

Searching Chemical Space with the Bayesian Idea Generator

Willem P. van Hoorn* and Andrew S. Bell

Department of Chemistry, Pfizer Global Research and Development, Sandwich Laboratories,
Sandwich, Kent CT13 9NJ, United Kingdom

Received February 25, 2009

The Pfizer Global Virtual Library (PGVL) is defined as a set compounds that could be synthesized using validated protocols and monomers. However, it is too large (10^{12} compounds) to search by brute-force methods for close analogues of a given input structure. In this paper the Bayesian Idea Generator is described which is based on a novel application of Bayesian statistics to narrow down the search space to a prioritized set of existing library arrays (the default is 16). For each of these libraries the 6 closest neighbors are retrieved from the existing compound file, resulting in a screenable hypothesis of 96 compounds. Using the Bayesian models for library space, the Pfizer file of singleton compounds has been mapped to library space and is optionally searched as well. The method is >99% accurate in retrieving known library provenance from an independent test set. The compounds retrieved strike a balance between similarity and diversity resulting in frequent scaffold hops. Four examples of how the Bayesian Idea Generator has been successfully used in drug discovery are provided. The methodology of the Bayesian Idea Generator can be used for any collection of compounds containing distinct clusters, and an example using compound vendor catalogues has been included.

INTRODUCTION

Like most companies in the pharmaceutical industry, Pfizer has in recent years invested heavily in combinatorial chemistry capacity.^{1–3} Using validated chemistry protocols, a collection of monomers compatible with chemistry, and a robotic synthesis and purification platform, a synthetic chemist can rapidly produce orders of magnitude more compounds than with traditional single compound synthesis.⁴ The initial focus of this effort was to dramatically expand the corporate screening file collection. The first generations of these combinatorial libraries were therefore predominately designed by chemical diversity. From the virtual library, the collection of compounds that could be made given the chemical protocol and monomers, only a small subset (a mixture of fully combinatorial and more sparse matrices) was synthesized. When any of these library compounds are identified as actives in a High-Throughput Screen (HTS), library chemistry allows for the rapid synthesis of close analogues from the remaining gaps in the virtual library. For most library protocols, only a tiny fraction of the virtual library has ever been synthesized, leaving enough scope to design targeted follow-on libraries. Indeed, it is estimated that the collection of protocols and monomers at Pfizer add up to a virtual library (PGVL) in the order of 10^{12} compounds.⁵ Since the chemistry protocol has been registered for all library compounds, a simple database look-up establishes which protocol should be used to follow up the HTS hits. The chemical space to look for close analogues equates to the virtual library of that protocol, which is typically in the order of 10^6 compounds and fully searchable.⁶ However, if hits are found which are not from a registered

library, the search space is the full Pfizer virtual library of 10^{12} compounds. The size of this chemical space is huge and growing. Even with massive parallel computing it is not a feasible option to search this space. If a computer grid would be available with 1000 CPUs and each of these CPUs would process 1000 compounds per second, a search through 10^{12} compounds would take more than 11 days, limiting the number of searches that can be performed in one year to at most 31.⁷

Bayesian statistics^{8,9} can be used to build single category activity models from large data sets like HTS screening results¹⁰ or multicategory models from collections of screening data.^{11–13} Other successful uses of Bayesian statistics have been a natural product-likeness score,¹⁴ reranking of protein–ligand docking results,¹⁵ identification of the most likely hit when compound mixtures have been screened,¹⁶ prediction of potential to covalently modify protein target in HTS campaigns,¹⁷ selectivity mining,¹⁸ and prioritizing compounds from a full HTS library based on fragment screening data.¹⁹

In this paper the Bayesian Idea Generator is described. At its heart is a multicategory Bayesian model of the Pfizer combinatorial library collection, based on the observation that compounds from the same library series have more in common e.g. template or bonds to disconnect than those from different series. We show that this novel application of Bayesian statistics leads to successful prediction of the library-provenance for a given compound without the requirement of (pre-coded) chemical knowledge.^{5,7} The Bayesian short-cut reduces the search space to a few protocols which can then be searched explicitly within minutes. The result of this search is a list of protocols illustrated with the most similar combinatorial compounds that have already been made using each protocol. This set of compounds can be readily screened,

* Corresponding author phone: +44-1304-648470; e-mail: willem.van.hoorn@pfizer.com.

allowing for quick evaluation of the potential of each individual virtual library either without chemical synthesis or by rapid synthesis using monomers available in our collection (~25,000). In addition, singleton chemical space can be mapped onto the library chemical space, allowing fast searching of this chemical space as well. The diversity of the answers leads in many cases to a scaffold jump, i.e. the hits can be considered a different chemical series, which can be beneficial for metabolic profile or patentability. Once an initial hit is found using this analysis, the methodology can be reapplied to discover further generations of leads by further file screening or directed synthesis. The usefulness of the Bayesian Idea Generator is illustrated with examples from past Pfizer projects.

METHODS

All calculations have been performed using Scitegic Pipeline Pilot.²⁰ The work described in this paper has been implemented in multiple stages using multiple versions of Pipeline Pilot starting from 4.5.2 SP1.

Bayesian Learning. Bayesian learning has been applied as implemented in Pipeline Pilot.⁸ The training set of a single-category Bayesian model must contain a subset of G “good” molecules of interest, in the case of activity modeling this could be the hits as defined by percent inhibition greater than a user-defined threshold. The remainder of the training set constitutes B “bad” molecules. The baseline probability of a random molecule from this set being active, $P(\text{Baseline}) = G/(G+B)$. All fingerprint features (~substructures) are calculated for all molecules in the training set. For each feature x , it is determined how often it is contained in a “good” (g_x) and how often in a “bad” molecule (b_x). This yields the uncorrected estimate of activity for x , $P(\text{Active}|x) = g_x/(g_x+b_x)$. This estimate becomes unreliable if the number of observations g_x+b_x becomes small. Therefore, L additional samples are added each with $P(\text{Baseline})$ chance of being “good”. If $L = 1/P(\text{Baseline})$ this is the Laplacian correction, which yields a corrected estimate of activity $P_{\text{corr}}(\text{Active}|x) = (g_x + L \cdot P(\text{Baseline})) / (g_x + b_x + L)$. For small g_x+b_x , this defaults to $P(\text{Baseline})$ which is the expected value for a random fingerprint feature x . The relative contribution for feature x compared to the baseline estimate is $P_{\text{final}}(\text{Active}|x) = P_{\text{corr}}(\text{Active}|x)/P(\text{Baseline})$. For each feature x present in the training set, the $\log P_{\text{final}}(\text{Active}|x)$ is by default only stored if the absolute value exceeds 0.05. This greatly reduces the size of the model by removing features that do not distinguish significantly between “good” and “bad”. For fingerprint features overrepresented in “good” molecules, $\log P_{\text{final}} > 0$. Features underrepresented in good molecules have $\log P_{\text{final}} < 0$. The Bayesian score of a molecule is calculated by adding all the $\log P_{\text{final}}$ values of the fingerprint features present in it. Multicategory Bayesian modeling is used for a training set with more than one subset of “good” molecules. For each category C , the equivalent of a separate single-category model is derived where all molecules C are “good”, and the remainder of the training set constitutes the “bad” molecules. In effect, for each fingerprint x , C separate $P_{\text{final}}(\text{Active}, C|x)$ values are calculated and stored. One of the weaknesses of the normalized probability scores is that they are not truly comparable across different models. However, building a model automatically performs a fast

leave-one-out cross-validation, and results of that process can be used to create absolute estimators that are comparable: “Enrichment” and “EstPGood”. The Enrichment value is interpreted as follows: if only molecules with relative Bayesian scores close to this compound’s score are tested, how much better (or worse) would the hit rate be compared to simply testing random molecules? The Enrichment score is biased toward categories with a small number of members. The EstPGood score is an estimate of the probability that a compound would be in this category if only molecules with relative Bayesian scores close to this compound’s score were tested. In contrast to the Enrichment score, the EstPGood score is biased toward categories with large numbers of members.²⁰

