ARTICLES

Project-Focused Activity and Knowledge Tracker: A Unified Data Analysis, Collaboration, and Workflow Tool for Medicinal Chemistry Project Teams

Marian D. Brodney,[†] Arthur D. Brosius,*,[†] Tracy Gregory,[†] Steven D. Heck,[†]
Jacquelyn L. Klug-McLeod,[‡] and Christopher S. Poss[§]

Data and Design Analytics, Computational Sciences Center of Emphasis, and Neuroscience Chemistry, Pfizer Global Research and Development, Eastern Point Road, Groton, Connecticut 06340

Received July 10, 2009

Advances in the field of drug discovery have brought an explosion in the quantity of data available to medicinal chemists and other project team members. New strategies and systems are needed to help these scientists to efficiently gather, organize, analyze, annotate, and share data about potential new drug molecules of interest to their project teams. Herein we describe a suite of integrated services and end-user applications that facilitate these activities throughout the medicinal chemistry design cycle. The Automated Data Presentation (ADP) and Virtual Compound Profiler (VCP) processes automate the gathering, organization, and storage of real and virtual molecules, respectively, and associated data. The Project-Focused Activity and Knowledge Tracker (PFAKT) provides a unified data analysis and collaboration environment, enhancing decision-making, improving team communication, and increasing efficiency.

INTRODUCTION

Much has been written about the recent explosion of data in the drug discovery arena and the efforts to integrate and present this data in a meaningful way to researchers.[1] High-throughput screening, the search for novel mechanisms of action, advances in our understanding of biological pathways, drug metabolism, and safety endpoints, and an increased use of in silico techniques have led to this data explosion. Coupled with an urgency to improve productivity and shorten drug discovery timelines, these increases call for better techniques for organization and storage of data and mechanisms for presenting it to the researcher in a way that fosters effective analysis and decision-making.

Reports have emerged from various companies describing the construction of data warehousing strategies and integrated data presentation platforms in response to these challenges.[2−5] Specifically, our medicinal chemistry project teams were seeking a system that would

- decrease the amount of time required to gather, organize, and transform measured and in silico data on *real* compounds;
- foster collaboration by capturing design hypotheses and *virtual* compound ideas across the entire project team;
- promote unbiased, data-driven decision-making by (a) calculating an agreed-upon set of in silico properties for every designed molecule and (b) integrating data for *real* compounds into the same environment in which design hypotheses are created and *virtual* molecules are designed;
- present data on *real* and *virtual* compounds in a visually engaging and customizable environment;
- support collaborative annotation and knowledge capture;
- provide basic project workflow and team communication support by capturing and reporting time stamps, contacts, statuses, and other metadata for molecules and designs.

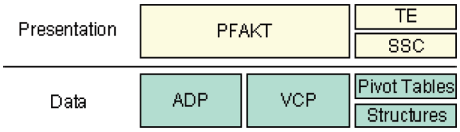


Figure 1. Dual-layer solution architecture: PFAKT, Project-Focused Activity and Knowledge Tracker; TE, Target Editor; SSC, Structure Search Client; ADP, Automated Data Presentation; VCP, Virtual Compound Profiler.

This work is an attempt to respond to the data explosion problem in general and to the specific requirements of our medicinal chemistry project teams. We describe a suite of integrated services and end-user applications that enable medicinal chemistry project teams to gather, organize, analyze, annotate, and communicate about both *real* and *virtual* compounds, their design hypotheses, measured and calculated data, and any knowledge gained through the study of those compounds. Our system supports all aspects of the design cycle: from conception of a design idea to creation, analysis, and selection of *virtual* candidate compounds for organic synthesis, to synthesis, biological testing, and data analysis, leading to the next round of design ideas.

The overall architecture of our solution may be seen as two “layers”; a data layer, which stores data relevant to the project team after it has been extracted from primary data sources and pivoted, and a presentation layer, which provides the user-interface components (Figure 1). The Automated

[*] Corresponding author e-mail: arthur.d.brosius@pfizer.com.
[†] Data and Design Analytics.
[‡] Computational Sciences Center of Emphasis.
[§] Neuroscience Chemistry.

10.1021/ci9002443 CCC: $40.75
Published on Web 11/09/2009

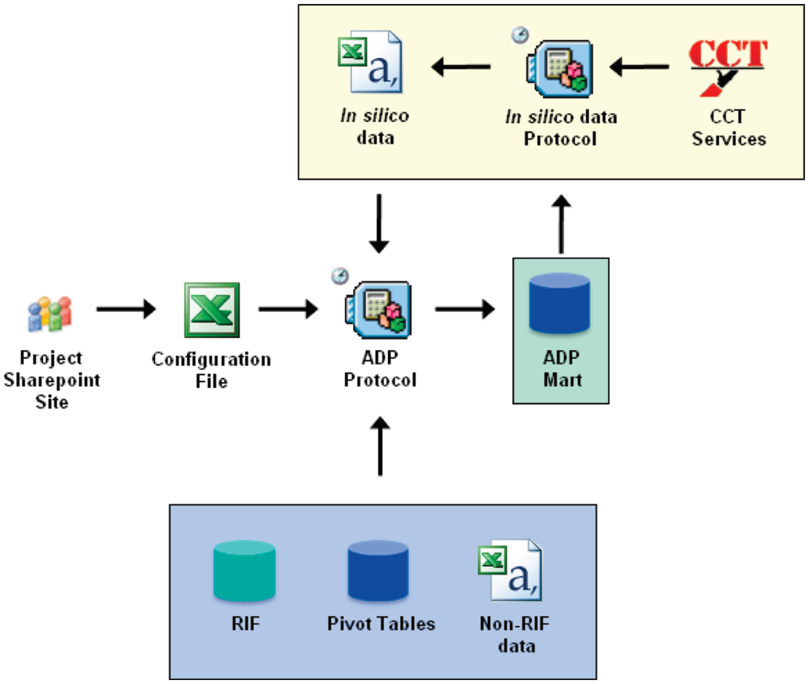


Figure 2. Schematic of data flow in the ADP process.

Data Presentation (ADP) and Virtual Compound Profiler (VCP) support real and virtual molecules, respectively, and form the core of the data layer, with a number of additional database tables providing specialized data. The main component of the presentation layer is the Project-Focused Activity and Knowledge Tracker (PFAKT) system. PFAKT is accompanied by the Target Editor (TE) and Structure Search Client (SSC) applications, which provide the ability to register virtual molecules and search for chemical structures in PFAKT, respectively. Below we will discuss the individual components the data layer and how this data is presented to the user in PFAKT.[6,7]

RESULTS AND DISCUSSION

Automated Data Presentation (ADP). Data for real compounds was aggregated into "pivoted" project-specific data marts in an Oracle database. Each row in the mart represents a single compound and its associated data, where the primary key is the compound's internal Pfizer ID number.

Two primary data sources are used to populate the ADP mart (Figure 2). The first is Pfizer's data warehouse, the Research Information Factory (RIF), which houses the majority of biological screening data. Assay data in the RIF environment is unpivoted, meaning there are an arbitrary number of rows in the database for a given Pfizer ID. Each row is distinguished from another by one or more designators—typically an assay code, an assay date, and an experiment ID. Certain subsets of RIF data from locally or globally shared biological assays are also prepivoted by a separate Pipeline Pilot process, resulting in several Oracle marts generally referred to as "pivot tables".[8] The prepivoting process is a performance enhancement technique that also facilitates data visualization (see the "Pivot Tables" section below). The second primary data source is Pfizer's Computational Chemistry Toolbox Services (CCT Services),[9] a grid computing based infrastructure that executes in silico model calculations for both real and virtual compounds.

The ADP process begins with a configuration file, implemented in Microsoft Excel. Project team members use this file to indicate which compounds to include in the mart, which data (both "wet" and in silico model data) to be retrieved, and any data transformations desired (e.g., selectivity ratios, log transformations, dose prediction equations, etc.). The configuration file resides in a Microsoft Sharepoint document repository where it can be accessed and updated by any project team member.

A generic Pipeline Pilot[10] protocol was written to perform the extraction, transformation, and data loading process. Top-level parameters are used to configure the protocol with project-specific links to the configuration file, database tables, and output folders. Despite the generic nature of the Pipeline Pilot protocol and the use of an external configuration file, we chose to create a copy for each project, allowing for the inclusion of project-specific customizations (typically complex data transformations).

