compounds in terms of basis products (BPs). The concept of basis products was first introduced by Shi, et al.[17] and is loosely defined as the products formed by combining all the reactants for a given reaction component with the simplest set of complementary reactant partners. It has been used as an efficient way to sample and explore the physical property space of large collections such as PGVL,[17] and to systematically apply structure-based library design across chemical space.[18] In addition, basis products have been used as one of the chemical space sampling methods to design Pfizer

dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579–589

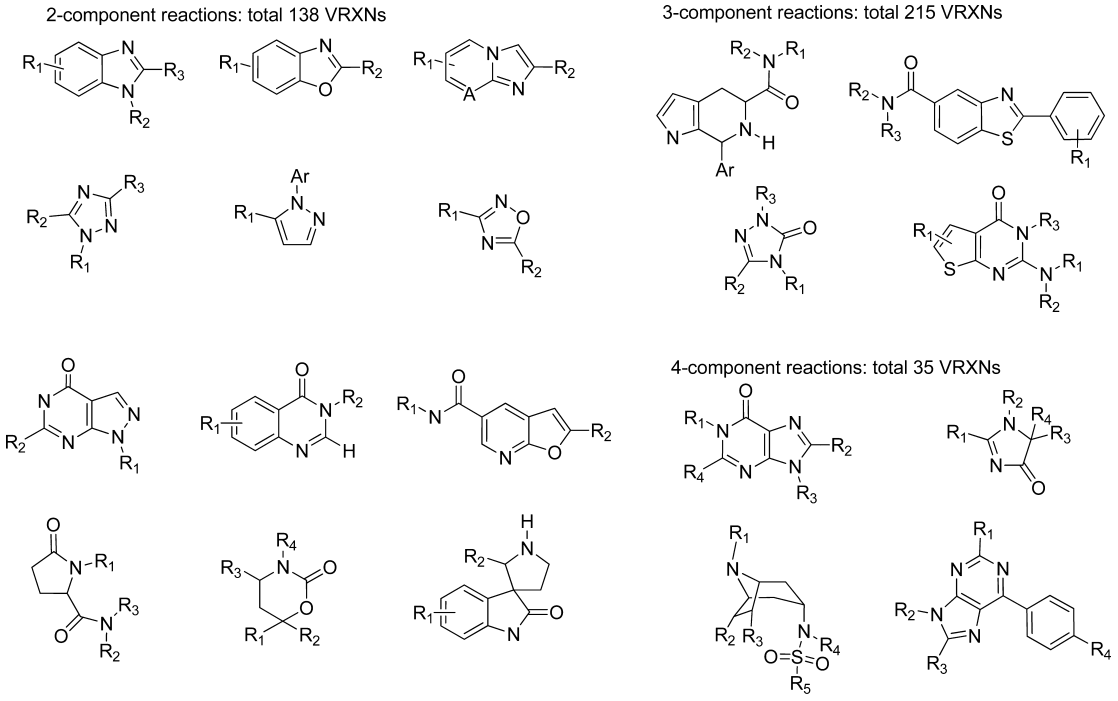


Figure 7. Summary and some examples of VRXNs that form heterocyclic rings.