Modeling Pfizer Data. A multicategory Bayesian model has been derived from the Pfizer combinatorial chemistry compound file which contains ~1.9 M compounds derived from 1805 libraries, together with a subset of 12,500 compounds that represent clean singleton chemical space.²¹ The model contains therefore 1806 categories. A random 50% was used as the training set. For each category (library), a separate file was stored containing the compound names, batch codes for ordering, and chemical structures represented as canonical SMILES.^{22,23} Two models were derived based on the ECFP_6 and FCFP_6 fingerprint descriptors,¹⁰ respectively.

The Pfizer liquid screening file of ~3 M compounds was processed as follows: First all compounds that were part of the combinatorial chemistry file were removed, followed by filtering by physical properties ($150 \leq \text{Mw} \leq 750$ and $\text{AlogP} \leq 7$) and the presence of any of 178 reactive chemical groups like aldehydes, peroxides, and the like. The resulting singleton file contained 1087447 compounds. These singleton compounds were mapped to library chemical space by assigning each to the highest scoring library from the multicategory Bayesian model based on ECFP_6 fingerprints. For each (predicted) library, a separate file was stored containing the compound names, batch codes for ordering, and chemical structures represented as canonical SMILES.^{22,23}

Implementation of Bayesian Idea Generator. The multicategory Bayesian model is used to make 16 predictions which libraries a given probe molecule could have come from. For each library, a Tanimoto nearest neighbor search is performed yielding the 6 closest compounds that have already been made. The same Tanimoto search is also performed within singleton compounds mapped to the same library. The Tanimoto searches are performed using the ECFP_4 or FCFP_4 fingerprints depending whether the Bayesian model to make the library predictions was based on the ECFP_6 or FCFP_6 fingerprints. The Bayesian Idea Generator is published as a Web service and users can enter structures as Pfizer internal structure code, SMILES string, or mol file or use an interactive interface to IsisDraw to sketch structures. The number of library predictions (default 16) and the number of neighbors retrieved in the Tanimoto search (default 6) is user-tunable, as is the fingerprint flavor of the model. The final output consists of a pdf report generated using the Pipeline Pilot reporting collection and a csv file which contains all information needed to order compounds for biological screening. For one input probe this will yield 96 compounds (Figure 1), which is the number of compounds that fit on a standard biological screening plate.

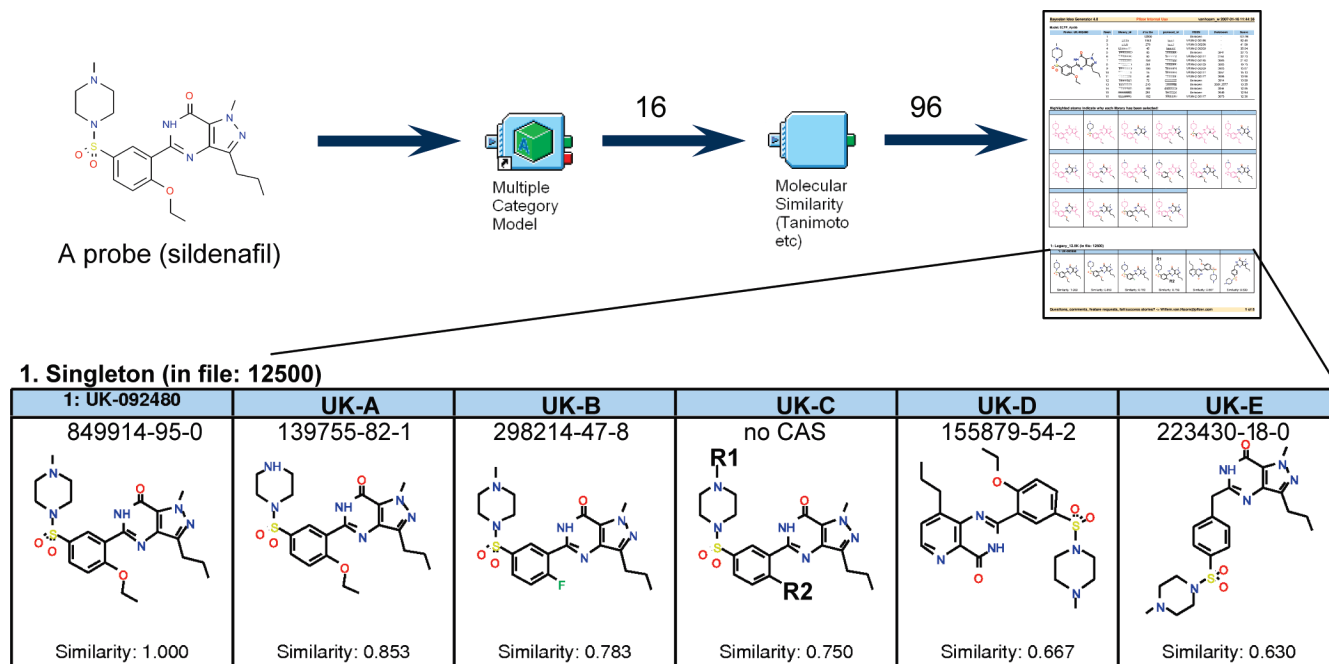


Figure 1. Workflow of Bayesian Idea Generator. For a given probe, 16 Bayesian predictions of library ID are made. For each of the libraries the 6 most similar existing compounds are identified with a Tanimoto nearest neighbor search yielding a total of 96 compounds. The outcome is summarized in a pdf report. For sildenafil (Viagra), the singleton library is the highest ranking library and the 6 nearest neighbors are shown (which includes sildenafil). For the purposes of this example, Pfizer internal structure codes have been replaced with CAS numbers.