The protocol, which runs every 15 min during working hours (see the "Automation" section below), is first responsible for performing a number of status and error checks, including checking for the presence of the configuration file and the ADP mart. The protocol then determines whether the configuration file has been updated since the last protocol execution by comparing the file's modification time stamp with a "last run" time stamp stored in an Oracle audit table. The automated process then downloads a copy of the configuration file to the Pipeline Pilot server and reads it. Next the protocol queries the RIF to determine if new data has been loaded since the protocol's last run. If no new data of interest to the team has been loaded, the protocol halts and runs again in 15 min.

If new assay data is detected, the protocol builds a list of compounds of interest to the team. Compound lists are generated by one or more of the following three methods, as indicated in the configuration file: (a) by including all compounds with screening data for a given assay or assays,

(b) by including all compounds that have been registered in Pfizer's corporate compound database under a particular project code, and (c) via lists of Pfizer IDs to explicitly include (or exclude).

Data for this aggregated compound list is extracted from the RIF and optionally from flat files or additional specialty databases, etc. Unpivoted data is automatically pivoted by the protocol, where a combination of the assay or transformation name, assay code, and endpoint type are used to dynamically construct a unique column name (a "property" in Pipeline Pilot terms), resulting in one record per Pfizer ID.

The following types of data are retrieved and incorporated into the ADP output by this process:

- Primary assay data: endpoint value (e.g., IC50, EC50, K_i, etc.), number of tests ("n"), % effect at max and min dose, and first test date for each assay, as defined per assay in the configuration file.
- In vitro AMDET data: human liver microsome $t_{1/2}$, Caco flux, cell viability, etc. This data is sourced from the pivot tables.
- Precalculated molecular property data stored in the RIF: log D, H-bond donor count, etc.
- In vivo ADME endpoint data: iv rat clearance, volume of distribution, $t_{1/2}$, etc.
- Compound registration info: most recent batch creator, registration date, amount made, purity, etc.
- In silico model data, e.g., predicted human liver microsome $t_{1/2}$, predicted potency etc.

In silico data is managed differently from other data streams. Since the results of in silico model predictions do not change unless the models themselves are rebuilt, we store a file of aggregated model results for each project team. A separate helper Pipeline Pilot protocol regularly checks for new compounds in the team's ADP mart and processes them through their set of in silico models. The file of aggregated results is then merged in to the ADP protocol. Since many models are rebuilt with some regularity, we also use this helper protocol to completely rebuild the file of in silico results as needed.

After all measured and calculated data has been retrieved, a number of different transformations are optionally performed on the data and molecular structures represented.

- Data transformations: ratios, averages, log transforms, inverse, binning, ligand efficiency,[11] and "priority combine" (e.g., if value A exists, report it, otherwise report value B), etc. Parameters for performing these calculations are supplied by the team in the configuration file.
- Series tagging: molecules in the ADP can be "tagged" with a chemical series designation using substructures and series tags supplied by the team.
- R-group deconvolution: similar to series tagging, this process uses team-defined substructures that are additionally marked with numbered substitution points (R1, R2, etc.). Molecules in the ADP are tagged with a "core" name and the IUPAC name and SMILES of each R group in columns named "R1_NAME", "R1_SMILES", "R2_NAME", "R2_SMILES", etc.

With the data extraction and transformation steps complete, the data is then prepared for loading into the ADP mart. Column names are made Oracle-compliant by purging disallowed characters and enforcing Oracle's 30-character limit. Two SQL statements are constructed dynamically: (a) a CREATE TABLE statement and (b) an INSERT statement. We chose to delete and recreate the table each time the ADP ran. In this way, we minimized the complexity of dynamically generating UPDATE and INSERT statements when data changed and/or ALTER TABLE statements when the table structure changed. Thus, each time the ADP runs to completion, it issues a DROP TABLE statement to the Oracle database and immediately creates a (possibly different) table using the CREATE TABLE statement. Each row in the data stream is then written to the new table using the INSERT statement with dynamic variable substitution of the actual data values.

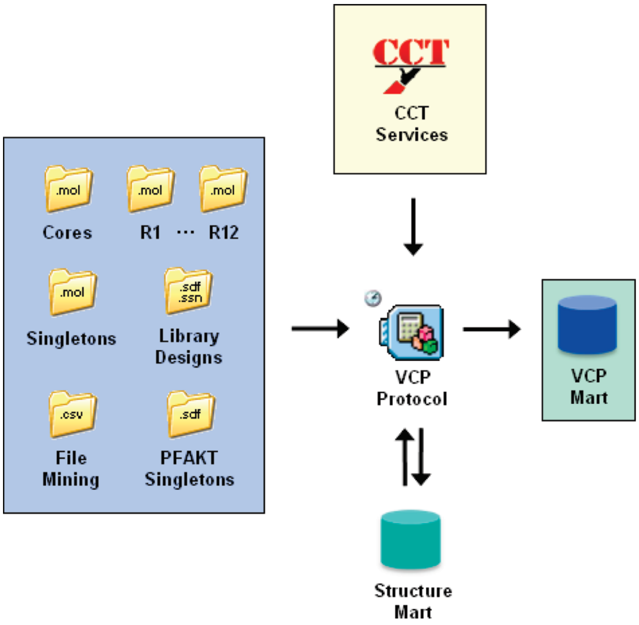


Figure 3. Schematic of data flow in the VCP process.

After the mart has been written, the ADP audit table is updated with the appropriate time stamps. The data is then output into several delimited text files that facilitate interoperability with Spotfire[12] and other internal data visualization packages. Finally, the Pipeline Pilot protocol uploads these files to the team's Sharepoint document repository. The entire process can take anywhere from 20 min to 2 h, depending primarily on the number of columns in the data set. A typical ADP mart holds 5000−20 000 records.

Virtual Compound Profiler (VCP). In an effort to foster a collaborative environment for medicinal chemistry design teams, we created a process we call the VCP (Figure 3). This companion to the ADP process provides a way for designers on the team to create, annotate, share, and analyze properties of potential targets for chemical synthesis and/or testing, so-called "virtual" compounds. The VCP encourages in silico design by calculating in silico properties for all virtual molecules. And since the team chooses the set of properties that will be calculated across all molecules, each potential synthetic target is evaluated using the same criteria, regardless of who designed it. Finally, by electing the same set of in silico models for both the ADP and VCP, both data sets can be more readily integrated, encouraging the use of in silico and measured properties of real compounds when deciding which compounds to synthesize next.

As in the ADP, the VCP process utilizes a Pipeline Pilot protocol, and the final output is a pivoted, project-specific mart in an Oracle database. Each row in the database represents a unique chemical structure.

Because ideas for new medicinal chemistry targets come about in a variety of different ways, we have provided a variety of methods for populating the VCP with virtual compounds, as follows:

1. Ideas for individual molecules can be drawn in a chemical drawing package (e.g., ISIS/Draw[13]) and exported in the mol file format. The user saves this file in a specific subfolder of a shared network folder created for their project team.

2. Pre-enumerated collections of virtual compounds may be exported in .sdf format and/or uploaded to the shared folder in one of the file formats utilized by Pfizer's internal library design tool, PGVL Hub.[14]

3. The team may design "core" fragments where the core molecule is decorated with numbered R groups (R1, R2, etc.) and complementary numbered R group fragments where the attachment points are indicated with Z groups (Z1, Z2, etc.). Cores and R group fragments are stored in .mol format in subfolders labeled "Cores", "R1", etc. The VCP Pipeline Pilot protocol performs a full cross-enumeration of all cores with all R group fragments.

4. The results of structure searches, docking algorithms, or other such file-mining exercises performed on the set of "real" compounds in the Pfizer corporate compound database may be included. A delimited file containing Pfizer ID numbers is added to a subfolder on the shared folder.

5. Individual molecules may be entered directly into PFAKT using the Target Editor application. PFAKT periodically exports an SD file of these molecules for incorporation into the VCP. This method will be discussed in more detail below.

A copy of the VCP Pipeline Pilot protocol is created for each project team and customized with top-level parameters, indicating the location of the team's shared network folder, the project code, etc. Every 15 min, the protocol is triggered and checks first to see if any modifications have been made to the files in the shared folder. If no modifications are detected, the protocol stops and checks again in 15 min. If changes are detected, the protocol reads from the various subfolders, performing enumerations of cores and R-groups and converting lists of Pfizer IDs to structures as necessary.

