



BRADSHAW: a system for automated molecular design

Darren V. S. Green¹ · Stephen Pickett¹ · Chris Luscombe¹ · Stefan Senger¹ · David Marcus¹ · Jamel Meslamani² · David Brett³ · Adam Powell³ · Jonathan Masson³

Received: 10 June 2019 / Accepted: 5 October 2019 / Published online: 21 October 2019
© The Author(s) 2019

Abstract

This paper introduces BRADSHAW (Biological Response Analysis and Design System using an Heterogenous, Automated Workflow), a system for automated molecular design which integrates methods for chemical structure generation, experimental design, active learning and cheminformatics tools. The simple user interface is designed to facilitate access to large scale automated design whilst minimising software development required to introduce new algorithms, a critical requirement in what is a very fast moving field. The system embodies a philosophy of automation, best practice, experimental design and the use of both traditional cheminformatics and modern machine learning algorithms.

Keywords Automated design · Cheminformatics · Experimental design · Active learning

Introduction

The search for efficient and effective drug design strategies has been a constant feature of the scientific literature since the concept of rational discovery was introduced by Elion and Hitchings [1–3]. The field of quantitative structure–activity relationship (QSAR) analysis [4–6] developed alongside the rational approach, with the goal of being able to use chemical structures and biological response to develop hypotheses, predictions and design experiments which would provide an efficient path to optimise chemical series into promising drug candidates. Pertinent to this

paper, the QSAR community even provided the first example of an automated molecular design system [7], using simplex optimisation and Hansch parameters. Limitations to the—generally-linear statistical techniques employed, use of parameters that were often derived from experiment and insufficient computational power to cope with the combinatorial nature of the chemical space to be interrogated were enough to inhibit the utility of classical QSAR approaches and researchers looked for different approaches to solve their drug design problems.

Protein structure based design techniques have featured heavily in the computer aided drug discovery toolset for more than 30 years, and have had some notable successes in delivering marketed drugs [8, 9]. However, utilising the knowledge of protein–ligand binding interactions to drive the creation of novel, bioactive chemical structures, so-called de novo design [10, 11], has not been straight forward even after 25 years of intense effort. Although 3D methods for design may have dominated, ligand-based QSAR approaches mirrored some of the capabilities [12, 13], for example de novo design from QSAR models (also sometimes known as “inverse QSAR”), and even the first AI-based drug design support system [14]!

In order to establish an efficient and effective computational drug design system, there are a number of fundamental elements that must be constructed:

The paper is dedicated to the memory of Dr. John Bradshaw, Medicinal Chemist, Cheminformatician, Mentor and Friend.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10822-019-00234-8>) contains supplementary material, which is available to authorized users.

✉ Darren V. S. Green
darren.vs.green@gsk.com

¹ Department of Molecular Design, Data and Computational Sciences, GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, UK

² Department of Molecular Design, Data and Computational Sciences, GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, PA 19426, USA

³ Tessella Ltd, Walkern Road, Stevenage, Hertfordshire SG1 3QP, UK

- (1) *Molecule ideation*. The generation of chemical structures which are relevant to the problem at hand (i.e. lead optimisation), are synthetically tractable and do not contain structural liabilities.
- (2) *Prediction*. The generation and application of computational models which cover the entire Target Compound Profile (TCP: primary target, selectivity, ADMET, physicochemical attributes). It is likely that the suite of models will include both statistical/QSAR models and physics-based models [15].
- (3) *Selection*. From the list of ideas generated and evaluated with the model suite, which ones should be made next in order to meet the TCP, or more likely, to generate knowledge that will best enable the project to design a molecule which does meet the TCP.
- (4) An infrastructure that will enable these elements to be coupled together in a robust, efficient manner and can handle the likely large numbers of chemical structures produced.

Fortunately, the science of cheminformatics [if this is broadly defined to encompass chemical databases, chemical structure generation (Q)SAR modelling, experimental design and the application of machine learning methods to all three of those areas] has advanced to the point where credible solutions exist for all of the fundamentals and it may be possible to join them together. That this was indeed possible was illustrated by the first example of a fully automated design project, which started with an approved acetylcholinesterase inhibitor drug and created brain-penetrable ligands with either specific polypharmacology or selectivity profiles for different G-protein-coupled receptors [16]. Several groups have published examples of automated design [17, 18], often coupled with automated synthesis/test systems [19] which are well suited to pilot this new design paradigm [20].

The emergence of Deep Learning methods has quickly added a variety of techniques for QSAR modelling [21–25], molecule ideation [26–31] and synthetic tractability [32, 33]. The rapid rise of publication volume in this field indicates that more is to be expected and recent suggestions for standard benchmarks [34, 35] are welcome.

It is not only Deep Learning methods that can effectively utilise large SAR datasets. Knowledge based cheminformatics methods are able to mine and reapply chemical transformations in the form of Matched Molecular Pairs, MMPs [36–42], and automatically discover SAR patterns and suggest new substituents to optimise a biological response [43–45].

From the perspective of an industrial group, the rapid development of new algorithms and capabilities may provide a step change in our ability to generate and evaluate large numbers of novel chemical structures. The challenge

is to build a system which can integrate multiple approaches from disparate disciplines, make it robust enough to be used by multiple people across a portfolio of projects, make it simple to add or remove algorithms and make the tools accessible without repeated training as new code is added. Because the new algorithms are capable of generating very large numbers of ideas that will overwhelm a “selection by visual inspection” approach, a different way of working must be adopted. Here there are opportunities beyond just using new algorithms: to build in best practices (e.g. safety alerts [46–48], physicochemical properties [49–51], multi parameter optimisation), to automate the expert so they might spend more time on harder problems or work on more projects simultaneously and to greatly reduce the time and resources expended on end-user tools and training. This last point, to reduce the amount of time medicinal chemists spend using tailored computational software, might be considered controversial. However, a recent study from the banking sector [52] found that return on investment to corporations from end-user software development is often marginal, whilst there was significant value derived from investments in automation, and in a modern world where a virtual assistant can order your shopping via speech recognition, it is surely reasonable to be able to ask “Alexa, which are my best R groups?”.