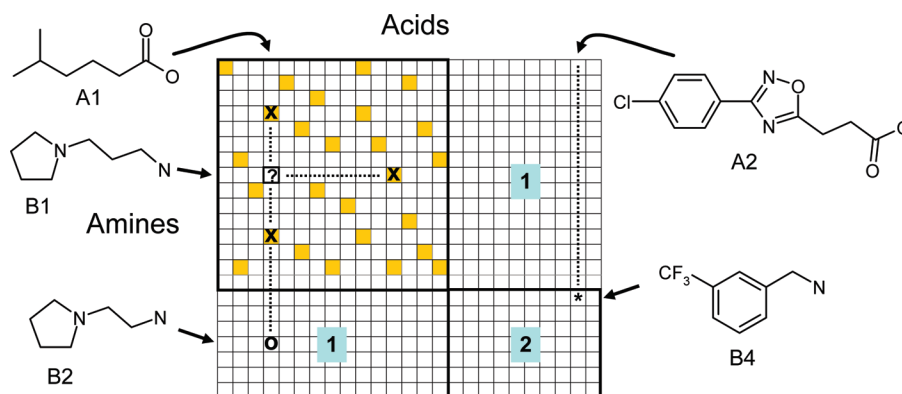


Figure 2. Coverage of the Bayesian model for a hypothetical amide library derived from amines and acid monomers. The model is built from the compounds that have been synthesized (yellow squares). This chemical space is fully covered by the Bayesian model since nearly all fingerprint features of any virtual compound (square marked with "?") are shared by at least one compound from the training set (squares marked with "X"). Virtual products in areas 1 share at least one monomer with a compound from the training set. In the case of compound "O", the new monomer B2 is very close to previously used B1. Only compounds from area 2 can be considered outside the scope of the model, but only because they have few fingerprint features in common with the existing products as shown for compound "*" where monomers A2 and B4 are unlike previously used monomers.

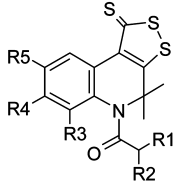
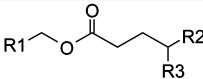
By all likelihood, these 96 compounds are not the closest analogues that could be made using the library protocol because the real library tends to be much smaller than the virtual library for any given protocol (Figure 2). However, they are immediately available for screening. If any of the compounds from one of the 16 libraries show the desired activity, even a weak one, there is a strong rationale for designing and synthesizing a targeted library exploiting the full virtual library. The default output of the Bayesian Idea Generator therefore consists of 16 screenable hypotheses. In practice the existing compounds are often considered too far away from the pharmacophore the chemist is working on, and therefore a few close target compounds are made as singletons. When successful, these are followed up with a full library. The average search takes just over 5 min on

standard server hardware. Variations of the runtimes are mainly due to different sizes of the libraries that have to be searched by the Tanimoto similarity searches. The Bayesian predictions have a throughput of multiple compounds per second, but loading the model in memory takes about 1 min. Generation of the pdf report takes about 3 min.

RESULTS AND DISCUSSION

Bayesian Models and Nearest Neighbor Searches of Compound Clusters. The aim of this research is to build a Bayesian model that predicts which Pfizer library protocol most likely yields similar compounds to a given input structure. As a proof of concept, a similar multicategory Bayesian model has been derived from publicly available

Table 1. Two Example Clusters Found in the Vendor Catalogue File^a

Cluster	Core	Compounds in training set	Compounds in test set	Predicted in top 5 by Bayesian model
10		119	112	112
18		6	7	4

^a Cluster 10, with a distinct common core, is similar to a template-based combinatorial library, while cluster 18 contains compounds that could have been made by ester coupling. The larger size of the cluster as well as the lower diversity from the common core leads to better predictions for cluster 10 compared to cluster 18 as shown by the number of correct top 5 predictions of the test set.

compounds in the Asinex and Maybridge vendor collections. Together these collections contain 186967 unique compounds as determined by canonical smiles. The vendor compounds were clustered using default Pipeline Pilot clustering, and only clusters with 10 or more compounds were retained. This results in 4370 clusters containing a total of 166790 compounds, which is comparable to the number of libraries in the Pfizer compound file (but an order of magnitude smaller with respect of number of compounds). Visual inspection of a subset of the clusters showed that they are a mixture of (template-based) library-like clusters, with a distinct core ring system coupled to small R-groups, and more loosely defined groups of molecules which have some ring or functional group in common not unlike the results of an ester coupling library (see Table 1). A random 50% of the compounds (83482) was used to build a multicategory Bayesian model using ECFP_6 descriptors and cluster number as category property. This model produces a ranked list of clusters a compound most likely belongs to. The top 5 most likely clusters were calculated for the remaining 83308 compounds. For 49938 of them, the highest ranking cluster number was the correct one, and for a further 20691 compounds the correct cluster number was found within the top 5 predictions. The throughput of this model is between 5 and 10 compounds per second. This example shows that the principle of building a multicategory Bayesian model of clusters of similar compounds can be used to predict which cluster a compound belongs to.

The same training and test sets of compounds were used to investigate if a Tanimoto nearest neighbor search could be used to predict cluster membership. For each compound from the test set, the 100 closest neighbors from the training set were calculated by ECFP_4 similarity. The top 5 cluster numbers found among the nearest neighbors were retained. This calculation is faster, but a nonsignificant worse prediction rate was observed compared to the Bayesian predictions.²⁴ The highest ranked prediction was correct for 50024 compounds, and for an additional 20522 compounds the correct cluster number was within the top 5 predictions. All 70546 correct predictions were from clusters with 6 or more members in the training set. In about 20% of the cases (14741/70546), fewer than 6 members of the

cluster were found in the top 100 nearest neighbors. The aim of research was to deliver to the user a balanced set of compounds with the same number of nearest neighbors for each cluster. This could be part accomplished by generating a larger neighbor list than the top 100, but there will still not be a guarantee that at least 6 neighbors are found for each cluster and it will lead to a drop in throughput. Performing a second nearest neighbor search will be necessary for each cluster in the top 5 predictions where fewer than 6 neighbors have been identified. When the training set is the Pfizer combinatorial chemistry file, a similarity search has to be performed against a growing data set with currently stands in the millions which makes this search slower. The Bayesian model for this data set also increases in size and deriving it most certainly takes longer, but predicting for a single compound is still less than a second. For larger data sets the similarity search is therefore expected to be of similar or slower speed than the Bayesian predictions. In summary, the work described in the paper could have been implemented with similarity search, but the latter will not scale as well for large data sets. The protocol to perform the above calculations is available as Supporting Information.

Building and Validation of Bayesian Models of the Pfizer Library File. A set of 6 multicategory Bayesian models were built from 50% of the Pfizer combinatorial chemistry file of 1806 libraries using the library identifier as category descriptor. The aim of these models is to predict which library protocol is most likely to yield compounds similar to a given input molecule. To test the performance of the models, the remaining 50% of the compounds (just under 1 million) was used as a test set. A random 1% (9452 compounds) was selected from these, and the predicted top 5 libraries (out of a possible 1806) were compared to the real library each compound originated from (see Table 2). All six Bayesian models find the correct library identifier within the top 5 predictions with >97% accuracy. The Enrichment score, which biases results toward small-sized libraries in the training set, does perform worst of the methods regardless which fingerprint (ECFP/FCFP) was utilized. The EstPGood scores, which have a bias toward larger libraries, were better than the Normalized_Probability

Table 2. Recall Rates of Parent Library ID of Random 9452 Compounds by 6 Different Bayesian Models

model	found in top 1	found in top 5	not in top 5
ECFP_Normalized_Probability	9068, 96%	9411, 99.6%	41, 0.4%
ECFP_Enrichment	5692, 60%	9247, 97.8%	205, 2.2%
ECFP_EstPGood	8372, 89%	9439, 99.9%	13, 0.1%
FCFP_Normalized_Probability	8920, 94%	9367, 99.1%	85, 0.9%
FCFP_Enrichment	6093, 64%	9344, 98.9%	108, 1.1%
FCFP_EstPGood	8547, 90%	9441, 99.9%	11, 0.1%