Each structure is then checked for uniqueness against the Structure Mart table.[15] The Structure Mart table resides in an Oracle database equipped with the Accord Chemistry Cartridge[16] and is the definitive source for VCP structures across all projects. Each molecule is identified by a serial number, project code, and a human-readable "VCP_ID" which is unique for a molecule within a given project. Chemistry data is stored in the binary Accord and text-based MOL and Pipeline Pilot's CANONICAL SMILES[10] formats. The repository currently holds over 50 000 structures.

If the structure is novel within that project, a VCP_ID is generated and the molecule is loaded into the Structure mart. If the molecule is found in the Structure mart, its VCP_ID is retrieved from the database.

Novel molecules are then sent by the protocol to CCT services for in silico model calculations. A standard set of properties (e.g., Log *D*, MW, TPSA, rule-of-five, rotatable bonds, etc.) and the project-elected models are calculated. As with the ADP, this data is cached in delimited text files. Rather than recalculating these values for molecules the VCP has previously processed, the protocol instead merges in these cached properties from the delimited files. As before, if models are rebuilt, Cheminformatics staff can set a parameter to have these files regenerated for the entire set of VCP molecules.

Using the same team-supplied core and R-group definitions employed in the ADP (see the Automated Data Presentation (ADP) section above), each VCP molecule is assigned series and zero or more R-group labels. Employing series and R-group deconvolution procedures on both ADP and VCP compounds further facilitates the integration of real and virtual molecules during medicinal chemistry design decisions.

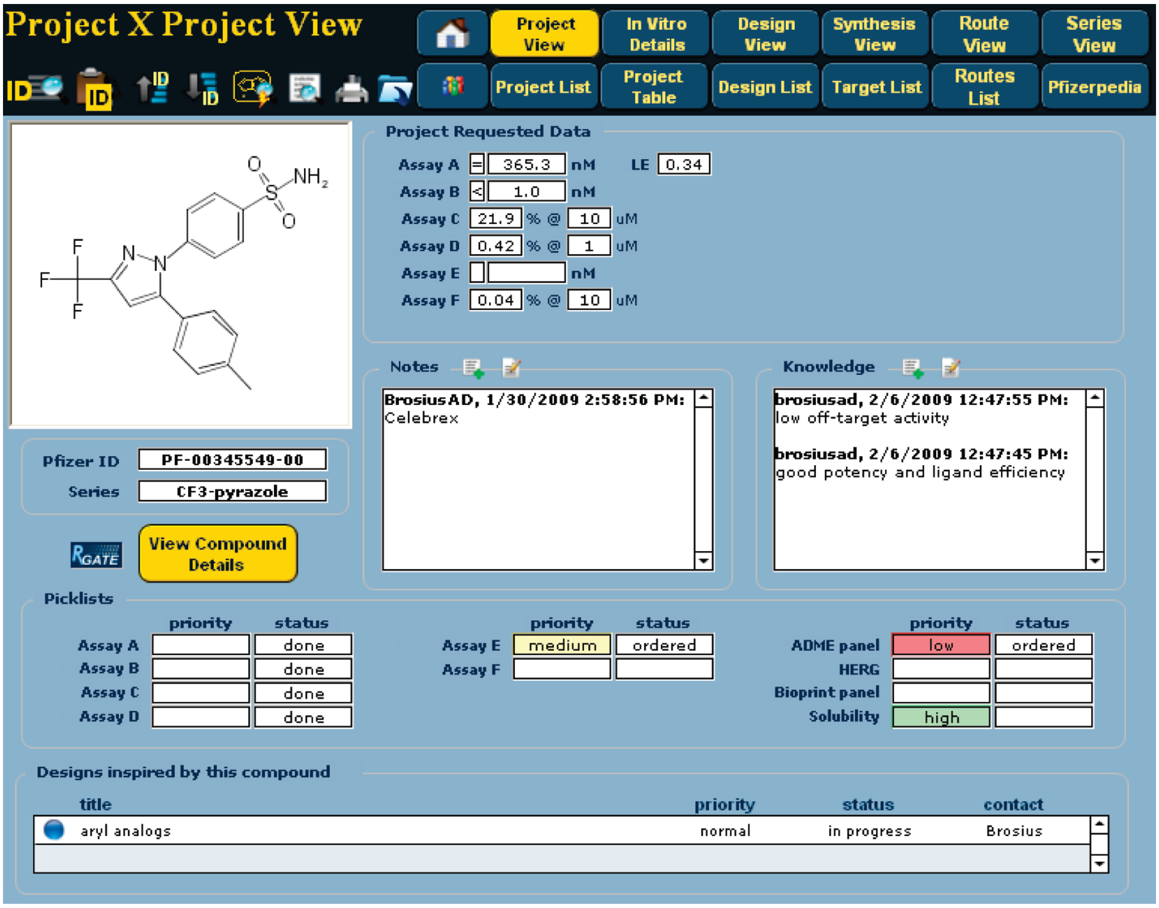
Since the purpose of aggregating these molecules together is to facilitate analysis and discussion about which molecules to synthesize and test, it follows that a certain number of these "virtual" molecules eventually become "real"; that is, they will be registered in the Pfizer corporate compound database. For this reason, the VCP protocol has a reference to the team's ADP output in order to determine if any of the molecules it is processing have received Pfizer IDs. If a match is found,[17] the Pfizer ID becomes a property of that VCP molecule. PFAKT uses the PFIZER_ID as a foreign key to link virtual molecules (and the designs from which they derive) to real molecules and their experimental data (vide infra).

In a nearly identical process to that employed in the ADP protocol, the VCP molecules with their associated data are then written to Oracle. Again, as before, the data is also written to several delimited text files and uploaded to the team's Sharepoint site.

To support the creation of virtual compounds from within PFAKT, two additional Pipeline Pilot protocols have been developed. One protocol performs uniqueness checking and registration of new molecules in the Structure mart. The second calculates the standard set of in silico models mentioned previously and issues an INSERT statement to merge that data into the team's VCP mart. Both protocols are called from within the Target Editor application using Pipeline Pilot's Client .NET SDK.

Pivot Tables.[8] As mentioned above during the discussion of the ADP process, we elected to extract, pivot, and load a subset of RIF data to facilitate both the ADP process and the visualization of such data within the PFAKT application. In vitro ADMET assays, which are typically performed by centralized global or site-wide groups, gene family screening panels (kinases, PDEs, etc.), cross-pharmacology panels such as the BioPrint profile,[18] chemical purity, and inventory from our local liquid compound logistics operation were included in this process.

Unlike the ADP process, where a generic SQL query can be used to retrieve data from the RIF for any arbitrary biological assay, construction of the pivot tables requires SQL queries specific to each assay being extracted. The process begins with generating a list of assay codes to be included in the particular pivot table. This list is then fed to a Pipeline Pilot protocol that analyzes the RIF and outputs a SQL query that can be used to pivot the data for that assay and create unique column names. These SQL queries are



Project Requested Data

Assay A	=	365.3	nM	LE	0.34
Assay B	<	1.0	nM		
Assay C	21.9	% @	10	uM	
Assay D	0.42	% @	1	uM	
Assay E			nM		
Assay F	0.04	% @	10	uM	

Pfizer ID: PF-00345549-00

Series: CF3-pyrazole

Notes

BrosiusAD, 1/30/2009 2:58:56 PM:
Celebrex

Knowledge

brosiusad, 2/6/2009 12:47:55 PM:
low off-target activity

brosiusad, 2/6/2009 12:47:45 PM:
good potency and ligand efficiency

Picklists

	priority	status
Assay A		done
Assay B		done
Assay C		done
Assay D		done

	priority	status
Assay E	medium	ordered
Assay F		

	priority	status
ADME panel	low	ordered
HERG		
Bioprint panel		
Solubility	high	

Designs inspired by this compound

title	priority	status	contact
● aryl analogs	normal	in progress	Brosius

Figure 4. "Project View" layout in PFAKT. This layout provides a view of the structure, Pfizer ID, and ADP data for a "real" compound and provides a link to any related design records. Each layout in PFAKT has a toolbar that facilitates common activities such as list management, sorting, structure searching, reporting, and navigation.

then read by a second Pipeline Pilot protocol, which executes the queries and builds the pivot tables themselves.

Unlike the ADP and VCP processes, where the tables are deleted and recreated each run, data in the pivot tables is dynamically updated. We took this approach because the structure of the tables themselves is static between runs of the protocol. Thus, when data is retrieved for a compound that does not yet have an entry in the pivot tables, a simple INSERT statement is used. When data changes or is added for a pre-existing compound, the protocol generates an UPDATE statement. The pivot table protocols are triggered every 15 min. Loading of new data and compounds takes on the order of 1−5 min per table.