In this paper we describe a solution to this challenge: BRADSHAW (**B**iological **R**esponse **A**nalysis and **D**esign **S**ystem using an **H**eterogenous, **A**utomated **W**orkflow), a system for automated molecular design which integrates methods for chemical structure generation, experimental design, active learning and cheminformatics tools. The project was inspired by a paper [53] which describes a system through which “A computer language, ALEMBIC, is used to collate the ideas of the scientists. The resulting list of potential molecules is then parametrised using whole molecule descriptors. Based on these descriptors, appropriate statistical techniques are used to generate sets of molecules retaining the maximum amount of the information inherent in all possible combinations of the scientists ideas”. If an algorithm can be considered as an additional scientist, BRADSHAW can be considered as a direct descendant of that heritage Glaxo system.

System Architecture

The BRADSHAW system is an integration between a number of external components. Its primary purpose is to orchestrate the running of data pipelines over compound sets (“Tasks”) and chaining the inputs/outputs of these Tasks to form designs.

The BRADSHAW system uses Docker containers to integrate various data management Tasks and the UI. The User

Interface is implemented using Angular CLI (version 1.1.1) [54].

The Tasks are described by an interface to a web service with a set of parameters, with the expected columns in input and output files also defined. This means that an administrator can easily create new Tasks without redeployment of the system. The administrator defines the parameters for a Task using a form in the user interface. This generates a form from which users can use these Tasks to build their design workflows for lead optimization with their chosen parameter values. An API interface is then imposed on web services when implementing the Tasks in order to integrate with BRADSHAW. The choice of technology for the web services is not imposed by the system; however typically these are built in Pipeline Pilot [55]. However, the BRADSHAW system itself knows nothing about Pipeline Pilot and any web-service could be used. Request parameters are passed in a single JSON object so that other web-services can be written if desired. We have developed utility components to read and validate the request JSON and to make it straightforward for developers to integrate with the call-back functions. The requirements for any protocol are minimal. Essentially the protocol must interpret the request JSON and send a call-back message to a defined url once complete. Status updates can also be sent using the call-back. This provides a way to give users feedback in the UI for long running Tasks.

The input of compound sets is from integration with either with LiveDesign [56] or via a file upload. Users can view the content of any files whether inputs or intermediate results in a design to review the work as they go along. This includes the handling of large compound sets via file streaming.

A user constructs a design by sequential addition of Tasks to a workflow and therefore can deploy as few or as many computational steps as they require. Some Tasks have restricted connectivity, which is governed through the required inputs/outputs of the files of each strategy. This enables us to impose best practice where we feel it is justified (for example, the Molecule Generator must always be followed by a Filter step to remove structures which do not meet agreed standards). Design workflows then, by definition, are validated. For simplicity, the designs are made up of a linear chain of Tasks. However, when a user wants to reuse the output of one part of a design they can use that as the input to multiple new designs, this permits branching flows.

Once a user runs a design there is immediate feedback to the user via web sockets allowing them to view the progress and logs from strategies and see the intermediate results as they are produced. There is also constant feedback on the health of running strategies via a beating heart in the user interface. The active feedback is enabled via a message queue in the system to push the status updates back to the user (Fig. 1).

A number of Tasks have been implemented and these are described in the next section.

Tasks

In BRADSHAW a Task is the term used to identify a particular scientific process. A Task may be complex and long running or relatively simple and fast. However, it performs one particular function, for example generating molecules from a lead compound, filtering generated molecules on properties, filtering molecules on substructure, selecting a subset of compounds using experimental design, preparing a file for analysis. The options for running a Task are configurable through the BRADSHAW UI (Fig. 2). However, one of the main design criteria is that the Task should encapsulate “best practice” by default. Extensive user configuration is not a primary design goal, for example, the Compound Molecule Filter contains defined settings for Oral and Inhaled drug-like properties. The user can provide their own XML configuration file to override these defaults but that should be manipulated outside of BRADSHAW.

All Tasks are implemented as Pipeline Pilot protocols and many are run standalone in other applications with the BRADSHAW protocol wrapping existing protocols or components.

The system is designed to make the deployment and integration of new scientific Tasks as easy as possible. Because the area of molecule design and machine learning algorithms is in a period of rapid change, we fully expect our current range of Tasks to have a high rate of turnover, or at least be modified to reflect progress in the scientific literature and experience of application in real projects. BRADSHAW will enable us to make cutting edge science accessible and integrated with other necessary workflows, with minimal software development or end user re-training. We have both test and production systems, so new Tasks can be thoroughly evaluated before being pushed to the production environment.

Molecule generation

The main purpose of this Task is to generate compounds from one or several lead molecules. At present this Task is focussed on lead optimisation, with relatively modest changes made to the lead molecules (a future goal is to add a “lead-hop” molecule generator, which may include additional generative methods). The methods implemented at the moment are a combination of established cheminformatics methods and more recent deep generative models. However, this Task is the poster child for a fast moving area of research, and we expect to add or replace methods on a regular basis.