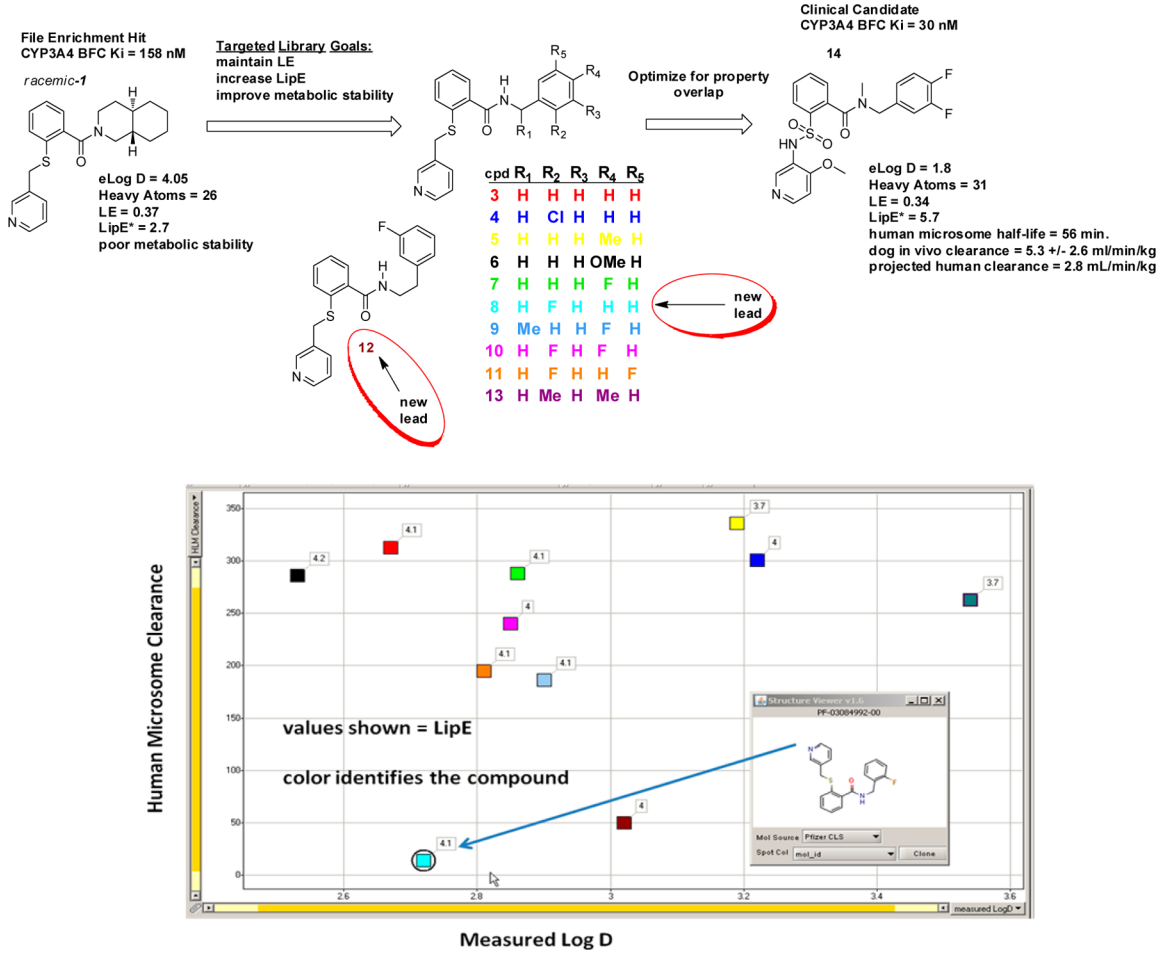


Figure 8. Discovery of clinical candidate **14** as a CYP3A4 inhibitor.