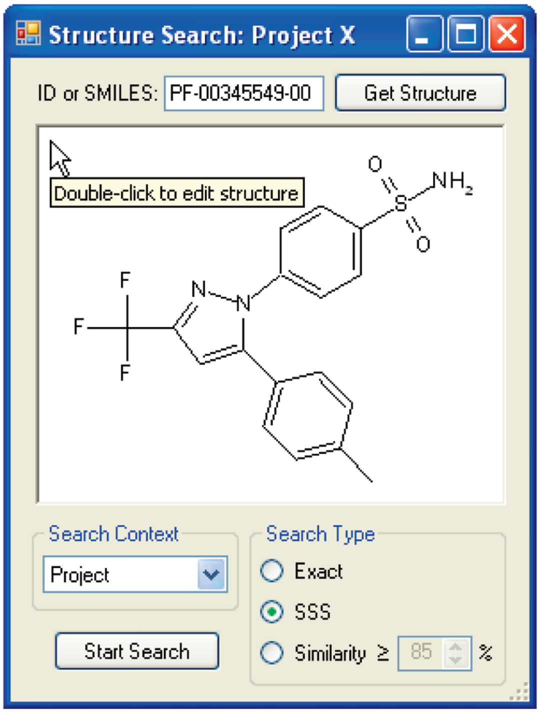
Periodically, new assays come online and are incorporated into this process. When we add new assays we perform a complete rebuild of that pivot table by deleting and rewriting the table. This process takes on the order of 15−30 min per table.

Automation. The Pipeline Pilot protocols that drive the ADP, VCP, and pivot table processes are triggered at defined intervals using a dedicated computer and the Microsoft Windows Scheduled Tasks utility. Schedules are staggered so as to minimize load on the Pipeline Pilot server infrastructure, with long-running tasks, such as rebuilding a team's in silico data file, scheduled for noncore hours. The Scheduled Tasks utility monitors the status of every job and will prevent overlapping executions of the same protocol.

The PFAKT application uses a syncing process to import the latest compounds and data from the ADP and VCP marts and to notify the VCP process of new molecules created with the Target Editor application. These syncing processes, which are written in FileMaker's ScriptMaker language,[19] are also triggered every 15 min by a Scheduled Task.

Project-Focused Activity and Knowledge Tracker (PFAKT). The PFAKT application is a unified data analysis, collaboration, and workflow tool, built on the FileMaker 9 platform.[19] In addition to providing views of ADP, VCP, pivot tables, and chemical structure data, team members use PFAKT to collaborate on medicinal chemistry design ideas and to record notes and catalog acquired knowledge about real and virtual compounds, designs, synthetic routes, and chemical series. Teams can prioritize candidates for chemical synthesis and assign those Targets to chemists in the laboratory. Real compounds can be queued for additional biological testing and sent to Pfizer's compound ordering system. Users can view and print reports of design records and their associated virtual compound properties, charts of the real and in silico properties of a chemical series, or Gantt charts to help identify bottlenecks in the design cycle. Data can be easily "pushed" to other visualization and analysis packages such as Spotfire, PCAT,[20] and Microsoft Excel.

Dozens of views are built-in to the PFAKT system, and with minimal training, users may customize built-in views or create new views that present the data in a way that fits



ID or SMILES: PF-00345549-00

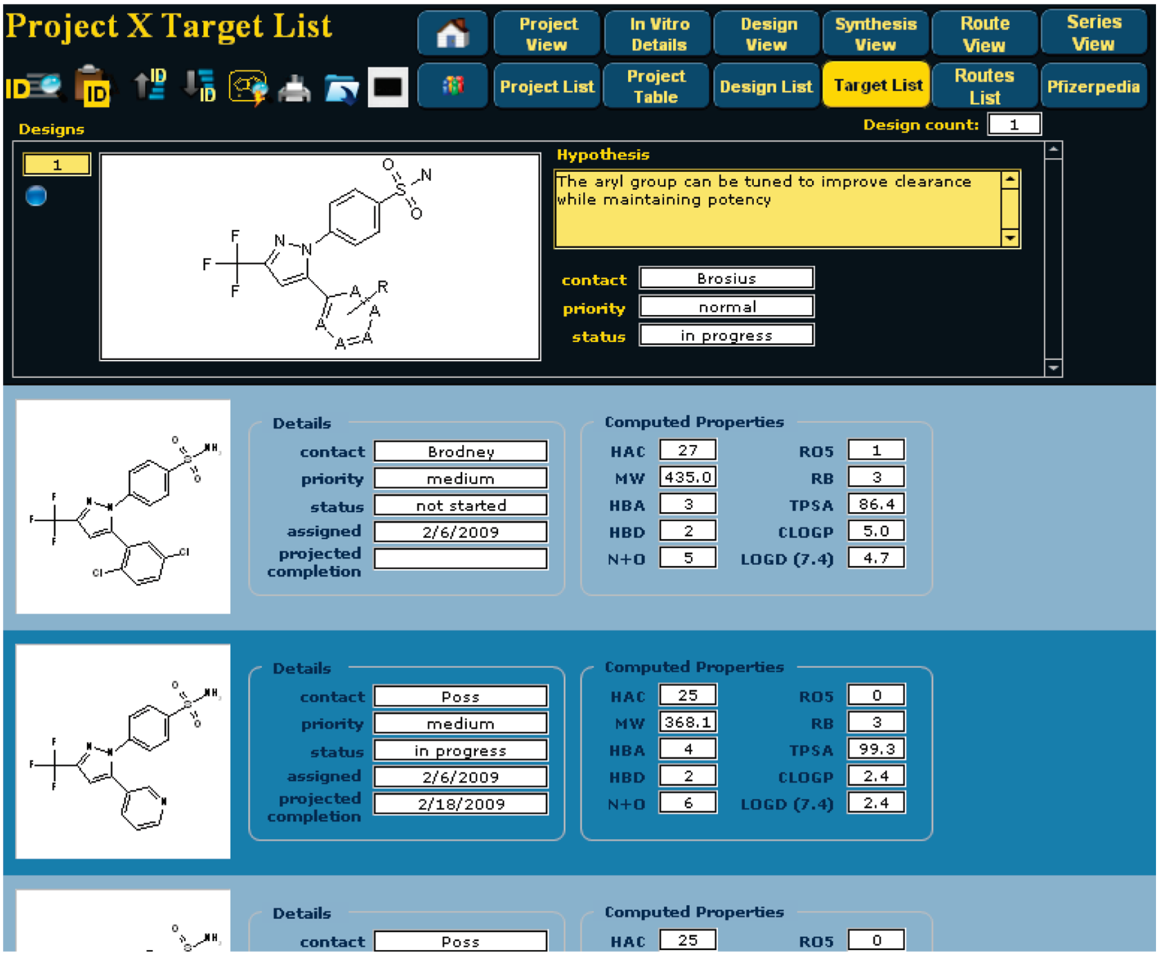
Search Context: Project

Search Type: ○ Exact ● SSS ○ Similarity ≥ 85 %

Figure 5. Structure Search Client (SSC) application. The SSC provides chemical structure searching services to PFAKT.

their data analysis needs. FileMaker's client/server architecture ensures that views and records added or modified made by one team member are seen by all other members in real time. The FileMaker platform provides an easy way to construct a relational database system from diverse data sources, including data stored externally in Oracle and other relational database management systems, and allows developers to add extensive automation to the user interface via the built-in scripting engine.

When a new project team forms, a copy of the PFAKT application file is made and customized by Cheminformatics staff to display that team's ADP and VCP data. Records in the ADP mart are mirrored as records in the "Project" sections of PFAKT. Here team members can view the data gathered as part of the ADP process. Searches may be performed on numerical fields (e.g., assay data, in silico model results), text fields (e.g., Pfizer ID, series, notes, status, etc.), and by structure. An example view of Project data is shown in Figure 4. This particular view is showing data from the PROJECT FileMaker database table (e.g., "Notes", "Assay E priority"), data from the ADP Oracle mart (assays A−F), data from the Pfizer Compound Lookup service (which renders the compound structure), and data from a related record in the DESIGN FileMaker table (see "designs inspired by this compound"), illustrating how multiple data sources are assembled to construct the PFAKT interface.