Fig. 1 Overview of the BRADSHAW system

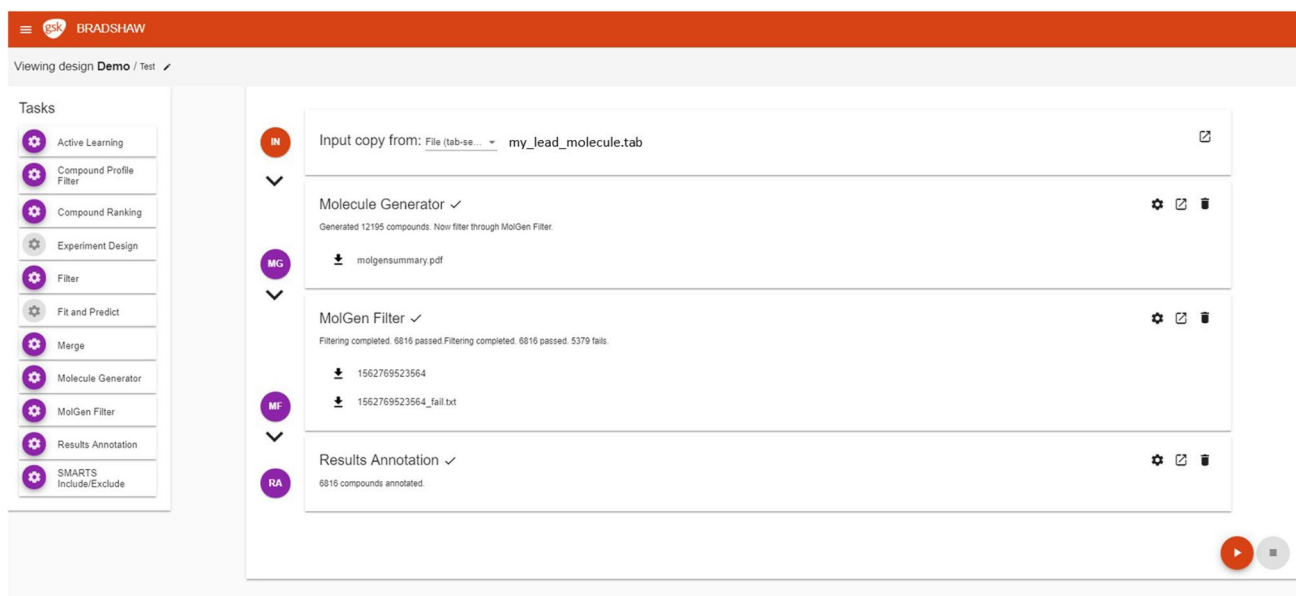
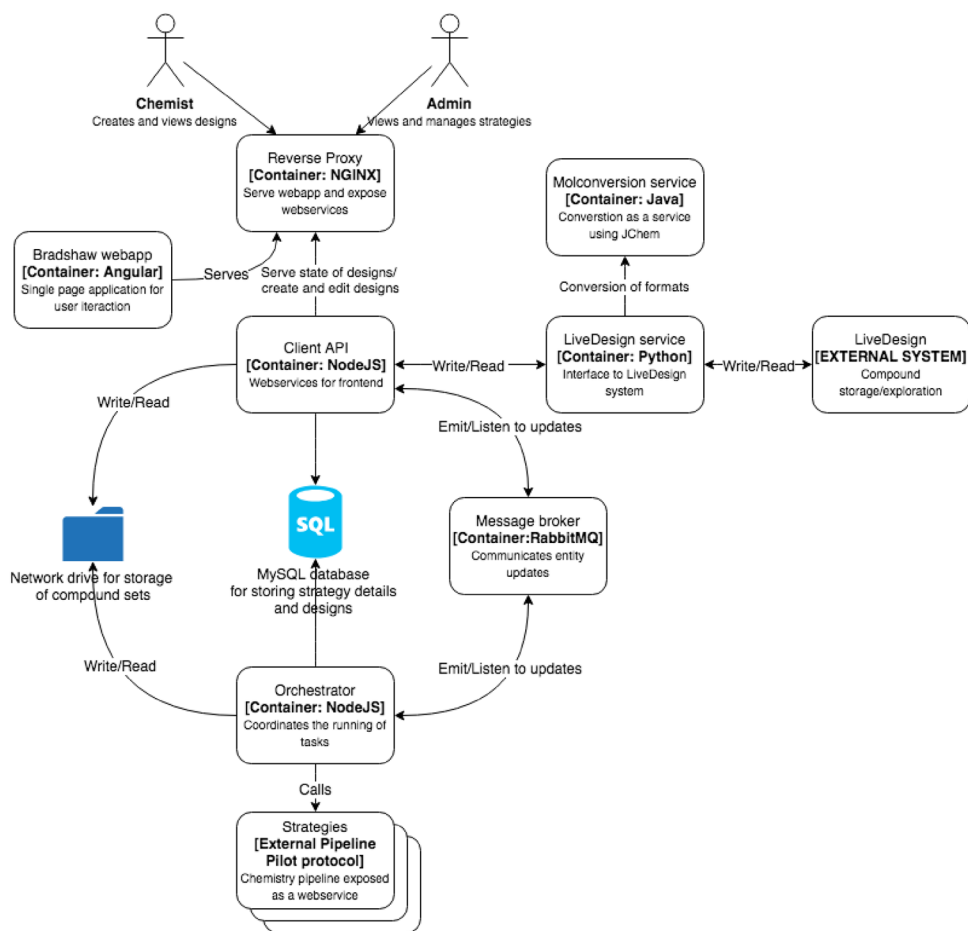


Fig. 2 The BRADSHAW User Interface. Available Tasks are on the left, and are coloured if they can be added to the current workflow

The user does not control which methods are run. However, there is an option to limit the output by similarity to a user-defined pool of compounds either by excluding or including generated molecules that are similar. Different criteria are defined for either case as explained below.

Molecule generation methods available currently are:

- (A) *MMP transforms*. We have an extensive collection of molecular transforms captured in the GSK BioDig system constructed using the algorithm of Hussain and Rea [37]. The input molecule is fragmented according to the fragmentation rules and BioDig is searched for relevant transforms. These transforms are then applied to the input molecule in a sequential fashion. For example a phenyl ring may be changed to a heterocycle or substituted with a methyl group.
- (B) *BRICS*. BRICS is a methodology developed to fragment molecules at defined chemically labile bonds [57]. This extends the well known RECAP approach [58, 59] with some additional disconnections and has been implemented in RDKit [59]. As implemented in RDKit BRICS works by fragmenting molecules and then recombining the fragments in an iterative fashion. As the fragments are recombined randomly and only to a predefined depth the output molecules can bear little resemblance to the input molecules. For our use case we prefer to modify the input molecules in a systematic fashion. Thus the RDKit implementation has been extended in Python to provide additional methods that provide greater control over the generation phase, removing the random combinations in BRICS and imposing a sequence of recombination which ensures that application of the procedure will generate molecules with similar topology to the input (and has the reassuring feature that a fragmented input molecule will always regenerate the input molecule!). In a separate process the BRICS fragmentation has been applied to the GSK collection. The fragments are collected into a pool according to the atom labels that indicate possible recombinations with frequency and molecular properties (heavy atoms, molecular weight). The modified algorithm is used to generate structures by randomly or exhaustively replacing a particular fragment in the input molecule with fragments from the pool that share the same labels, so generating new molecules for further evaluation. The BRICS bond definitions are very specific for certain atom types, ring types and substitution patterns so options are given to allow some flexibility when recombining. For example, BRICS differentiates bonds to aromatic and aliphatic rings. An option has been added to remove this restriction when combining fragments thus allowing a phenyl to be replaced by cyclohexyl. Fragments can also be

inserted (allowing for compound growth) or null fragments used to remove a fragment.

- (C) *Similarity searching*. Input molecules are searched against a number of internal and externally sourced datasets of available compounds, GSK collection, eMolecules [60], EnamineREAL [61]. Searches are performed using the MadFast application from ChemAxon [62] which provides results in seconds even for the largest databases. Two different fingerprints are used: ChemAxon path-based fingerprint 2048 bits, path 7, 4 bits per feature (CFP7) and the ChemAxon implementation of ECFP4 [63] with appropriately defined cutoffs [38].
- (D) *RG2SMI*. A deep generative model that generates molecules that have the same Reduced Graph as the input molecule [27]. Up to 1000 molecules are generated per input molecule.

As mentioned above there is an option to filter compounds according to similarity to an external pool of compounds or the input molecules. If the desire is to filter to similar compounds then a number of similarity measures are used: Tanimoto similarity using CFP7 and ECFP4 fingerprints and identical Reduced Graph [64]. The use of multiple techniques is to ensure all molecules of potential interest are carried forward. Where the user wishes to exclude similar compounds just the CFP7 is used at a tighter threshold, to ensure that only close analogues are excluded.

Compound profile filter

Multi parameter design is built in to the system. Compounds are filtered according to a set of property criteria as appropriate to the program objectives. The default options are preconfigured for Oral or Inhaled compounds. The user can provide a custom configuration by supplying their own XML file. A large range of molecular properties, predictive models for ADMET liabilities, permeability, solubility are available through GSK internal web-services, QSAR Workbench models [65] or as native Pipeline Pilot descriptors. The Task also runs GSK specific substructural filters to identify undesirable chemotypes [66]. It is a requirement of the system that the molecule generator output be processed by this Task. The intention is that the system never designs a molecule that is not predicted to be compatible with the program objectives, and does not produce molecules that contain chemical structure liabilities known to the organisation.

Compound rank

Compounds are ranked according to a user-defined multi parameter scoring algorithm. The properties and applicable ranges are defined via an XML file (an example file is

provided in the Supplementary Material). The ideal score for any property is 1 and decreases linearly from the defined min/max according to the gradient. Scores for all properties are summed and compounds are ranked according to the score. Alternatively the compounds are ranked according to the Pareto rank of the scores. Optionally the output can be limited to the top N compounds (with Pareto this will return all compounds from the appropriate Pareto front so more compounds may be returned).

SMARTS filtering

The program may be interested in particular chemotypes or indeed in avoiding particular substructures. This Task allows user-defined SMARTS to be entered and used to either include or exclude matched compounds. SMARTS filtering is implemented using the ChemAxon java toolkit.

Merge

A relatively simple and important Task. This allows addition of chemical structures which have been generated outside of BRADSHAW [for example, an enumeration of compounds that can be made using an available intermediate and accessible building blocks, or a LiveReport (the LiveDesign term for a user defined set of molecules and data) which is the result of a medicinal chemistry brainstorm session]. In this way, the machine and human generated ideas can be processed with the same workflow and subjected to exactly the same selection methods. All ideas are still equal and human creativity is added to the machine.

Active learning

A key component of the system is the ability to incorporate knowledge from previous rounds of screening and to suggest the most appropriate compounds for the next iteration. Active learning [67] provides a framework for such an approach. We have implemented an Active Learning protocol based on the modAL Python library [68]. A set of molecules, most likely derived from the molecule generation and filtering Tasks described above is used as the pool and a second set of molecules with measured activity data is defined for model building. The Task automatically builds and validates a QSAR model and uses it to predict the activity of the molecule pool. The output is a list of suggested molecules for synthesis at the next chemistry iteration, annotated as to whether the molecule was selected to “Explore” or “Exploit” from the QSAR model. The current Task uses Random Forests or the XGBoost variant, where the uncertainty in voting patterns in the trees is used to define if a molecule is in the “Explore” or “Exploit” category. The user has control over how much Exploration is performed in the

design step, as at present we do not have sufficient experience to set these automatically. Ideally this balance between Explore and Exploit would be set algorithmically and is the subject of current research.

Experimental design

At the start of a program or when exploring a new series there may be insufficient data to initiate an Active Learning cycle. We have implemented an experimental design Task that provides an efficient and informative mechanism for selecting compounds for a first iteration of lead expansion.

BRADSHAW utilises Design of Experiments (DOEs) based approaches for exploratory chemical array scenarios where the full ($M \times N$) array cannot be synthesized for practical reasons. By treating each monomer in the array as a categorical factor of the design, a balanced fractional array design can be generated. Once synthesized and measured, the results can be statistically analyzed to assess the additivity of SAR and then determine the contribution of the monomers to potency, selectivity and other properties of interest. This novel approach can be successfully used to understand and exploit the SAR of a late stage lead optimization program.