dx.doi.org/10.1021/co300096q | *ACS Comb. Sci.* 2012, 14, 579−589

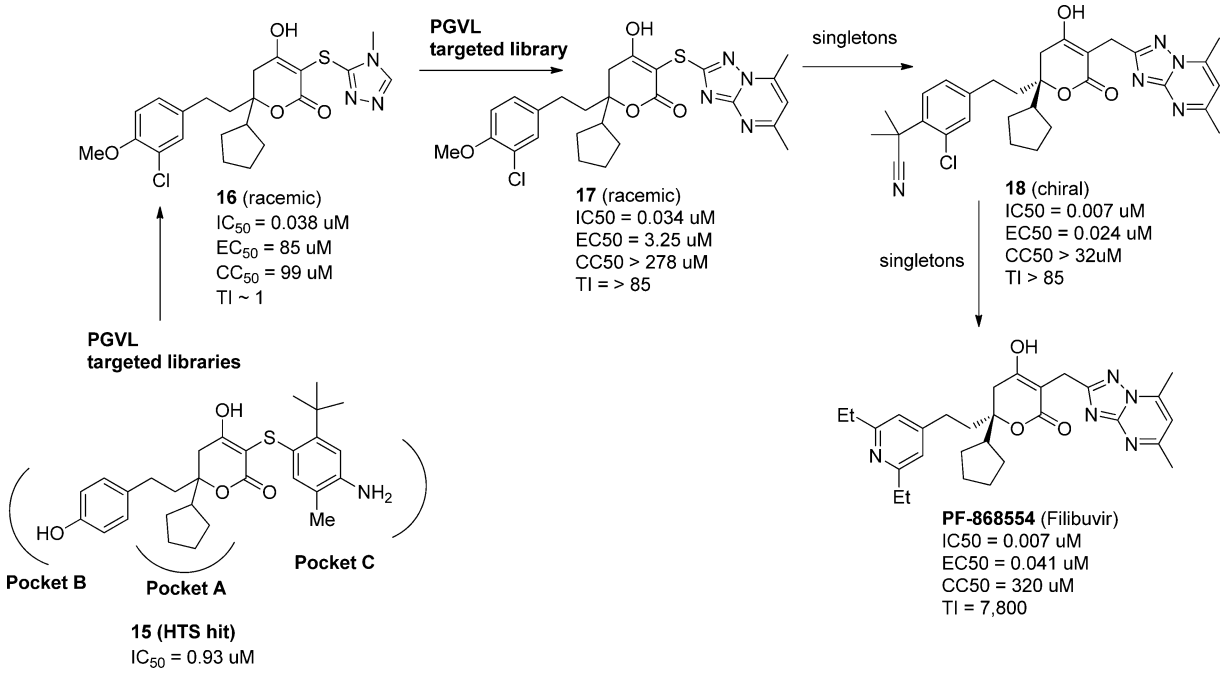


Figure 9. Discovery of Filibuvir from HTS hit **15**.

fragment screening collection[19] and to perform 2D similarity searches across the entire 10^{14} PGVL chemical space.[20]

Heterocyclic compounds are common scaffolds in drug discovery. A common practice in parallel medicinal chemistry is to decorate a heterocyclic scaffold using standard transformations such as amide bond formation, Suzuki reactions, ether formation, etc. Alternatively, the ability to form the heterocyclic ring while bringing together diversity elements on the periphery of the core is another powerful tool since it often does not require scale-up of a highly engineered heterocyclic scaffold. Samples of the 388 heterocyclic ring formation reactions within PGVL are shown in Figure 7, broken down by the number of components used. Although many of these reactions represent standard fare within the repertoire of a skilled medicinal chemist, the capture of this information into a searchable corporate database affords the luxury of not reinventing the wheel each time a reaction protocol is sought. In addition, for those less skilled, the database can provide a source for new ideas while browsing desirable substructures. With the systematic data capturing in a searchable format, design chemists can select from over 1000 reaction families (VRXNs) and over 3000 synthetic protocols, each with its own sets of filtered reactant lists, created to bias a design towards higher degree of synthetic success.

In summary, the PGVL system has been shown to be quite powerful, robust, and flexible. It has proven itself capable of handling reaction complexity from simple to complex using the hierarchical data structures shown in Supporting Information Figures 1 and 3. The one reaction class not handled well by PGVL is in the case of metalation and CH activation reactions where reactive sites on a large number of potential aromatic and heteroaromatic C−H sites are determined by pK_a of the CH carbon acid, by directing group effects and other properties.[21] The vast diversity of this reaction class is difficult to generalize to cover all cases. In those difficult situations, other approaches not solely based on substructural queries to identify the reactive sites might offer possible solutions.[22] However, PGVL does easily handle specific cases where structurally similar ArCH reactants are used. This is the situation most often encountered in library design targeted towards advancement of a specific lead series for a specific target.

PART 5. SOME EXAMPLES OF PGVL CHEMISTRY AND ITS APPLICATIONS AND IMPACT ON DRUG DISCOVERY

Examples of the application of PGVL to drug discovery have previously been cited.[23] Because PGVL has been used over many years, and not just for large library designs but also for designs and evaluations of smaller virtual compound collections, including singletons, the impact of the tool itself on drug design is difficult to quantify. The platform we established can provide both chemistry ideas and rapid hit follow-up via parallel synthesis, thereby increasing their productivity. An example of the use of PGVL at Pfizer involved the CYP3A4 project which had the goal to develop an oral PK enhancer that reversibly and selectively inhibited the activity of the CYP3A4 enzyme, a major human enzyme involved in the clearance of many drugs and drug candidates. The product was expected to improve the PK properties of coadministered antiretroviral or oncology agents, primarily metabolized by CYP3A4. Our efforts led to the discovery of compound **14**, a clinical candidate that entered first-in-human clinical trials in 2007. Figure 8 illustrates the discovery of **14** which started with compound **1**. Interestingly, **1**, itself came from Pfizer's file enrichment initiative.