Design count: 1

Designs: 1

Hypothesis: The aryl group can be tuned to improve clearance while maintaining potency

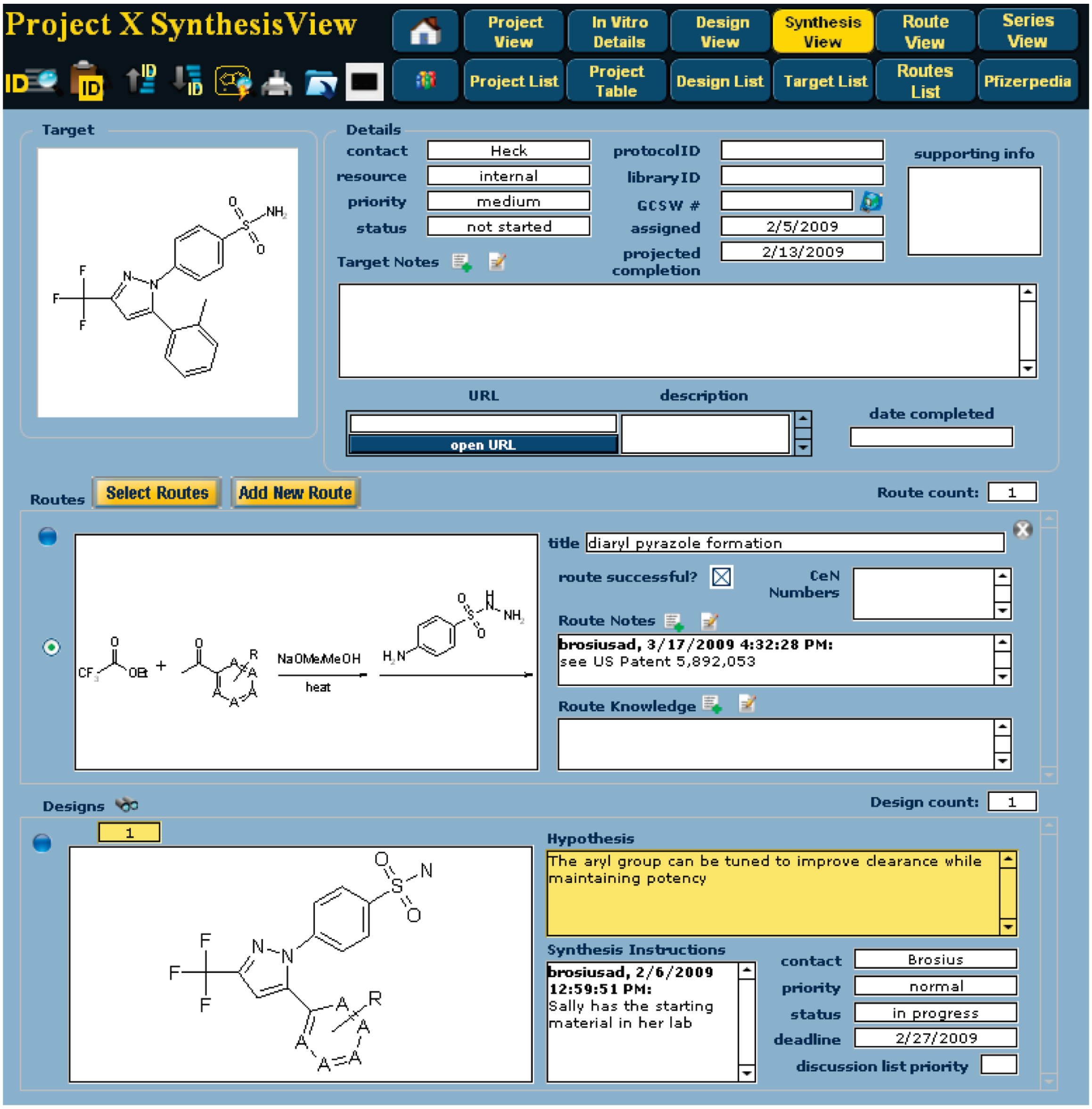
contact: Brosius
priority: normal
status: in progress

Details		Computed Properties			
contact	Brodney	HAC	27	RO5	1
priority	medium	MW	435.0	RB	3
status	not started	HBA	3	TPSA	86.4
assigned	2/6/2009	HBD	2	CLOGP	5.0
projected completion		N+O	5	LOGD (7.4)	4.7

Details		Computed Properties			
contact	Poss	HAC	25	RO5	0
priority	medium	MW	368.1	RB	3
status	in progress	HBA	4	TPSA	99.3
assigned	2/6/2009	HBD	2	CLOGP	2.4
projected completion	2/18/2009	N+O	6	LOGD (7.4)	2.4

Details		Computed Properties			
contact	Poss	HAC	25	RO5	0

Figure 6. "Target List" layout in PFAKT. Records in the VCP mart are presented as Targets in PFAKT. The Target structure, contact, priority, and other metadata, along with the results of in silico model calculations are visible here. Targets may be associated with Designs, which are displayed at the top of the Target list. Designs will be discussed below.



contact Heck
resource internal
priority medium
status not started
protocolID
libraryID
GCSW #
assigned 2/5/2009
projected completion 2/13/2009

Route count: 1

title diaryl pyrazole formation

route successful? ☒

Route Notes
brosiusad, 3/17/2009 4:32:28 PM:
see US Patent 5,892,053

Design count: 1

1

Hypothesis
The aryl group can be tuned to improve clearance while maintaining potency

Synthesis Instructions
brosiusad, 2/6/2009 12:59:51 PM:
Sally has the starting material in her lab

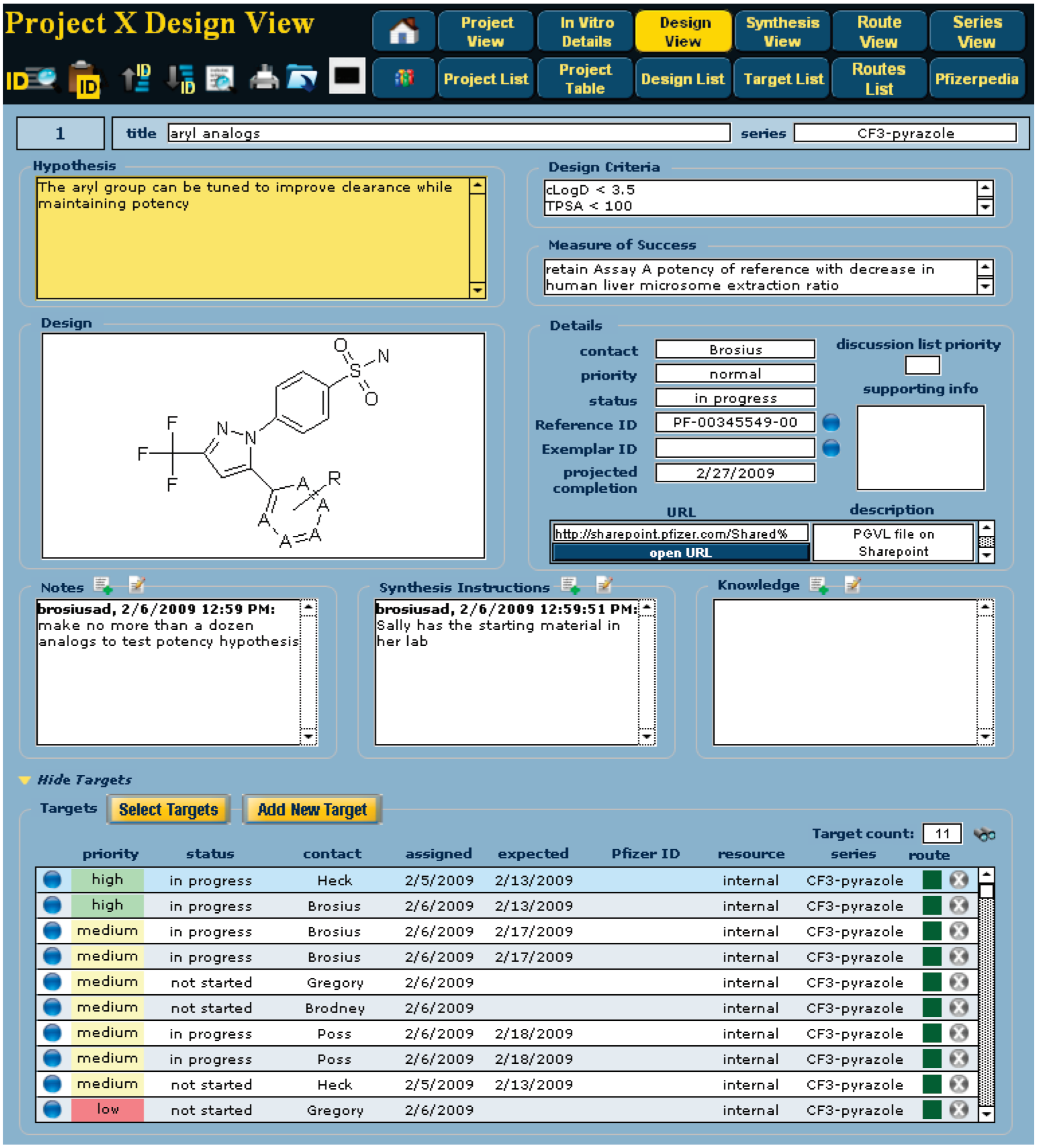
contact Brosius
priority normal
status in progress
deadline 2/27/2009
discussion list priority

Figure 7. "Synthesis View" layout in PFAKT, showing a Target and its associated metadata. Targets may be associated with one or more Routes and/or Designs, which will be discussed below.