The approach combines well established experimental design techniques as a first step to achieve a well balanced design scheme. The appropriate scheme is utilized in a second library optimization step that allows the incorporation of whole molecule properties or other considerations into the choice of final products. In the standard workflow the user defines a combinatorial array of molecules for synthesis from available building blocks using design tools outside of BRADSHAW such as LiveDesign. LiveDesign is used to define the scaffold(s) and R-group definitions and the annotated LiveReport used as input to the Task.

To create an experimental design each R Group position must be treated as a categorical factor and each monomer at that position as a categorical level of that factor. The objective of the Design generation is to sample across this categorical feature space in such a manner that a robust and objective assessment of the monomers' contribution to the response can be made. For the purposes of a Medicinal Chemistry Lead Optimization problem we only need two or three factors but the ability to specify large numbers of levels per factor is required. The experimental design that produced by BRADSHAW is an equally sampled, incomplete balanced block design (colloquially known as a “sparse” array) in which each monomer at a particular position is treated equally, i.e. it is sampled the same number of times as others in that position. The sampling rate will be different at different R-Groups unless the number of levels is identical.

Once the number of levels has been specified at each R Group position the only decision that needs to be made is

the size of the fraction to be sampled in order to explore the categorical feature space. For example, if RG1 has 24 monomer levels and RG2 has 32 monomer levels then it is possible to extract a 1/8th fraction in which each RG1 monomer is chosen 4 times and each RG2 monomer is chosen 3 times. The number of compounds sampled is 96. This is shown visually in Fig. 3 where each spot represents a compound to be synthesized from this virtual array.

The Pipeline Pilot protocol uses R to build either an equal sampling design (two R-group positions) or a factorial design for the defined number of compounds. The output from the design is essentially a list of coordinates in the dimension of the virtual library. The designs themselves are agnostic to the actual characteristics of the individual reagents. They only know the dimensionality of the problem and the number of items in each dimension. Thus, the user needs to assign actual physical reagents to each position in the final design. This problem has been addressed in keeping with the automated philosophy behind BRADSHAW. We have developed and implemented a novel library design algorithm that is completely generic in nature in that the algorithm makes no assumptions about the shape of the design, combinatorial or non-combinatorial. The user defines the product map as a set of coordinates within the virtual library based upon the chosen experimental design, specifying which points of the virtual library should be included in the design. The algorithm assigns reagents to these positions in such a way as to optimize user-supplied product-based properties for the resultant library, starting with random assignment followed by a steepest descent optimiser.

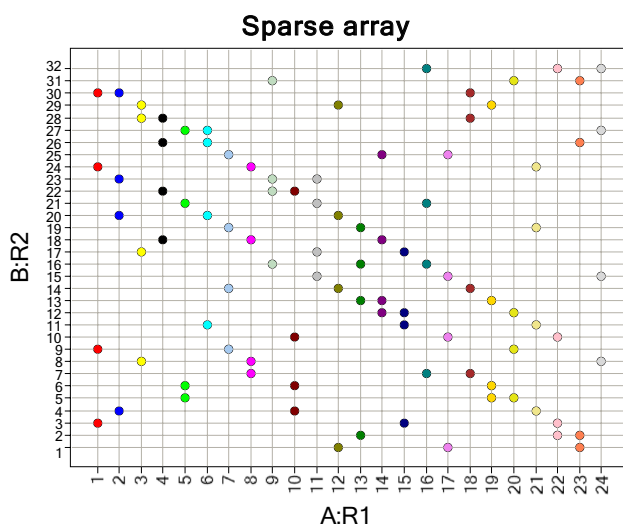


Fig. 3 Example of a sparse array design for a two component library

Fit and Predict

The SAR generated from a Sparse Array is ideally suited for a classical Free-Wilson analysis [69]. In keeping with the automated philosophy, the Fit and Predict Task is used to both build the QSAR model and predict the activity of all combinations of R groups in the data set.

Results annotation

A lesson learned in our pilot projects is that it is most important to annotate the selected compounds in order to facilitate a smooth project team discussion:

- What parts of the structure are being modified? Annotating in this way can integrate the output of a machine-driven process back into the thought process of a medicinal chemist and therefore aid assimilation and acceptance of the ideas.
- What are the predictions and what is the confidence?
- What is the likely synthetic tractability? Note that at this score is used as an annotation, rather than a filter, because none of the published methods perform well enough on our in-house data for them to be considered predictive. The current implementation uses the SCScore [33] algorithm. At some stage in the future we expect that this will change, and at that point the synthetic tractability would be moved further up in the design workflow, either within the molecular generators or as a compound filter.
- More general annotation is based on the GSK SIV framework [70]. Compounds are clustered using several methods: sphere exclusion clustering, framework clustering [64] at different levels. Compound properties are computed such as PFI [51].

Examples

Molecule generation of adenosine A2A antagonists

The discovery and optimisation of Adenosine A2A antagonists [71, 72] is the first example of successful structure based design using crystal structures of a GPCR. The discovery of the clinical candidate, AZD4635 (compound 1), was based on an initial hit, the commercially available 5,6-diphenyl-1,2,4-triazine-3-amine (compound 2). Optimisation was driven using crystallography, computational methods and state of the art medicinal chemistry thinking. 176 Chemical structures are exemplified in the Heptares Patent [73], as retrieved from SureChEMBL [74]. BRADSHAW was fed 5,6-diphenyl-1,2,4-triazine-3-amine as an input and using a single pass of molecule generation (only making

changes to the one starting structure), filtered according to our Oral Drug protocol and keeping the 1,2,4-triazine-3-amine headgroup, produced 10,675 structures. 72/176 of the exemplified structures were generated and a further 63 close analogues (ECFP4 Tanimoto > 0.8) generated which were not exemplified in the patent (Scheme 1). The BioDig molecular generator produced a minority of the ideas (317 structures), but these were very efficient in mapping to the patented structures (29 exact matches). In a real-world application, this set would be added to by further molecular generation after the first SAR is generated. However, we believe this small example is a good demonstration that BRADSHAW generates the “right” kind of molecules, which are both relevant to the project and synthetically tractable.