Compound **1** can be characterized as a metabolically unstable, lipophilic hit (experimental log D = 4.05) with moderate potency (IC_{50} = 158 nM measured in the presence of 7-benzyloxy-trifluoromethylcoumarin [BFC] as the CYP3A4 substrate). Goals were immediately set to address the deficiencies of **1**. These goals included: (1) enhancing lipophilic efficiency (LipE),[24] (2) reducing clearance, and (3) gaining selectivity for CYP3A4 over other CYP isoforms. Figure 8

dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579−589

provides a summary of the effort to optimize **1** that involved the use of structure-based drug design (SBDD), traditional and parallel medicinal chemistry and synchronized multiparameter optimization. The impact of parallel medicinal chemistry is emphasized in Figure 8 to align with the topics of this paper. One strategy employed right away involved synthesis of several amide side-chain libraries to improve LipE and metabolic stability. From the initial amide libraries, we gained rapid SAR information as depicted in the SpotFire plot of Figure 8 and we also discovered two new leads (**8** and **12**) with improved LipE (4.1 and 4.0) and greatly enhanced metabolic stability (HLM clearance of 50 and 14 mL/min/kg respectively). When compared to the other benzylic amides shown in the SpotFire plot of Figure 8, the enhanced metabolic stability of **8** is striking and likely due to the ortho-fluoro group blocking a site of metabolism. Compared to the parent compound, **3**, (HLM Cl > 300 mL/min/kg), compound **8** is a stable outlier. The discovery of ortho-fluoro-substituted benzylamide (**8**) and meta-fluoro phenethylamide (**12**) as metabolically stable outliers was information we gained very early in the program through the use of parallel medicinal chemistry. Building upon this knowledge and keeping in mind further LipE enhancement, led to the replacement of the thioether linker by the sulfonamide to reduce lipophilicity. Finally, the unsubstituted pyridine was replaced by the 4-methoxypyridine to slow down metabolic N-oxidation, further improving clearance. Ultimately, **14** was identified for progression into further studies leading to its nomination as a clinical candidate.[25] Prior to publication of the patent application,[26] there were only four substances in the Chemical Abstracts database with the 3-pyridyl benzamide substructure of **1**. This substructure is one that the casual observer would not immediately recognize as novel because of its structural simplicity. This surprising finding shows that, using a simple two-step procedure of amide bond formation and thiol alkylation, novel and useful structures such as **1** can serve as hits for lead optimization. In addition to serving the file enrichment cause, PGVL also accelerated our lead finding efforts by organizing synthetic information, calculated physicochemical properties, and enabled design attributes which facilitated the rapid design and synthesis of the targeted libraries discussed above.

Another example of a successful drug discovery program impacted through the skilled use of PGVL is the discovery of the HCV-polymerase inhibitor, PF-868554 (Filibuvir) shown in Figure 9.[27] The discovery of Filibuvir began with HTS hit **15**.[27c] A targeted library designed to explore the SAR of the C-pocket was launched using heterocyclic thiols to examine the interaction of various heterocycles with protein residues in the C-pocket. This led to the discovery of 1,3,4-triazole **16** having improved properties relative to the aniline hit. Although **16** showed 38 nM potency in the HCV polymerase enzymatic assay, the antiviral potency was weak with a therapeutic index (TI) equal to 1. A second round of C-pocket SAR exploration via another targeted library identified the 5,7-dimethyl-[1,2,4]-triazolo[1,5-a]pyrimidine fragment of **17**.

This unique heterocycle led to a 26-fold improvement in antiviral potency and an 85-fold improvement in TI. Efforts to find smaller S-linked and C-linked heterocycles with equivalent antiviral potency was unsuccessful and this unique triazolo[1,5-a]pyrimidine became a clinical candidate. These efforts were, of course, facilitated by the use of PGVL as the project team generated virtual compounds and analyzed their calculated properties prior to synthesis. Included in this development effort was enablement of parallel C-linked chemistry using a dimethylamine-borane reduction to form the carbon−carbon bond of the C-linked pyrone using aldehydes (protocols LJ0430 and LJ0476). The new protocols were captured by the PGVL system ensuring the searchable retrieval of synthesis information for present and future discovery efforts at Pfizer.