Chemical structure searching in PFAKT is provided via the Structure Search Client (SSC) application (Figure 5). The SSC, launched via a button on the PFAKT toolbar, allows the user to sketch a structure from scratch or use a seed structure obtained by providing an ID or SMILES string and pressing the "Get Structure" button. After selecting a search context (choices include "Project" and "Targets", corresponding to ADP and VCP entries, respectively) and a search type, clicking the "Start Search" button will perform the query. A search context of "Project" issues a query to Pfizer's Compound Search service, whereas a search context of "Targets" initiates a search of the Structure mart (see the Virtual Compound Profiler (VCP) section above). Each service returns a list of matching Pfizer IDs, which are then displayed by PFAKT. Exact, substructure (SSS) and similarity-based search "types" are supported in both search contexts. As each structure service is built on the same chemical cartridge technology,[16] search results are consistent across search contexts. An "exact" search returns only the stereoisomer, tautomer, and/or ionization state drawn by the user.[15] "Similarity" searches use Accord's default fingerprint definitions and the Tanimoto similarity coefficient.

Data from nearly a dozen disparate FileMaker and Oracle databases is seamlessly presented to PFAKT users, providing an integrated analysis environment. In addition to the data sources mentioned above, additional data for each record in the Project sections of PFAKT can be drawn from the pivot tables (see the Pivot Tables section above) and from a variety of RIF tables that hold compound batch creator, inventory, screening status, and the like. Records in PFAKT and each of these auxiliary tables are joined by PFIZER_ID in one-to-one or one-to-many relationships.

Records in the VCP data mart are represented by records in the TARGET FileMaker table and are visible in "Target" and "Synthesis" views in PFAKT (Figures 6 and 7). In these views, users analyze the results of in silico model calculations, indicate relative priority for further follow-up, assign Targets to synthetic chemists, and record the status of ongoing synthesis efforts. As in the Project views, Notes and Knowledge sections are provided (see Figure 7). With the



1 title aryl analogs series CF3-pyrazole

Hypothesis
The aryl group can be tuned to improve clearance while maintaining potency

Design Criteria
cLogD < 3.5
TPSA < 100

Measure of Success
retain Assay A potency of reference with decrease in human liver microsome extraction ratio

Design

Details

contact	Brosius
priority	normal
status	in progress
Reference ID	PF-00345549-00
Exemplar ID	
projected completion	2/27/2009

discussion list priority

supporting info

URL	description
http://sharepoint.pfizer.com/Shared% open URL	PGVL file on Sharepoint

Notes
brosiusad, 2/6/2009 12:59 PM:
make no more than a dozen analogs to test potency hypothesis

Synthesis Instructions
brosiusad, 2/6/2009 12:59:51 PM:
Sally has the starting material in her lab

Knowledge

Hide Targets

Targets Select Targets Add New Target

Target count: 11

	priority	status	contact	assigned	expected	Pfizer ID	resource	series	route	
●	high	in progress	Heck	2/5/2009	2/13/2009		internal	CF3-pyrazole	🟩	⊗
●	high	in progress	Brosius	2/6/2009	2/13/2009		internal	CF3-pyrazole	🟩	⊗
●	medium	in progress	Brosius	2/6/2009	2/17/2009		internal	CF3-pyrazole	🟩	⊗
●	medium	in progress	Brosius	2/6/2009	2/17/2009		internal	CF3-pyrazole	🟩	⊗
●	medium	not started	Gregory	2/6/2009			internal	CF3-pyrazole	🟩	⊗
●	medium	not started	Brodney	2/6/2009			internal	CF3-pyrazole	🟩	⊗
●	medium	in progress	Poss	2/6/2009	2/18/2009		internal	CF3-pyrazole	🟩	⊗
●	medium	in progress	Poss	2/6/2009	2/18/2009		internal	CF3-pyrazole	🟩	⊗
●	medium	not started	Heck	2/5/2009	2/13/2009		internal	CF3-pyrazole	🟩	⊗
●	low	not started	Gregory	2/6/2009			internal	CF3-pyrazole	🟩	⊗

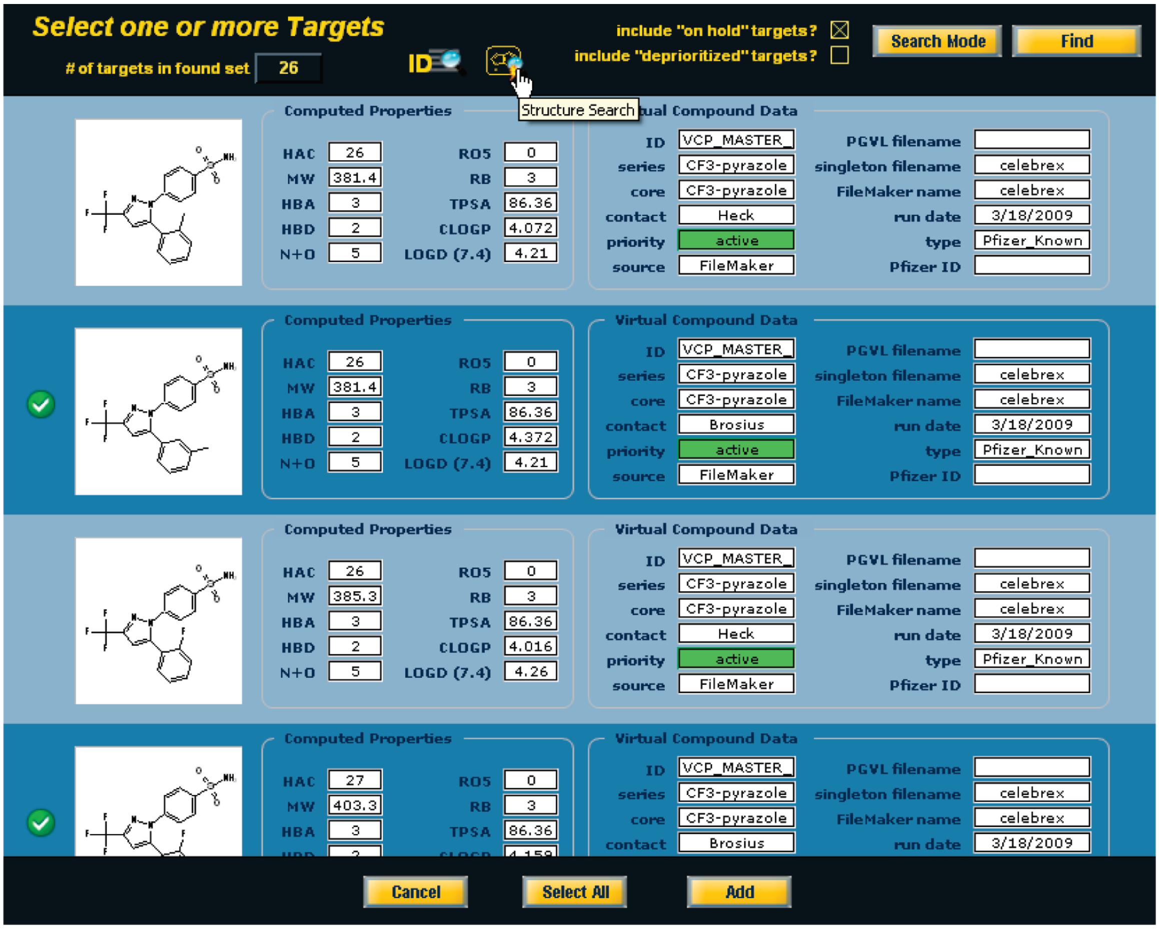
Figure 8. "Design View" layout in PFAKT.

click of a button, Target data can be pushed to alternative visual analysis tools like Spotfire or PCAT and the list of selected Targets later retrieved in PFAKT for assignment to a synthetic chemist, association to a Design, etc.