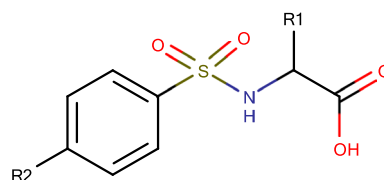
Experimental design of MMP-12 inhibitors

A recent publication [18] disclosed a full 2500 member combinatorial library with associate biological data. The dataset is particularly relevant as there are a number of compounds

that could not be synthesized successfully. The library was based on a core template as shown in Scheme 2.

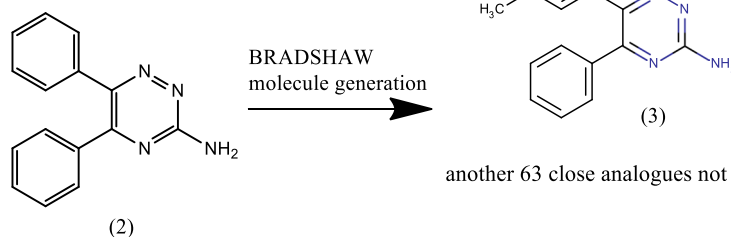
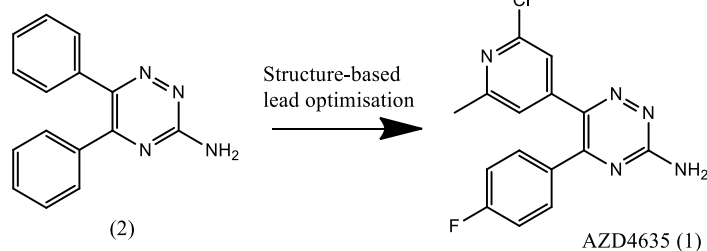
With such a complete array it is straightforward to identify the most potent compounds and the key monomers which are contributing to improved assay performance. The question that one could ask is whether the key findings from this particular chemical array could be obtained from only a fraction of the compounds?

Using BRADSHAW, an incomplete balanced block design is generated where by each of the 50 monomers in each position was selected twice and only twice in the

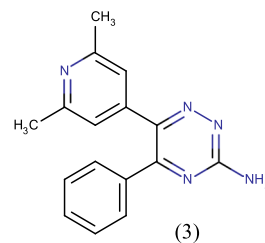


Scheme 2 The core template for a 50×50 array targeted against MMP12

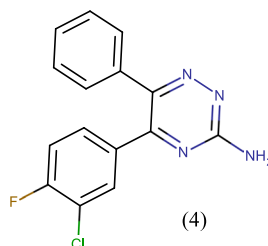
Scheme 1 Molecule generation for Adenosine A2A antagonists



72/176 exemplified structures, for example:



another 63 close analogues not exemplified, for example:



design. This produced a design in which 100 compounds were sampled from the 2500 array set. The balanced nature of the design meant that each monomer was equally leveraged and would thus in theory allow an unbiased assessment of the monomers potential to add or subtract value in terms of the assay response.

The created design pairs up monomers in R1 and R2 but the assignment of which monomer is allocated to which labelled Level can either be done randomly or by using an algorithmic approach to optimise the final assignment so as to produce a sampled set of compounds which are more optimally selected against a chosen property, e.g. lipophilicity, predicted permeability etc. For the purposes of this exercise we optimised the final selection to meet the design constraints (each monomer chosen twice and only twice) and also to maximise the leadlikeness of the molecules using a desirability function which combines counts

of hydrogen bond acceptors, donors and rotatable bonds, clogP, PFI, polar surface area and predictions of hERG and p450 inhibition.

Once the final compound set had been identified their respective potency values were retrieved from the full 2500 array data set. 64 Data records were found with measured potency data from the 100 compounds identified for synthesis. This is in line with the overall attrition rate in the completed array. A selection of the more interesting compounds is shown in Table 1, with the full list available in the Supplementary Material.

Using the Fit and Predict Task, a Free-Wilson QSAR model was built and the MMP12 activity of the non-selected members of the full array were predicted. The top 10 predictions are shown in Table 2.

The designed sparse array represents just 2.5% of the full array and the Free-Wilson model is built on just two

Table 1 Compounds selected from the MMP12 Sparse Array design along with their biological data

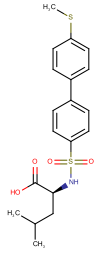

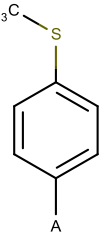
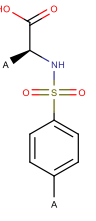
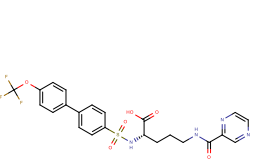

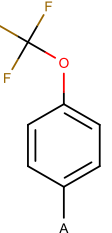
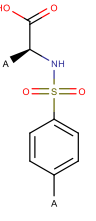
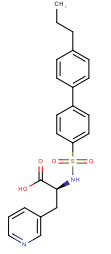
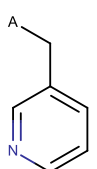
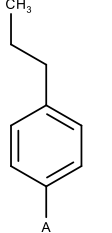
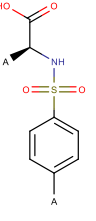
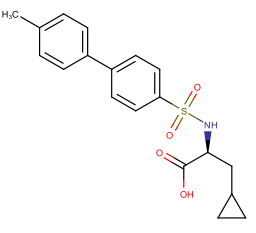
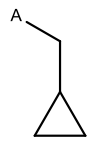
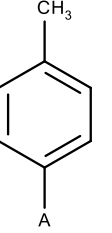
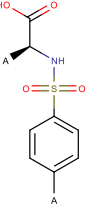
Compound Structure	ID	R1-smiles	R2-smiles	R0-smiles	MMP12_pIC50
	MOL1				7.8
	MOL4				7.3
	MOL5				7.1
	MOL7				6.9

Table 1 (continued)