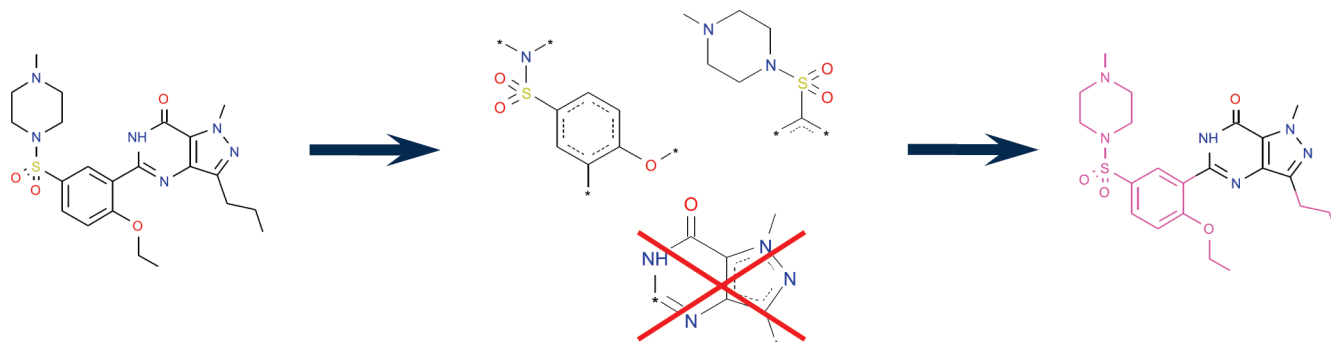
scores when looking at the top 5 results but worse when looking at the top ranking results. The differences between ECFP and FCFP fingerprint descriptors are minor as well, with the ECFP_Normalized_Probability scores outperforming all other scores when looking at the top 1. Since the differences between the models are minor, and the ECFP_4 fingerprints have been shown to be superior in similarity searches,²⁵ it was decided to use the ECFP_Normalized_Probability score as default.

Coverage of the Models into Virtual Compound Space. In the previous section it was shown that the Bayesian models are predictive for compounds that were made by library chemistry. In Figure 2 it is shown how a model that is derived from O(6) real compounds has nearly full coverage of O(12) virtual compounds. Since most of the original Pfizer libraries were designed by diversity, the synthesized real library contains most of the monomers of the virtual library or at least the most descriptive subset of monomers. All virtual compounds interpolated in the real compound space are therefore covered by the model; they do contain few fingerprint features that are not also present in at least one compound from the real library training set. This is illustrated by product “?” in Figure 2. Monomers that have been added to the monomer pool after the real library was synthesized are not covered by the model. However, a monomer might be absent, but it is only truly not covered if it does not share fingerprint features with any of the other monomers. For instance, monomer 1-pyrrolidinopropylamine B1 has been used before and is included in the model, but the new monomer 1-pyrrolidinoethylamine B2 does not contain fingerprint features not already present in it and is therefore covered well by the model for this library. Likewise, many monomers in the Pfizer file resemble each other. Typical examples are different lengths and branching of alkane chains and different halogenation pattern around aromatic rings. A significant proportion of the Pfizer libraries are based on a single template contained in all products. This partially

covered chemical space is indicated by “1” in Figure 2. Only compounds from area “2” are not covered by the model, but only if both monomers are structurally different from all other monomers as illustrated by the aromatic monomers A2 and B4 in Figure 2. This is a rare occasion, illustrated by the low number of wrongly predicted compounds in Table 2. The 41 mispredicted compounds from the “ECFP_Normalized_Probability” model were mostly from large amide coupling libraries where a large amine monomer had been used only once, and no template was present, similar to the vendor compounds from Cluster 18 in Table 1. The absence in the model of monomers that have consistently failed to yield products can be considered advantageous, for instance in the case when a compound can be created from a monomer X by two different library synthesis protocols. If monomer X has failed to yield products in one library, the Bayesian model will favor the other library where the monomer has been successful.

Visualization of Scores. The Bayesian score of a compound is the sum of the normalized probabilities of the fingerprint features, but it is just a number; most chemists are interested in what part of the molecule contributes to the score. The sum of the positive normalized probabilities of the fingerprint features covering a given atom can be calculated (Figure 3).²⁶ The parts of the molecule that contribute most to the Bayesian score can be visualized by highlighting atoms where the sum of the normalized probabilities exceeds a preset threshold (currently set to 1.0). At the top of the pdf report (Figure 1) there is a table where the input molecule is colored by the each of 16 library models. This offers a snapshot overview as to what substructures of the input molecule are also present in each library.

Inclusion of Singleton Chemical Space. The Bayesian model derived from library compounds is only predictive for library chemical space, but the Pfizer screening file also contains many singleton (individually synthesized) compounds. To capture this singleton chemical space, the multicategory Bayesian models have been supplemented with a “Singleton” library. This set of 12,500 compounds was not derived from combinatorial chemistry but is a representative subset of ~86k ultraclean singleton compounds.²¹ This set of compounds has been thoroughly filtered by physical properties and presence of undesirable (reactive) chemical groups. As a result of this filtering, the singleton compounds cannot easily be distinguished from library compounds by properties like molecular weight or presence of reactive

**Figure 3.** Coloring molecules by contribution to Bayesian score. Only fingerprint features that contribute positively (positive normalized probability) are retained. If the sum of the normalized probabilities exceeds 1.0 an atom is colored magenta.