CONCLUSIONS

During the past decade, the file enrichment initiative at Pfizer has generated more than one million new compounds for the screening collection. Over 3000 parallel synthesis protocols were categorized into more than 1000 virtual reactions (**VRXN**s). To enable the full potential of this corporate asset, an enterprise-level library design and cheminformatics system was created and named PGVL (**Pfizer Global Virtual Library**). This system has a front end GUI interface called PGVL Hub[3] and a back end chemistry data foundation system described in this paper. The PGVL system is maintained by reaction registrars who use their synthetic chemistry knowledge and expert chemical query construction skills to capture synthesis information into a searchable database. Key attributes of the system include chemistry-savvy automated reactant mining objects (**ARM Object**) for producing tailor-made reactant lists and virtual reaction objects (**VRXN Object**) for precise product enumeration, which forms the basis of a machine-based chemistry knowledge foundation that can be stored and reused. PGVL also forms the basis for speedy Hit follow-up and Lead Hopping via a set of Lead-Centric Mining (LCM) type of approaches (see a summary in the Table 13.6 of ref 20 for more details).

While recognizing the complexity of virtual chemistry space, our efforts with PGVL are aimed at organizing this complexity into a **synthesis knowledge system**, supplied with the chemistry engines for library design and lead-centric mining with desktop access[3] for efficient use and reuse. Although the full PGVL compound set is too large to be fully enumerated and stored as explicit molecules, any subset of PGVL can be explicitly formed and delivered on-the-fly to project chemists.

Even to the people with inside knowledge of PGVL such as the reaction registrars, the chemistry knowledge accumulated over the years is fairly enormous and complex. The complexity is also driven by the design goal of PGVL to fully capture chemistry knowledge and enable automated reactant mining and product formation without any additional user input. In a sense, the valuable investment made by a handful of highly experienced reaction registrars are fully leveraged by end users who tap into the PGVL chemistry knowledge and services for library design, product formation, and virtual compound searching and screening. Now scientists no longer have to remember more than 1000 reactions and 3000 protocols available in-house and go through ACD or other similar compound databases to select reactant for their library design and idea generation.

A potential drawback in chemistry knowledge capture is often the incompleteness of information. For template based synthetic protocols, often the chemist will only validate a single template, so the registrar has to infer that other similar templates will also work under the given reaction conditions. Or, the chemist may indicate that a reactant fails in reaction development, but that reactant may closely resemble other successful reactants, so the registrar may be unable to separate reactant failure due to chemistry or a bad reactant bottle. Often the protocol development chemist is consulted or "chemist

dx.doi.org/10.1021/co300096q | ACS Comb. Sci. 2012, 14, 579−589

intuition" is used by the reaction registrars to resolve ambiguities. One other practice involves updating the ARM object after a library is made. This is done only if the reaction registrar analyzes the library results and recognizes more chemistry knowledge can be extracted. This iterative refinement of the ARM object ensures it will lead to high quality ARM lists and hence higher synthetic success when used for library design.

One measure of usefulness lies in tracking the usage of PGVL within Pfizer. At its peak, PGVL had more than 1000 unique users and a steady 60−100 launches per day. Its functionality now resides within a next-generation library design tool with new features and the added benefit of tighter integration with other Pfizer desktop design tools.

■ ASSOCIATED CONTENT

Supporting Information

Figures showing (1) the data structure of automated reactant mining object (ARM Object), (2) ARM object editor, (3) the data structure of virtual reaction object (VRXN Object), (4) last step transformation, and (5)VRXN object editor. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: huqy@shhrp.com, qiyue.hu@gmail.com.

Present Address

[†]Shanghai Hengrui Pharmaceutical Co. Ltd. 279 Wenjing Road Shanghai 200245 People's Republic of China