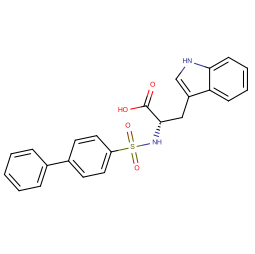
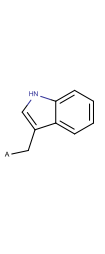
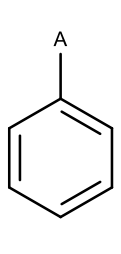
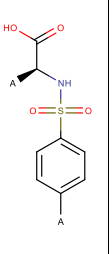
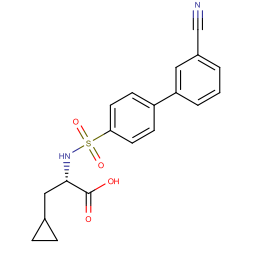
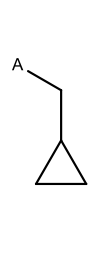
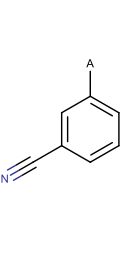
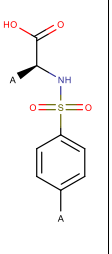
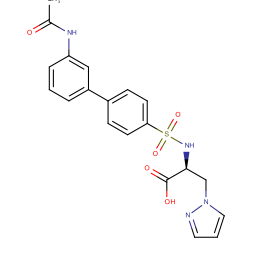
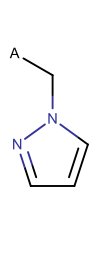
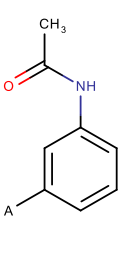
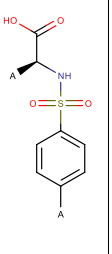
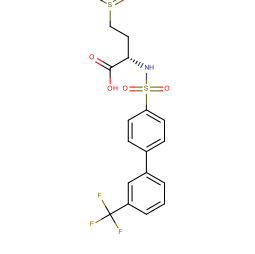
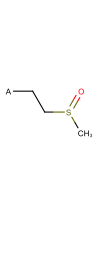
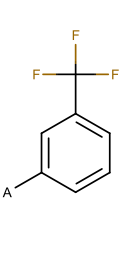
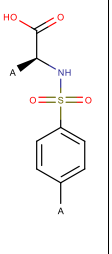
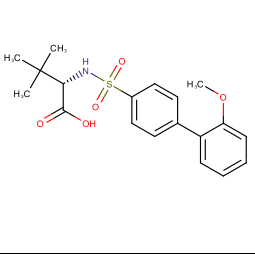
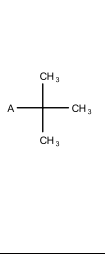
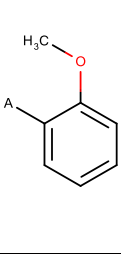
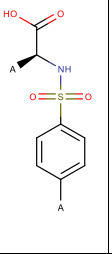
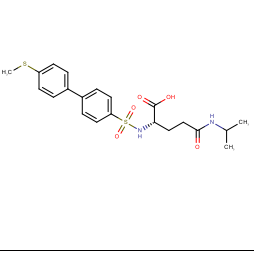
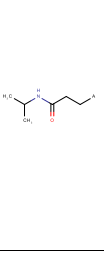
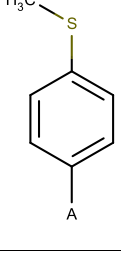
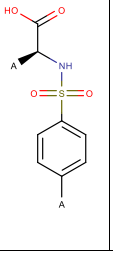
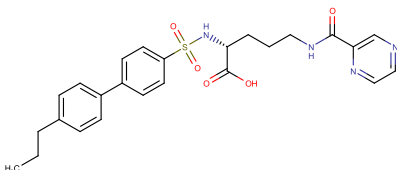
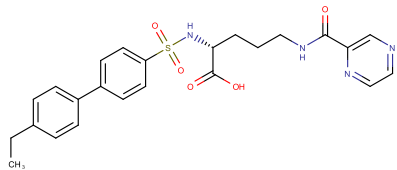
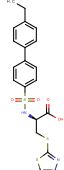
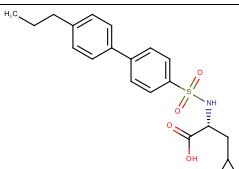
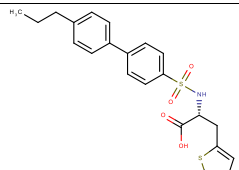
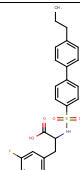
	MOL14				6.4
	MOL33				5.1
	MOL38				5
	MOL58				4.3
	MOL59				4.2
	MOL68				Could not be synthesized

Table 2 Compounds from the MMP12 set predicted to be the best actives using the Fit & Predict Task, along with their MMP12 activity

Compound Structure	ID	MMP12_PIC50
	MOL101	7.5
	MOL102	7.3
	MOL103	7.3
	MOL104	7.2
	MOL105	7.2
	MOL106	7.2

exemplars per monomer position. From this design, 8/10 of the top predicted molecules (that could be synthesized) have a $pIC_{50} > 7$. A comparison of the MMP12 activity for the whole data set, the sparse design and the Fit and Predict selections is shown in Fig. 4.