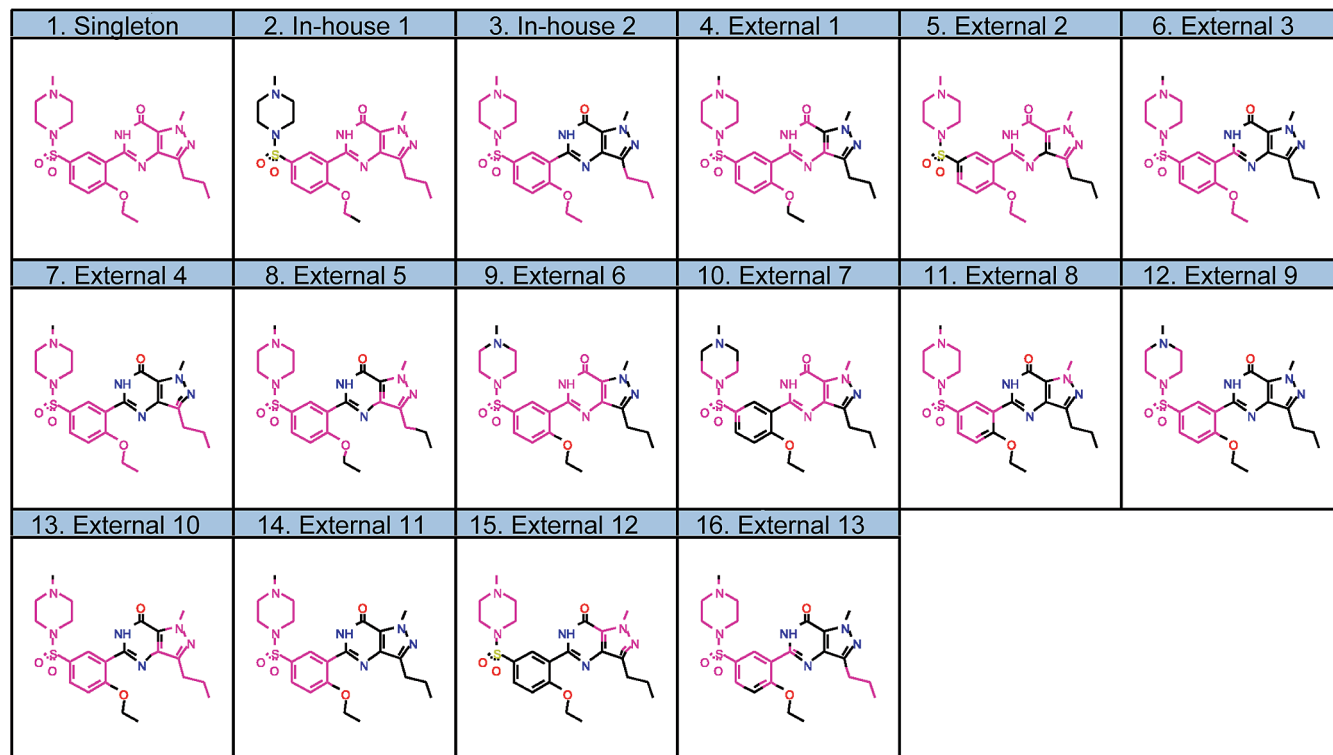


Figure 4. Bayesian score highlights for sildenafil. The pyrazolo[4,3-d]pyrimidine ring of sildenafil is only represented in the singleton library and one internal library.

functional groups that are incompatible with library protocol conditions since most libraries were filtered similarly during design. If there is a difference between the singleton and the library compounds it has to be the presence of a fingerprint feature (substructure) that is underrepresented in the library file. If the Bayesian score representing the singleton library is high for a given compound, it indicates that the compound is outside combinatorial chemical space. This is illustrated in Figure 4, where the pyrazolo[4,3-d]-pyrimidine ring of sildenafil is highlighted by the singleton model and one internal library model, but in none of the other library models. The singleton model can be used to identify substructures whose inclusion as a library template would increase the chemical space coverage of library chemistry.

The library models have been employed to map the Pfizer singleton screening file onto library file space by calculating the most likely library each compound could have come from. The predicted distribution of the singleton compounds over the libraries is shown in Figure 5. As was expected, the largest predicted library for the singleton compounds is the Singleton library (8%). From the 1806 libraries, all but 4 had at least one singleton compound assigned to it. The sizes of these 4 unmapped libraries are 1, 1, 2, and 11. Assuming that the libraries attempted for synthesis was significantly larger they can be considered failed libraries, and it is reassuring that no singleton compounds were mapped to them. However, there is no obvious correlation between size of the library and number of singleton compounds predicted to be from that library (data not shown). The mapped singletons are searched by similarity alongside the similarity searches of the “real” libraries. The results of these searches are an additional set of 96 compounds that can be submitted to a biological assay.

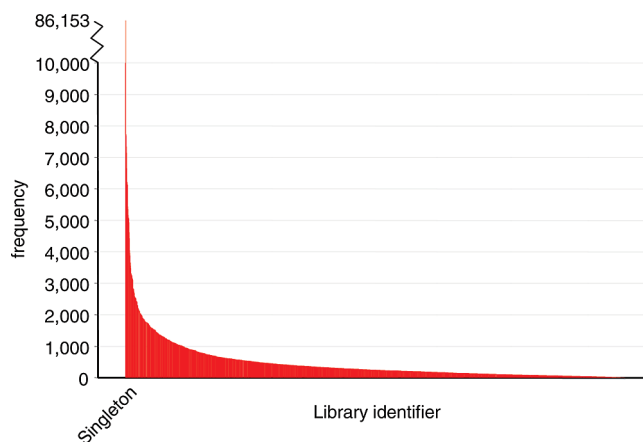


Figure 5. Mapping of 1087447 singleton compounds to 1806 libraries by Bayesian predictions. The largest predicted library is the Singleton library (86153). Only 4 libraries have no singleton compound assigned to it.

Chemotype Jumping. The ability of an in-silico search method to identify a novel chemical series with the same activity as a given series is known as scaffold or chemotype hopping.^{5,27–36} Chemotype jumping is desirable for a variety of reasons: the novel series might be easier to work on synthetically, might have fewer or at least different pharmacological risks, and/or might be free from intellectual property constraints. The Bayesian Idea Generator offers more chance of chemotype jumps compared to a 2D Tanimoto similarity search via three mechanisms:

First, the default output of 16 times 6 compounds is diverse since only 6 compounds have been selected from each library. Suppose the Pfizer file contains a single library that is not very diverse and similar to the input probe, the highest ranking output of the Tanimoto search will be dominated by compounds from that library. The same argument applies

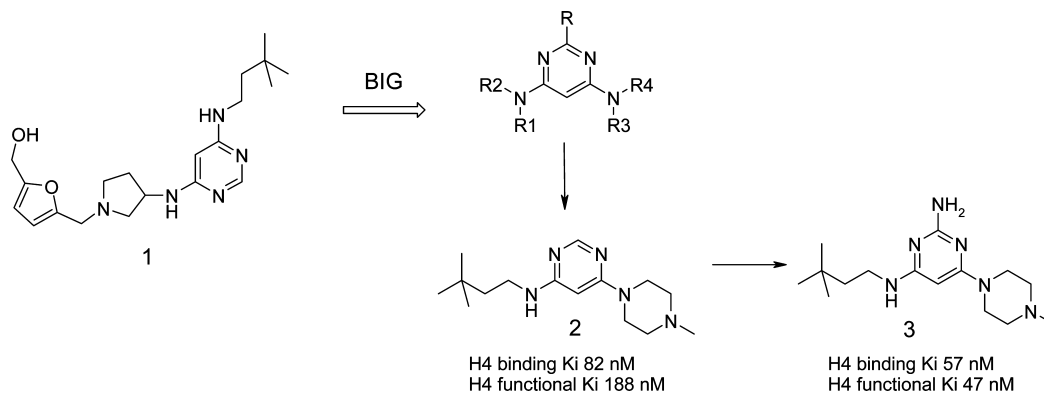


Figure 6. H4 antagonists. Application of the Bayesian Idea Generator on weak hit **1** identified a library of 4,6-diaminopyrimidines including compound **2**. Subsequent optimization led to more potent compounds like **3**.