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Dr. Ben Burke for helping with the final editing of the manuscript. We would also like to thank the following Pfizer colleagues for their dedicated effort and generous support provided to us during the construction and deployment of the PGVL knowledge system: Da Tse and Zi (Lily)Yang for their long-term support in computer hardware and software (Oracle, ISIS db, etc); Dr. Bo Chao for the accessing of the corporate compound databases and the related compound structure service that enabled global deployment of PGVL; Dr. Zhengdong Zhu and Dr. Michael McAllister for sharing their valuable knowledge and experience in maintaining the LiBrain chemical knowledge and data system, the predecessor of PGVL; Dr. Alex Polinsky and Dr. Larry Truesdale for supporting the continued development of LiBrain and PGVL system, and maintaining the close ties between the parallel synthesis production group and the LiBrain and PGVL systems; the chemists whose creativity really drives projects forward with the help of tools like PGVL Hub, Angelica Linton, Dr. Hui Li, John Tatlock, Dr. John Kath, and Dr. Ben Burke; and finally the dedicated members of the global VL reaction registrars, Shaughn Robinson, Christina Vermillion, Dr. Sridharan Sundaram, Caroline Landre-Smith, Dr. Seiji Nukui, Dr. Catherine Johnson, Dr. Ken Wu, Dr. John Clark, Dr. Ian Johns, Dr. Nunzio Sciammetta, Sarah Anstead, Dr. Derek Sheehan, Dr. Andrew Cronin, Dr. Mark Cox, Dr. Dave Powers, Dr. Michael Tollefson, Dr. Rachel Osborne, and Andrew Samant for their reaction registration efforts.

■ REFERENCES

(1) For historical reviews, see: (a) Hogan, J. C., Jr. Combinatorial Chemistry in Drug Discovery. *Nat. Biotechnol.* **1997**, *15*, 328−330. (b) Hall, S. E. The Future of Combinatorial Chemistry as Drug Discovery Paradigm. *Pharm. Res.* **1997**, 1104−1105. (c) Salemme, F. R.; Spurlino, J.; Bone, R. Serendipity Meets Precision: the Integration of Structure-based Drug Design and Combinatorial Chemistry for Efficient Drug Discovery. *Structure* **1997**, *5*, 319−324. (d) Floyd, C. D.; Leblanc, C.; Whittaker, M. Combinatorial Chemistry as a Tool for Drug Discovery. *Prog. Med. Chem.* **1999**, *36*, 91−163. (e) Beeley, N.; Berger, A. A Revolution in Drug Discovery. Combinatorial Chemistry Still Needs Logic to Drive Science Forward. *Br. Med. J.* **2000**, 581−582. For two more recent reviews, see: (f) Hughes, I. Combinatorial Chemistry in the Drug Discovery Process. In *Drug Discovery and Development, Vol. 1: Drug Discovery*; Chorghade, M. S., Ed.; John Wiley & Sons: 2006; pp 129−167. (g) Kennedy, J. P.; Williams, L.; Bridges, D. N.; Weaver, D.; Lindsley, C. W. Application of Combinatorial Chemistry Science on Modern Drug Discovery. *J. Comb. Chem.* **2008**, *10*, 345−354 and references within..

(2) Truchon, J.; Bayly, C. GLARE: A New Approach for Filtering Large Reagent Lists in Combinatorial Library. Design Using Product Properties. *J. Chem. Inf. Model.* **2006**, *46*, 1536−1548.

(3) Peng, Z.; Yang, B.; Mattaparti, S.; Shulok, T.; Thacher, T.; Kong, J.; Kostrowicki, J.; Hu, Q.; Na, J.; Zhou, J. Z.; Klatte, D.; Chao, B.; Ito, S.; Clark, J.; Sciammetta, N.; Coner, B.; Waller, C.; Kuki, A. PGVL Hub: An Integrated Desktop Tool for Medicinal Chemists to Streamline Design and Synthesis of Chemical Libraries and Singleton Compounds. *Methods Mol. Biol.* **2011**, *685*, 295−320.

(4) For the Pfizer file enrichment initiative, see: (a) Milne, G. M. Pharmaceutical Productivity: The Imperative for New Paradigms. *Annual Report of Medicinal Chemistry*; Academic Press: New York, 2003; Vol. *38*, Chapter 35, pp 383−396. (b) Estep, K. *File Enrichment and Hit Follow Up: Evolution and Examples*; ALA LabFusion: Boston, MA, 2004. (c) Smith, G. F. Enabling HTS Hit follow up via Chemo informatics, File Enrichment, and Outsourcing. In *High Throughput Medicinal Chemistry II*; MMS Conferencing & Events Ltd., Institute of Physics; London, 2006. This article is also on-line (http://www.mmsconferencing.com/pdf/htmc/g.smith.pdf).