The DESIGN table and its corresponding views provide an interface to record medicinal chemistry Design ideas (Figure 8). Here designers communicate to their team members a visual depiction of their Design, a hypothesis, the criteria used for selecting Targets for the Design, and suggest a "measure of success" to evaluate the Design after the hypothesis has been tested. After the Design hypothesis has been tested, the team may record what they've learned in the "Knowledge" section. The "Design" section of PFAKT encourages knowledge sharing across the team and promotes the use of the entire body of knowledge acquired by the team when starting the next round of medicinal chemistry design.

Designs often encompass a genus of potential synthetic molecules (Targets). In PFAKT, users associate Target records with a Design record through a simple visual selection process (Figure 9). Users can drill-down to the Targets of interest by searching by any of the VCP data associated with that Target, including its chemical structure.

Users may also register new Targets in PFAKT by pressing the Add New Target button, which appears in multiple views in PFAKT (e.g., see Figure 8). The Add New Target button launches the Target Editor application (Figure 10). Using techniques familiar to those used in the Structure Search Client, users enter a structure for the new Target and click the Add Target button. The Target Editor verifies the uniqueness of this structure,[15] registers it in the Structure mart, and creates a new Target record in PFAKT. Users are prompted with the option to associate the new Target with the currently selected Design record. Users attempting to register a nonunique structure will be presented with the option to associate the pre-existing Target record with the current Design. Periodically, Targets added to PFAKT in



Select one or more Targets

# of targets in found set 26

include "on hold" targets? ☒

include "deprioritized" targets? □

HAC	26	RO5	0	ID	VCP_MASTER_	PGVL filename	
MW	381.4	RB	3	series	CF3-pyrazole	singleton filename	celebrex
HBA	3	TPSA	86.36	core	CF3-pyrazole	FileMaker name	celebrex
HBD	2	CLOGP	4.072	contact	Heck	run date	3/18/2009
N+O	5	LOGD (7.4)	4.21	priority	active	type	Pfizer_Known
				source	FileMaker	Pfizer ID	

✓ Computed Properties / Virtual Compound Data

HAC	26	RO5	0	ID	VCP_MASTER_	PGVL filename	
MW	381.4	RB	3	series	CF3-pyrazole	singleton filename	celebrex
HBA	3	TPSA	86.36	core	CF3-pyrazole	FileMaker name	celebrex
HBD	2	CLOGP	4.372	contact	Brosius	run date	3/18/2009
N+O	5	LOGD (7.4)	4.21	priority	active	type	Pfizer_Known
				source	FileMaker	Pfizer ID	

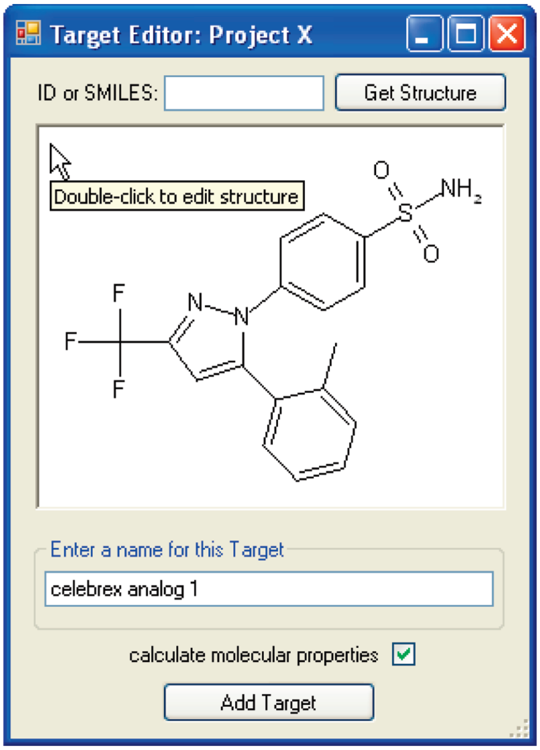
Computed Properties / Virtual Compound Data

HAC	26	RO5	0	ID	VCP_MASTER_	PGVL filename	
MW	385.3	RB	3	series	CF3-pyrazole	singleton filename	celebrex
HBA	3	TPSA	86.36	core	CF3-pyrazole	FileMaker name	celebrex
HBD	2	CLOGP	4.016	contact	Heck	run date	3/18/2009
N+O	5	LOGD (7.4)	4.26	priority	active	type	Pfizer_Known
				source	FileMaker	Pfizer ID	

✓ Computed Properties / Virtual Compound Data

HAC	27	RO5	0	ID	VCP_MASTER_	PGVL filename	
MW	403.3	RB	3	series	CF3-pyrazole	singleton filename	celebrex
HBA	3	TPSA	86.36	core	CF3-pyrazole	FileMaker name	celebrex
				contact	Brosius	run date	3/18/2009

Figure 9. Selecting Targets to associate with a Design.



Target Editor: Project X

ID or SMILES:

Enter a name for this Target: celebrex analog 1

calculate molecular properties ☑

Figure 10. The Target Editor (TE) application.

this manner are exported in SD format and incorporated into the team's VCP (see the Virtual Compound Profiler (VCP) and Automation sections above). The optional "calculate molecular properties" checkbox causes a predetermined set of approximately a dozen in silico models to be run and their results made visible in PFAKT in real time, without having to wait for the new Target to be incorporated into the team's VCP.

Since Designs and Targets are so closely linked, they are often shown in conjunction with one another in PFAKT layouts (Figures 6−8). Once associated, the properties of all Targets for a particular Design can be summarized and visualized to facilitate data analysis and decision making. One example is the radar plot shown in Figure 11, in which the black line represents the average values across six different calculated properties of the 11 Targets associated with Design record 1. The green regions inside the plot indicate the team's desired property space.

A ROUTE FileMaker table supports the sharing of synthetic route ideas. Route records can be associated with zero or more Targets, and each Target can be associated with zero or more Routes. Chemists can visually catalog and share useful synthetic routes, building the team's collective knowledge about how to most efficiently construct Target molecules (see Figure 7).

Finally, a SERIES table is provided. The ADP and VCP processes assign each molecule a series label (see the Automated Data Presentation (ADP) and Virtual Compound Profiler (VCP) sections above). PFAKT automatically generates a record for each unique SERIES record found within the ADP and VCP data sets for that team. Summary statistics

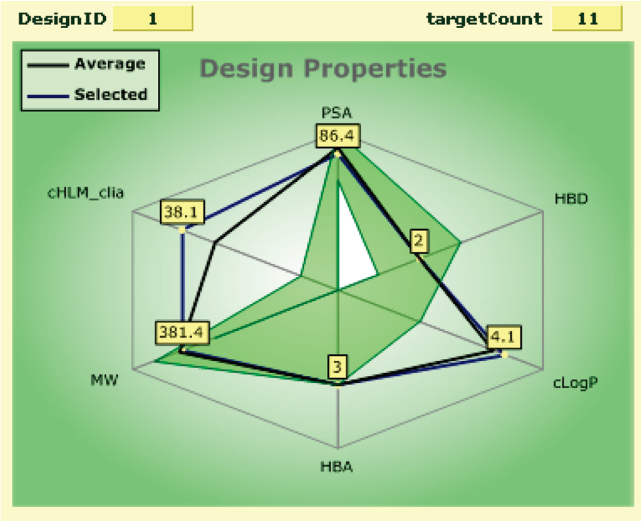


Figure 11. PFAKT automatically generates visualizations of Target properties associated to a Design.

(e.g., min/max/mean cLog *P*) and charts are generated by PFAKT to aid the comparison of one chemical series to another.

One of the unanticipated benefits of PFAKT has been the insight gained into how our medicinal chemistry teams operate on both a strategic and operational level. For example, on a strategic level, knowing the Design hypotheses, its associated Targets, the measured data on those Targets (via the ADP), and the amount of time taken from hypothesis generation to data collection is allowing us to assess whether or not the hypothesis was correct and how effectively and efficiently the hypothesis was tested by synthesis and testing of the associated Targets. We hope to translate the learnings from such analyses across multiple Designs and projects into more effective and efficient designs in the future.