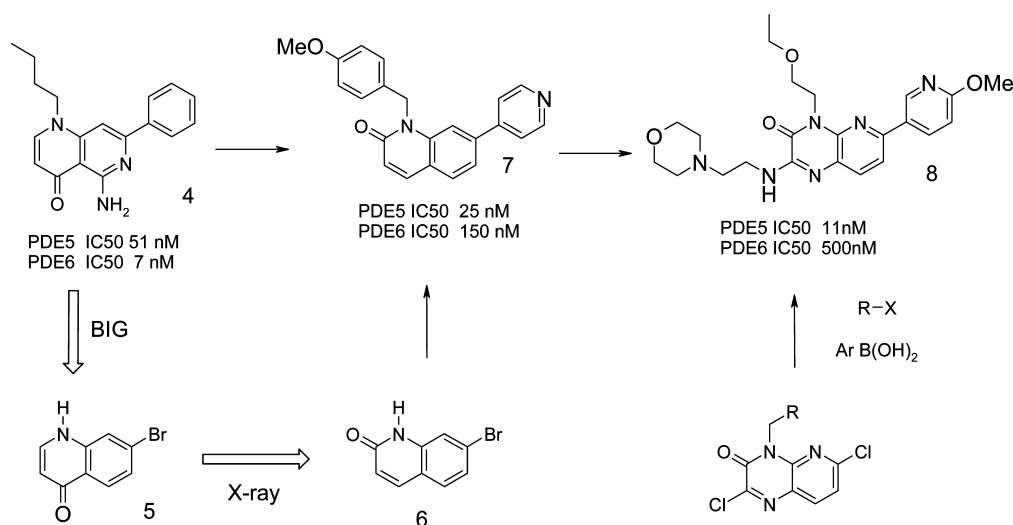


Figure 7. PDE5 inhibitors. Hit **4** was followed up by two libraries based on templates **5** and **6**. The chemistry was found by applying the Bayesian Idea Generator; the jump for 4-one quinoline **5** to 2-one quinoline **6** was conceived through studying the crystal structure of **4** in the PDE5 enzyme.

to close-in series of singleton compounds. Selecting only 6 compounds from one library prevents this domination by one series.

Second, the fingerprint features in a Tanimoto search are all equally important. This is not the case for Bayesian predictions: the normalized probability weights are large for fingerprint features that make one library distinct from other libraries. These high scoring features tend to be associated with library templates which are often heterocyclic ring systems. The high scoring features are less often simple functionalities like chlorinated phenyl, alkyl, etc. which are found in many libraries since they are part of the R-group monomers that have been used repeatedly. If the probe molecule contains a ring system that is similar to a template ring system from the combinatorial libraries, this will yield a high Bayesian score for the library even if the probe molecule contains other functionalities that are not found in this library. This is illustrated in Figure 4, where the libraries were found since they contain an aryl-piperazine-sulfonamide template or a ring system that is similar (in case of the first two identical) to the pyrazolo[4,3-d]pyrimidine of sildenafil. The perception of a series in the medicinal chemist's mind is more associated with template ring systems than R-groups; the focus of Bayesian scores therefore coincides with what chemists find important. The 96 compounds found by the Bayesian Idea Generator can be compared to the top 96 most

similar compounds to sildenafil identified by an ECFP₄ Tanimoto nearest neighbor search against the same compound file. The Tanimoto results are much less diverse: 1 compound from "In-house 2", 47 compounds from "In-house 1", and 48 compounds from the Singleton library. The latter can be explained by the large history of Pfizer working on PDE inhibitors. Of the 96 compounds found by the Bayesian Idea Generator, 13 were also in the top 96 by Tanimoto, a further 49 within the top 1000, 33 in the top 10000, and 1 ranked beyond 10000. The compound file that was searched contains ~2 million compounds, the compounds found by the Bayesian Idea Generator are still within the top 0.5% most similar compounds. The similarities of the compounds are high enough to have a reasonable chance of activity^{32,37} but low enough to provide novelty.

And finally, the nearest neighbor searches of the mapped singleton compounds provide an extra source of novelty on top of the mechanisms outlined above by a mechanism similar to the indirect similarity method described by Wale et al.³⁵ In this method, two compounds are ranked as similar if their respective classic nearest neighbor lists show a large degree of overlap, even if the two compounds themselves have low similarity. In the case of the mapped singleton compounds a similar mapping by association occurs: the probe that is mapped to library X is associated with other compounds that have been mapped in the same library.

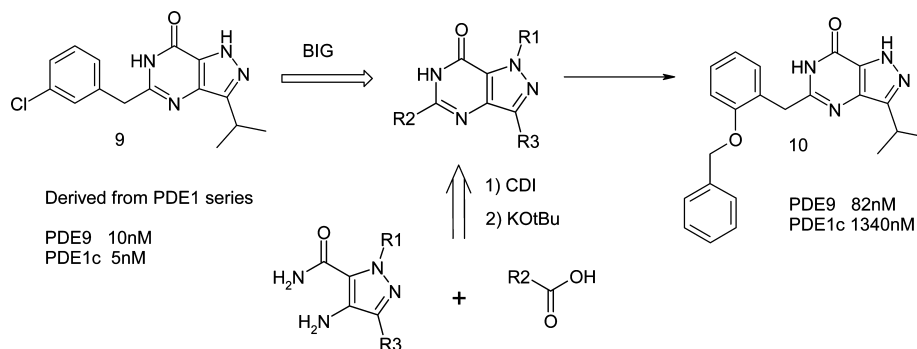


Figure 8. Introducing selectivity over PDE1c from nonselective hit 7.

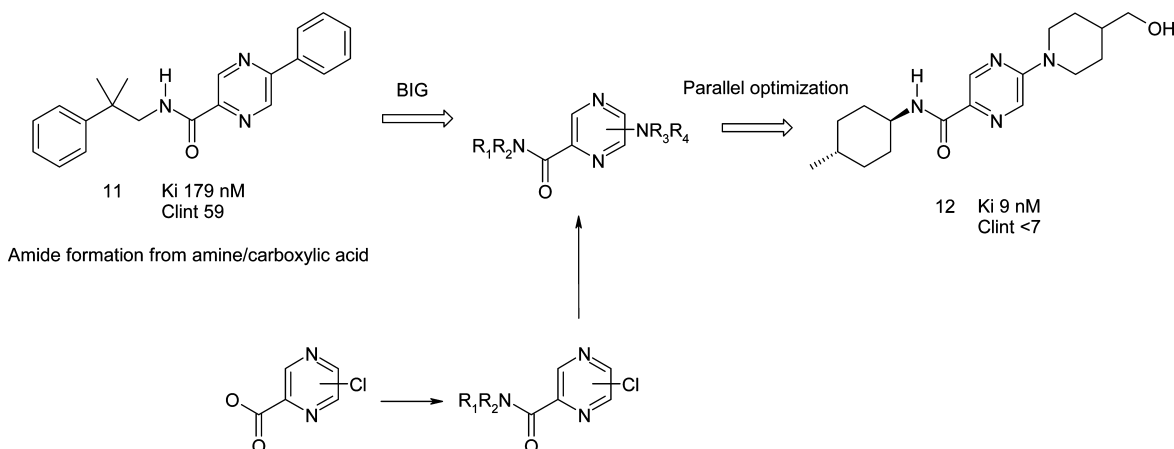


Figure 9. The mGluR1 antagonist **11** was found by optimizing an initial amide hit via replacement of the acid or amine portions during amide formation. A Bayesian Idea Generator search uncovered an alternative route which allowed for easier introduction of R-groups and rapid elaboration to lead **12**.