(5) For the Alanex effort in synthetic feasible virtual space, see: (a) Polinsky, P.; Feinstein, R. D.; Shi, S.; Kuki, K.; LiBrain: Software for Automated Design of Exploratory and Targeted Combinatorial Libraries. *Molecular Diversity and Combinatorial Chemistry*; Chaiken, I. M., Handa, K. D., Eds.; American Chemical Society: Washington, DC, 1996; pp 219−232. (b) Shi, S.; Kuki, K.; Zhou, J.; Na, J.; Thacher, T.; Yanovsky, A.; Polinsky, P. LiBrain, An Intelligent System for the High-Throughput Design of Combinatorial libraries in Drug Discovery, Poster Presentations at the Fifth International Conference on Chemical Structures, 1999, and the Second European Conference on Strategies and Technologies for Identification of NOVEL BIOACTIVE COMPOUNDS, 1998.

(6) (a) Andrews, K. M.; Cramer, R. D. Toward General Methods of Targeted Library Design: Topomer Shape Similarity Search with Diverse Structures as Queries. *J. Med. Chem.* **2000**, *43*, 1723−1740. (b) Cramer, R. D.; Soltanshahi, F.; Jilek, R.; Campbell, B. AllChem: Generating and Searching 10^{20} Synthetically Accessible Structures. *J. Comput.-Aided. Mol. Des.* **2007**, *21*, 341−350.

(7) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D.; Khachko, D.; Fokin, V.; Queen, C.; Zosimov, V. A Very Large Diversity Space of Synthetically Accessible Compounds for Use with Drug Design Programs. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 47−63.

(8) Manly, J. C. Managing laboratory automation: integration and informatics in drug discovery. *J. Autom. Methods Manage. Chem.* **2000**, *22*, 169−170.

(9) Lessel, U.; Wellenzohn, B.; Lilienthal, M.; Claussen, H. Searching Fragment Spaces with Feature Trees. *J. Chem. Inf. Model.* **2009**, *49*, 270−279.