Operationally, having a priority, status, contact, and assignment and completion dates for each Target that has been slated for synthesis (see Figure 8) allows a synthetic team leader to appropriately distribute the workload across the synthetic chemists and adjust staffing as priorities and timelines change. We have observed cases where PFAKT has illuminated gross inefficiencies in the way the synthetic team was operating, leading to substantial time savings in the synthetic operations of the team.

Comparisons to Other Industry Solutions. The system we have described has some concepts in common with other recently described solutions from our industry peers. For example, Johnson & Johnson's ABCD system[2] comprises a data warehouse, an elegant and elaborate user interface ("3DX"), and an external collaboration tool ("Explorer"). However, the two systems differ significantly in scope and purpose. ABCD's data warehouse and 3DX system would be best compared to Pfizer's RIF data warehouse and the accompanying query tool called RGate.[21] In contrast to ABCD, our system focuses on the data and collaboration needs of individual projects and adds the ability to manage new ideas that are spawned by the analysis of the data.

AstraZeneca has described a system[4] that is similar conceptually to the VCP and Design and Target portions of PFAKT, most notably in the ability to catalog and share design hypotheses and to calculate in silico properties of specific "target" molecules, all in a rich, collaborative user-interface environment. Another tool called "OnePoint",[5] developed at Pfizer's Sandwich, U.K. research site, has been used for general purpose project collaboration, including medicinal chemistry design.

CONCLUSION

We have described an integrated data analysis, collaboration, and workflow system for medicinal chemistry project teams. A series of automated processes gather, integrate, and store real and virtual molecules and their measured and calculated properties. Teams choose what molecules are retrieved in the ADP, what virtual molecules are added to the VCP, what data is retrieved and/or calculated, and what transformations are done to the data. All data for both real and virtual compounds of interest to the project team are brought together in a single user interface in the PFAKT application, enhancing decision-making, improving team communication, and increasing efficiency through the medicinal chemistry design cycle. Nearly 350 chemists at Pfizer's Groton Research site are users of the PFAKT system and the supporting technologies described herein, with expansions to other sites underway.

ACKNOWLEDGMENT

We thank Dan Kung, Bruce Lefker, Dave Griffith, and Matt Brown for helpful discussions, Steve Rieth and Dave Lees for IT support, Greg Bakken, Carmen Garcia-Irizarry, Dave Mack, Jared Milbank, and Soliant Consulting, Inc. for technical input. We also thank Alan Mathiowetz and Brian Bronk for support and encouragement throughout the course of this work.

REFERENCES AND NOTES

(1) Waller, C. L.; Shah, A.; Nolte, M. Strategies to support drug discovery through integration of systems and data. *Drug Discovery Today* **2007**, *12*, 634–639.

(2) Agrafiotis, D. K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E. P.; Konstant, P.; Leung, A.; Lobanov, V. S.; Marichal, P.; Martin, D.; Rassokhin, D. N.; Shemanarev, M.; Skalkin, A.; Stong, J.; Tabruyn, T.; Vermeiren, M.; Wan, J.; Xu, X. Y.; Yao, X. Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model.* **2007**, *47*, 1999–2014.

(3) Sander, T.; Freyss, J.; von Korff, M.; Reich, J. R.; Rufener, C. OSIRIS, an entirely in-house developed drug discovery informatics system. *J. Chem. Inf. Model* **2009**, *49*, 232–246.

(4) Robb, G. Hypothesis-driven drug design using wiki-based collaborative tools. Presented at the UK-QSAR and ChemoInformatics Group Meeting, May 14, 2009. http://www.documentarea.com/qsar/Grobb_Ukqsar_May09.pdf (accessed June 18, 2009).

(5) Barber, C. G.; Gardner, B.; Haque, N. "OnePoint"—Combining OneNote and SharePoint to facilitate knowledge-transfer. *Drug Discovery Today* **2009**, *14*, 845–850.

(6) All chemical structures appearing in this paper have been previously disclosed in one or more of the following publications: (a) Penning, T. D.; Talley, J. J.; Bertenshaw, S. R.; Carter, J. S.; Collins, P. W.; Docter, S.; Graneto, M. J.; Lee, L. F.; Malecha, J. W.; Miyashiro, J. M.; Rogers, R. S.; Rogier, D. J.; Yu, S. S.; Anderson, G. D.; Burton, E. G.; Cogburn, J. N.; Gregory, S. A.; Koboldt, C. M.; Perkins, W. E.; Seibert, K.; Veenhuizen, A. W.; Zhang, Y. Y.; Isakson, P. C. Synthesis and biological evaluation of the 1,5-diarylpyrazole class of cyclooxygenase-2 inhibitors: identification of 4-[5-(4-methylphenyl)-3-(trifluoromethyl)-1H-pyrazol-1-yl]benzenesulfonamide (SC-58635, Celecoxib). *J. Med. Chem.* **1997**, *40*, 1347–1365. (b) Talley, J. J.; Penning, T. D.; Collins, P. W.; Rogier, D. J., Jr.; Malecha, J. W.; Miyashiro, J. M.; Bertenshaw, S. R.; Khanna, I. K.; Graneto, M. J.; Rogers, R. S.; Carter, J. S. Substituted pyrazolyl benzenesulfonamides. U.S. Patent 5,466,823, Nov 14, 1995.

(7) All assay data appearing in the figures was randomly generated. All text appearing in the hypothesis, knowledge, notes, design criteria, measure of success and other such areas in the figures are for illustration purposes only and are not intended to reflect the properties of the molecules shown therein. The authors of this work played no role in the discovery or development of Celebrex or its analogues.

(8) The pivot table process described herein has been presented in preliminary form previously: Klug-McLeod, J. Data integration to facilitate project team decision making. Presented at (a) the Accelrys Science Forums 2009, Boston, MA, June 9, 2009, (b) the Massachusetts Biotechnology Council, Cambridge, MA, April 2, 2009, and (c) the 2009 Accelrys User Group Meeting 2009, February 19, 2009.

(9) Howe, W. J. An integrated desktop computing environment for medicinal and computational chemists. Presented at the American Chemical Society National Meeting, Philadelphia, PA, August 18, 2008; Abstract CINF 32.

(10) Pipeline Pilot, versions 6.1, 7.0, and 7.5; Accelrys, Inc.: San Diego, CA, 2006, 2008, and 2009, respectively.

(11) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.

(12) Spotfire DecisionSite, version 9.0; TIBCO Software Inc.: Palo Alto, CA, 2007.

(13) ISIS/Draw, version 2.4; Symyx Technologies Inc.: Sunnyvale, CA, 2001.

(14) Smith, G. F. Enabling HTS hit follow up via chemo informatics, file enrichment, and outsourcing. Presented at the Royal Society of Chemistry High Throughput Medicinal Chemistry II Symposium, May 9, 2006. http://www.mmsconferencing.com/pdf/htmc/g.smith.pdf (accessed March 19, 2009).

(15) A structure is considered “unique” if it represents a novel stereoisomer, tautomer, or ionization state of a previously registered compound. Mixtures of stereoisomers (typically racemic mixtures of enantiomers) are also considered distinct from the individual stereoisomers themselves. The fact that tautomeric forms are considered distinct is seen as a limitation and will be addressed in a future version of the VCP.

(16) Accord Chemistry Cartridge, version 7.0; Accelrys, Inc.: San Diego, CA, 2008.

(17) The current implementation attempts to match VCP and ADP molecules by comparison of the Pipeline Pilot CANONICAL SMILES representations of the parent fragments (non-salt forms). Unfortunately, this method may fail to match molecules with multiple chiral centers and/or multiple tautomeric forms, based on differences in how these molecules were drawn during the original registration process. Future development is planned to address these shortcomings.

(18) Cerep. http://www.cerep.fr/cerep/users/pages/ProductsServices/Industrialization.asp (accessed March 19, 2009).

(19) FileMaker Pro, version 9.0; FileMaker, Inc.: Santa Clara, CA, 2007.

(20) PCAT, Pfizer Compound Analysis Tool, version 3.5; Pfizer Global Research & Development: Cambridge, MA, 2009; PCAT is a tool for clustering, organizing, and visualizing molecules with their associated properties and biological activities.

(21) RGate, version 2.7.1.3.0; Pfizer Global Research & Development: Groton, CT, 2009.

CI9002443