Project Examples.^{38–42} The Bayesian Idea Generator has been used successfully in multiple projects, and some of the results are outlined here. In a H4 antagonist project, screening of a subset of the Pfizer compound file identified **1** as a weakly potent H4 antagonist (Figure 6). Due to the low ligand efficiency of this hit compared with other comparable agents, we sought to reduce the molecular weight of the hit without compromising potency.⁴³ Application of the Bayesian Idea Generator identified a library of 4,6-diaminopyrimidines, which had not been screened. Compound **2** was found to be more active and with a lower molecular weight than the initial hit, significantly improving ligand efficiency compared to compound **1**. The identification of a validated library protocol by the Bayesian Idea Generator allowed subsequent follow-up utilized a range of amines and alternative 4,6-dichloropyridine templates, yielding more potent analogues such as **3**. In this case the library of 4,6-diaminopyrimidines was identified because it shares the pyrimidine template with target molecule **1**. This would also have been obvious disconnection points for retrosynthesis; however, replacement of the amino-pyrrolidine group with a piperazine group is less obvious, and the similarity of compounds **1** and **2** is only 0.39.

The second example is from a PDE5 inhibitor project. As part of our effort to discover a novel series of long-acting PDE5 inhibitors, a high throughput screen of our compound file identified compound **4** (Figure 7). The compound was a singleton hit obtained from a commercial source, but at the time there was no literature route available to the compound or its analogues.⁴⁴ Fortunately, sufficient of the original dry sample remained to obtain a cocrystal structure with PDE5,

which showed that compound **4** had achieved a novel binding mode without any of the hydrogen-bonding interactions made by other series of inhibitors. The Bayesian Idea Generator suggested that it would be possible to synthesize related quinolinones and their aza-substituted analogues using a two-step procedure from known 1H-7-bromoquinolin-4-one **5**. Based on the cocrystal structure of **4**, we also included the corresponding 1H-7-bromoquinolin-2-one **6** as a template in our initial follow-up library, which identified the novel, potent PDE5 inhibitor **7**. Subsequent rounds of synthesis investigated alternative templates featuring additional heteroatoms in the bicyclic skeleton and, guided by structural information, a site for the introduction of an additional substituent leading to lead compound **8**.

The third example is from a PDE9 inhibitor project. As part of our search for selective inhibitors from recently discovered cyclic nucleotide phosphodiesterase inhibitors, we cross-screened a subset of inhibitors of the well characterized PDE isozymes. In this example, we identified 8H-3-isopropyl-7-(3-chlorophenylmethyl)pyrimidin-4-one **9** as a nonselective inhibitor of both PDE1c and PDE9. Application of the Bayesian Idea Generator identified a parallel synthesis protocol based on a two-step amide bond formation and cyclization of 2-amino-heterocyclic carboxamides with carboxylic acids. A subsequent library chemistry follow-up of the initial screening hit identified the 2-position of the phenylmethyl as a region of space where substitution was tolerated by PDE9 but not by PDE1, resulting in lead compound **10**.

The final example is from an mGluR1 antagonist project. After targeted screening of the Pfizer file an initial amide

hit was optimized by using a simple amide bond forming protocol, resulting in lead compound **11** (Figure 9). This hit is less potent but moderately more metabolically stable compared to the initial hits (human liver microsomal metabolic stability, Clint, expressed as $\mu\text{L}/\text{min}/\text{mg}$). A search by the Bayesian Idea Generator located a previously synthesized array of ~ 6000 compounds made by amide coupling followed by amine displacement of the chloropyrazine core. The full array was screened and yielded moderately potent hits. In the new parallel chemistry protocol all R groups in the final product are introduced from amine monomers. The double use of the amine monomers significantly increases the design space compared to the previous protocol. Exploitation of this increased design space by parallel chemistry lead to the discovery of the potent and stable lead molecule **12**.

CONCLUSIONS

The Bayesian Idea Generator offers a fast and accurate method to identify which library protocol was used to synthesize a given input molecule by applying Bayesian statistics. This solves the issue of searching the vast chemical space of the Pfizer virtual library in a manageable time. The output of the Bayesian Idea Generator contains the most probable library protocols listed together with example molecules from the Pfizer collection. This set of answer molecules shows the right balance between similarity and diversity for lead hopping: close enough to the (presumably active) input molecule to be active, diverse enough to be novel. Inclusion of a model for singleton space provides a measure whether the input molecule is within or outside library chemical space. Mapping of the singleton file to the library file has enabled searching the singleton file with the same technology. No chemical knowledge of how to make compounds has been explicitly encoded in the Bayesian Idea Generator, it is therefore completely unbiased, and new library protocol can be added by simply rederiving the Bayesian models. The output consists of compounds that have been synthesized before, and in contrast to many *de novo* design tools the chemical do-ability of these is per definition high. Since samples of many of these compounds are still available, the applicability of a compounds derived from a library can be tested before synthesizing new compounds. The method described here is generally applicable to compound collections with subcategories of molecules that have features in common like activity classes but also compounds claimed in a patent.

ACKNOWLEDGMENT

We thank Pfizer colleagues Bruce Lefker, John Kath, Dafydd Owen, Dan Kung, Graham Smith, Jens Loesel, and Kevin Dack for their willingness to use untested software, challenges and useful suggestions. David Rogers (Scitegic/Accelrys) is thanked for implementing and significantly increasing the performance of multicategory Bayesian learning.

Supporting Information Available: Pipeline Pilot protocol to perform the calculations described in the section "Bayesian models and nearest neighbor searches of compound clusters". This material is available free of charge via the Internet at <http://pubs.acs.org>.

REFERENCES AND NOTES

- (1) Everett, J.; Gardner, M.; Pullen, F.; Smith, G. F.; Snarey, M.; Terrett, N. The application of non-combinatorial chemistry to lead discovery. *Drug Discovery Today* **2001**, 6, 779–785.
- (2) Milne, G. M. Pharmaceutical Productivity - The Imperative of New Paradigms. In *Annual Reports in Medicinal Chemistry*; Doherty, A. M., Ed.; 2003; Vol. 38, pp 383–396.
- (3) Smith, G. F. Enabling HTS Hit follow up via Chemoinformatics, File-Enrichment, and Outsourcing. Presented at High Throughput Medicinal Chemistry II [Online], London, 2006. MMS Conferencing & Events Ltd <http://www.mmsconferencing.com/pdf/htm/cg.smith.pdf> (accessed June 17, 2009).
- (4) Kennedy, J. P.; Williams, L.; Bridges, T. M.; Daniels, R. N.; Weaver, D.; Lindsley, C. W. Application of Combinatorial Chemistry Science on Modern Drug Discovery. *J. Comb. Chem.* **2008**, 10, 345–354.
- (5) Boehm, M.; Wu, T.-Y.; Claussen, H.; Lemmen, C. Similarity searching and scaffold hopping in synthetically accessible combinatorial chemistry spaces. *J. Med. Chem.* **2008**, 51, 2468–2480.
- (6) Van Hoorn, W. P. Library Design by Bayesian Modelling. Presented at 2005 Pipeline Pilot User Group Meeting [Online], San Diego, 2005. Pipeline Pilot Presentation Archive. http://media.accelrys.com/scitegic/protected/presentationArchive/UGM2005/CustomerPDFs/Pfizer_LibraryDesign_WillemvonHoorn.pdf (accessed June 17, 2009).
- (7) Nikitin, S.; Zaitseva, N.; Demina, O.; Solovieva, V.; Mazin, E.; Mikhalev, S.; Smolov, M.; Rubinov, A.; Vlasov, P.; Lepikhin, D. 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) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of Kinase Inhibitors Using a Bayesian Model. *J. Med. Chem.* **2004**, 47, 4463–4470.
- (9) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naive Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, 46, 1945–1956.
- (10) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screening* **2005**, 10, 682–686.
- (11) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naive Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 170–178.
- (12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, 24, 805–815.
- (13) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on ChemoGenomics Databases. *J. Chem. Inf. Model* **2006**, 46, 1124–1133.
- (14) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, 48, 68–74.
- (15) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results. *J. Med. Chem.* **2004**, 47, 2743–2749.
- (16) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high-throughput screening data using a naive Bayes classifier. *J. Biomol. Screening* **2004**, 9, 32–36.
- (17) Metz, J.; Huth, J.; Hajduk, P. Enhancement of chemical rules for predicting compound reactivity towards protein thiol groups. *J. Comput.-Aided Mol. Des.* **2007**, 21, 139–144.
- (18) Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for computer-aided chemical biology. Part 3: analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug Des.* **2008**, 71, 518–28.
- (19) Crisman, T. J.; Bender, A.; Milik, M.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Fejzo, J.; Hommel, U.; Davies, J. W.; Glick, M. "Virtual Fragment Linking": An Approach To Identify Potent Binders from Low Affinity Fragment Hits. *J. Med. Chem.* **2008**, 51, 2481–2491.
- (20) Pipeline Pilot, version 4.5.2SP1; Accelrys: San Diego, 2005.
- (21) Yeap, S. K.; Walley, R. J.; Snarey, M.; Van Hoorn, W. P.; Mason, J. S. Designing Compound Subsets: Comparison of Random and Rational Approaches Using Statistical Simulation. *J. Chem. Inf. Model.* **2007**, 47, 2149–2158.
- (22) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28, 31–36.