(10) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical

Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(11) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.
(12) Pipeline Pilot, Accelrys, Inc. http://accelrys.com/products/pipeline-pilot/.
(13) Specially PLP extension to ISIS Draw: For example, the Sn2 reaction is "sp2", "sp3" advanced techniques are used. A good example would be a simple Sn2 phenol alkylation where Cl, Br, and I are used attached to a carbon, which is annotated as "sp3". In the case of deprotection step the "clip" annotation can be used on the atoms of the protection group which are attached to the protected atom to indicate that all atoms beyond the annotated atoms should not be present in the final product structures.
(14) ISIS-Draw, Accelrys, Inc. http://accelrys.com/products/pdf/ISISDRAW.pdf.
(15) (a) Leach, A. R.; Bradshaw, J.; Green, D. V. S.; Hann, M. M. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1161−1172. (b) Yasri, A.; Berthelot, D.; Gijsen, H.; Thielemans, T.; Marichal, P.; Engels, M.; Hoflack, J. REALISIS: A Medicinal Chemistry-Oriented Reagent Selection, Library Design, and Profiling Platform. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2199−2206.
(16) Available Chemicals Directory (ACD), Accelrys, Inc. http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html.
(17) Shi, S.; Peng, Z.; Kostrowicki, J.; Paderes, G.; Kuki, A. Efficient Combinatorial Filtering for Desired Molecular Properties of Reaction Products. *J Mol. Graph. Model.* **2000**, *18*, 478−496.
(18) Zhou, Z.; Shi, S.; Na, J.; Peng, Z.; Thacher, T. Combinatorial Library-Based Design with Basis Products. *J. Comput-Aided Mol. Des.* **2009**, *23*, 725−736.
(19) Lau, W.; Hepworth, D.; Magee, T.; Du, J.; Bakken, G.; Miller, M.; Hendsch, Z.; Thanabal, V.; Kolodziej, S.; Xing, L.; Hu, Q.; Narasimhan, L.; Love, R.; Charlton, M.; Hughes, S.; Van Hoorn, W.; Mills, J.; Withka, J. Design of a Multipurpose Fragment Screening Library using Molecular Complexity and Orthogonal Diversity Metrics. *J. Comput.-Aided Mol. Des.* **2011**, *25*, 621−636.
(20) Hu, Q.; Peng, Z.; Kostrowicki, J.; Kuki, A. LEAP into the Pfizer Global Virtual Library (PGVL) Space—Creation of Readily Synthesizable Design Ideas Automatically. *Methods Mol. Biol.* **2011**, *685*, 253−276.
(21) Klis, T.; Lulinski, S.; Serwatowski. Remote-Substituent-Directed Metalations of Arenes. *J. Curr. Org. Chem.* **2008**, *12* (17), 1479−1501.
(22) (a) Gasteiger, J.; Pforner, M.; Sitzmann, M.; Hollering, R.; Sacher, O.; Kostka, T.; Karg, N. Computer-assisted Synthesis and Reaction Planning in Combinatorial Chemistry. *Perspect. Drug Discovery Des.* **2000**, *20*, 245−264. (b) Branan, M.; Pop, L.; Willard, X.; Horvath, D. Reactivity Prediction Models Applied to the Selection of Novel Candidate Building Blocks for High-Throughput Organic Synthesis of Combinatorial Libraries. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1119−1127.
(23) (a) Teng, M.; Zhu, J.; Johnson, M. D.; Chen, P.; Kornmann, J.; Chen, E.; Blasina, A.; Register, J.; Anderes, K.; Rogers, C.; Deng, Y.; Ninkovic, S.; Grant, S.; Hu, Q.; Lundgren, K.; Peng, Z.; Kania, R. S. Structure-Based Design of (5-Arylamino-2H-pyrazol-3-yl)-biphenyl-2′,4′-diols as Novel and Potent Human CHK1 Inhibitors. *J. Med. Chem.* **2007**, *50*, 5253−5256. (b) Peng, Z.; Hu, Q. Design of Targeted Libraries against the Human Chk1 Kinase Using PGVL Hub. *Methods Mol. Biol.* **2011**, *685*, 321−336.
(24) Ryckmans, T; Edwards, M.; Horne, V.; Correia, A.; Owen, D.; Thompson, L.; Tran, I.; Tutt, M.; Young, T. Rapid Assessment of a Novel Series of Selective CB(2) Agonists Using Parallel Synthesis Protocols: A Lipophilic Efficiency (LipE) Analysis. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 4406−4409.
(25) Xu, L.; Desai, M. C. Pharmacokinetic Enhancers for HIV Drugs. *Curr. Opin. Invest. Drugs* **2009**, *10* (8), 775−786.
(26) Patterson, D.; Sakata, S.; Nambu, M.; Patel, Leena.; Tatlock, J. Preparation of Pyridinylaminosulfonylarylcarboxamides As Cytochrome P450 3A4 Inhibitors. PCT Int. Appl. WO 2007034312 /A2 20070329, 2007.
(27) (a) Li, H.; Tatlock, J.; Linton, A.; Gonzalez, J.; Jewell, T.; Patel, L.; Ludlum, S.; Drowns, M.; Rahavendran, S. V.; Skor, H.; et al. Discovery of (R)-6-Cyclopentyl-6-(2-(2,6-diethylpyridin-4-yl)ethyl)-3-((5,7-dimethyl-[1,2,4]triazolo[1,5-a]pyrimidin-2-yl)methyl)-4-hydroxy-5,6-dihydropyran-2-one (PF-00868554) as a Potent and Orally Available Hepatitis C Virus Polymerase Inhibitor. *J. Med. Chem.* **2009**, *52* (5), 1255−1258. (b) Johnson, S.; Drowns, M.; Tatlock, J.; Linton, A.; Gonzalez, J.; Hoffman, R.; Jewell, T.; Patel, L.; Blazel, J.; Tang, M.; et al. Synthetic Route Optimization of PF-00868554, An HCV Polymerase Inhibitor in Clinical Evaluation. *Synlett* **2010**, *5*, 796−800. (c) Li, H.; Tatlock, J.; Linton, A.; Gonzalez, J.; Borchardt, A.; Dragovich, P.; Jewell, T.; Prins, T.; Zhou, R.; Blazel, J.; Parge, H.; Love, R.; Hickey, M.; Doan, C.; Shi, S.; Duggala, R.; Lewis, C.; Fuhrman, S. Identification and Structure-Based Optimization of Novel Dihydropyrones As Potent HCV RNA Polymerase Inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16* (18), 4834−4838.