- (23) 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.
- (24) The numbers quoted are from a run on Linux server hardware, when running the same protocol on a Windows-based server a slightly different clustering is obtained and the similarity search performs marginally better than the Bayesian predictions.
- (25) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- (26) Rogers, D. Personal communication.
- (27) Barker, E. J.; Buttar, D.; Cosgrove, D. A.; Gardiner, E. J.; Kitts, P.; Willett, P.; Gillet, V. J. Scaffold Hopping Using Clique Detection Applied to Reduced Graphs. *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
- (28) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569–6583.
- (29) Bohl, M.; Loeprecht, B.; Wendt, B.; Heritage, T.; Richmond, N. J.; Willett, P. Unsupervised 3D Ring Template Searching as an Ideas Generator for Scaffold Hopping: Use of the LAMDA, RigFit, and Field-Based Similarity Search (FBSS) Methods. *J. Chem. Inf. Model.* **2006**, *46*, 1882–1890.
- (30) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.
- (31) Jenkins, J. L.; Glick, M.; Davies, J. W. A 3D Similarity Method for Scaffold Hopping from Known Drugs or Natural Ligands to New Chemotypes. *J. Med. Chem.* **2004**, *47*, 6144–6159.
- (32) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.
- (33) Tsunoyama, K.; Amini, A.; Sternberg, M. J. E.; Muggleton, S. H. Scaffold Hopping in Drug Discovery Using Inductive Logic Programming. *J. Chem. Inf. Model.* **2008**, *48*, 949–957.
- (34) Venhorst, J.; Nunez, S.; Terpstra, J. W.; Kruse, C. G. Assessment of scaffold hopping efficiency by use of molecular interaction fingerprints. *J. Med. Chem.* **2008**, *51*, 3222–3229.
- (35) Wale, N.; Watson, I. A.; Karypis, G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 730–741.
- (36) Zhang, Q.; Muegge, I. Scaffold Hopping through Virtual Screening Using 2D and 3D Similarity Descriptors: Ranking, Voting, and Consensus Scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.
- (37) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.
- (38) Bell, A. S.; van Hoorn, W. P. Parallel enabling our singleton file using a statistical approach. Abstracts of Papers, 236th ACS National Meeting, Philadelphia, PA, United States, August 17–21, 2008, 2008, ORGN-537.
- (39) Hughes, R. O.; Walker, J. K.; Cubbage, J. W.; Fobian, Y. M.; Rogier, D. J.; Heasley, S. E.; Blevis-Bal, R. M.; Benson, A. G.; Owen, D. R.; Jacobsen, E. J.; Freskos, J. N.; Molyneaux, J. M.; Brown, D. L.; Stallings, W. C.; Acker, B. A.; Maddux, T. M.; Tollefson, M. B.; Williams, J. M.; Moon, J. B.; Mischke, B. V.; Rumsey, J. M.; Zheng, Y.; MacInnes, A.; Bond, B. R. Yu, Y. Investigation of aminopyridiopyrazinones as PDE5 inhibitors: Evaluation of modifications to the central ring system. *Bioorg. Med. Chem. Lett.* [Online] 2009, in press, corrected proof. ScienceDirect. <http://dx.doi.org/10.1016/j.bmcl.2009.06.004> (accessed June 23, 2009).
- (40) Owen, D. R.; Walker, J. K.; Jon Jacobsen, E.; Freskos, J. N.; Hughes, R. O.; Brown, D. L.; Bell, A. S.; Brown, D. G.; Phillips, C.; Mischke, B. V.; Molyneaux, J. M.; Fobian, Y. M.; Heasley, S. E.; Moon, J. B.; Stallings, W. C.; Joseph Rogier, D.; Fox, D. N. A.; Palmer, M. J.; Ringer, T.; Rodriguez-Lens, M.; Cubbage, J. W.; Blevis-Bal, R. M.; Benson, A. G.; Acker, B. A.; Maddux, T. M.; Tollefson, M. B.; Bond, B. R.; MacInnes, A.; Yu, Y. Identification, synthesis and SAR of amino substituted pyrido[3,2b]pyrazinones as potent and selective PDE5 inhibitors. *Bioorg. Med. Chem. Lett.* 2009, in press, corrected proof. ScienceDirect. <http://dx.doi.org/10.1016/j.bmcl.2009.06.012> (accessed June 23, 2009).
- (41) DeNinno, M. P.; Andrews, M.; Bell, A. S.; Chen, Y.; Eller-Zarbo, C.; Eshelby, N.; Etienne, J. B.; Moore, D. E.; Palmer, M. J.; Visser, M. S.; Yu, L. J.; Zavadski, W. J.; Michael Gibbs, E. The discovery of potent, selective, and orally bioavailable PDE9 inhibitors as potential hypoglycemic agents. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2537–2541.
- (42) Owen, D. R.; Dodd, P. G.; Gayton, S.; Greener, B. S.; Harbottle, G. W.; Mantell, S. J.; Maw, G. N.; Osborne, S. A.; Rees, H.; Ringer, T. J.; Rodriguez-Lens, M.; Smith, G. F. Structure-activity relationships of novel non-competitive mGluR1 antagonists: A potential treatment for chronic pain. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 486–490.
- (43) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9*, 430–431.
- (44) Prezent, M. A.; Dorokhov, V. A. Boron chelates as intermediates in the synthesis of new functionalized pyridines and pyrimidines from a,a-dioxoketene amins. *Boron Chem. Beginning 21st Century, [Proc. Int. Conf. Chem. Boron]*, 11th **2003**, 91–93.

CI900072G