Active Learning of MMP12 inhibitors

Active learning is a powerful technique which is suitable to guide iteration experimentation such as Lead Optimisation and is a natural follow on to an initial Sparse Array design. As a demonstrator and comparator to the Fit and Predict example, the output of the MMP12 Sparse Array was used as a seed for a BRADSHAW Molecule

Generation and Active Learning workflow. The top 5 most active compounds from the Sparse Array were used as seeds for the Molecule Generator. 53,707 molecules were generated, filtered using the Oral molecule profile and a SMARTS query to only consider molecules with the phenyl-sulphonamide-acid core. Those molecules from the full MMP12 set that were not part of the sparse array were Merged, giving a total of 7385 compounds. These were fed as the selection pool to an Active Learning Task, which used the MMP12 pIC_{50} s from Sparse Array design as the training set. An XGBoost model was built with 100 trees, and the Active Learning was asked to select 50 molecules, with 80% Exploit and 20% Explore (chosen because the QSAR model from a sparse design

Table 2 (continued)

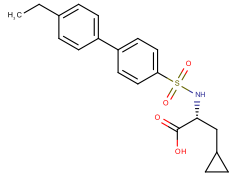
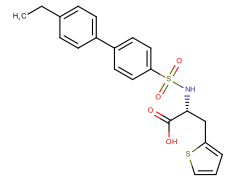
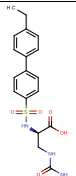
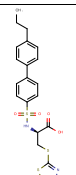
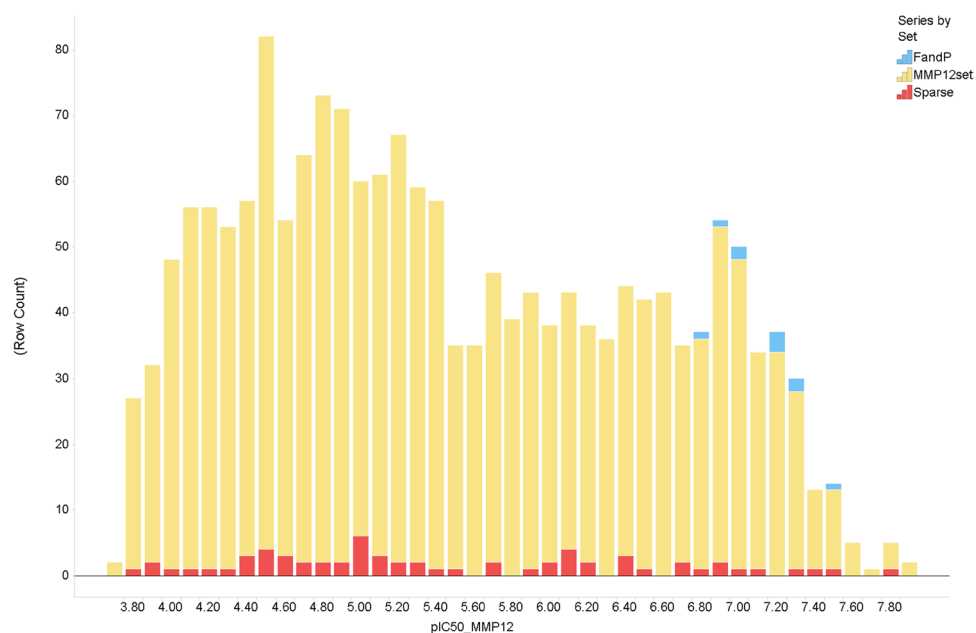
	MOL107	7
	MOL108	7
	MOL109	6.9
	MOL110	6.8

Fig. 4 MMP12 activity distribution across the MMP12 data set (yellow), the sparse design (red) and the Fit and Predict selections (blue)



is generally good, favouring Exploit, and our experience with Active Learning is that some Explore is always a good idea). A selection of the resulting molecules are shown in Tables 3 and 4. 26 of the compounds selected are present in the experimental MMP12 data, including one of the most potent compounds at pIC₅₀ of 8. For the

generated structures, the best biphenyl substituents (CS- and propyl-) are selected amongst the molecules and the SAR around alkyl amino-acids is sampled in some detail.

Table 3 Molecules from the full MMP12 data set selected by the Active Learning Task

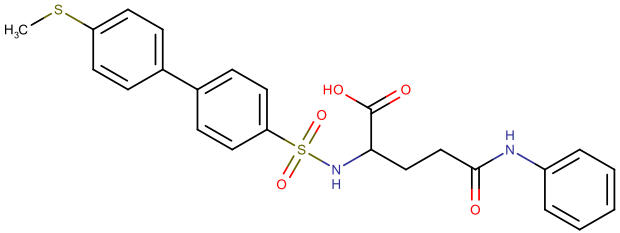
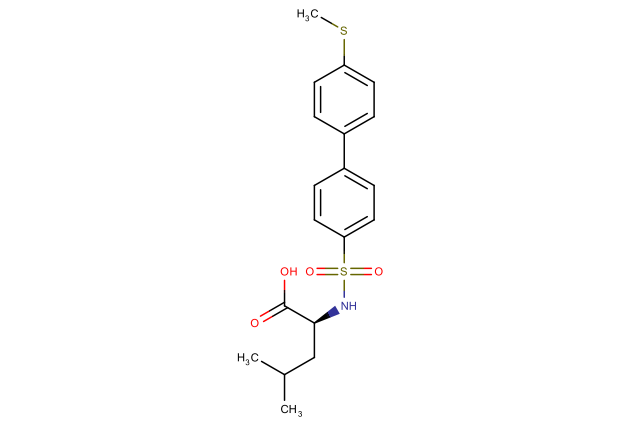
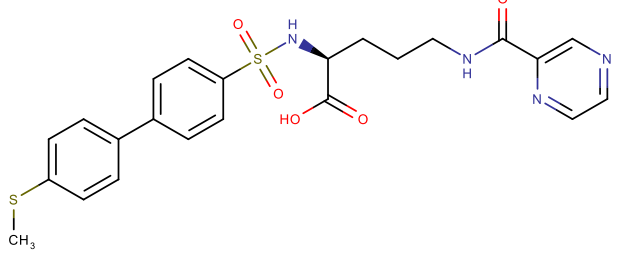
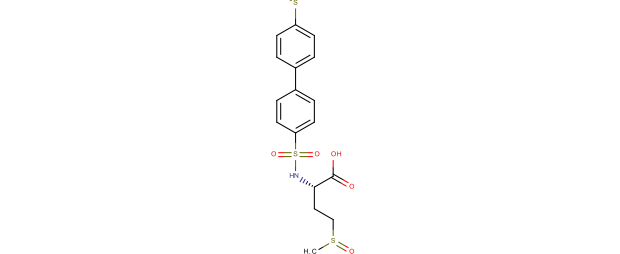
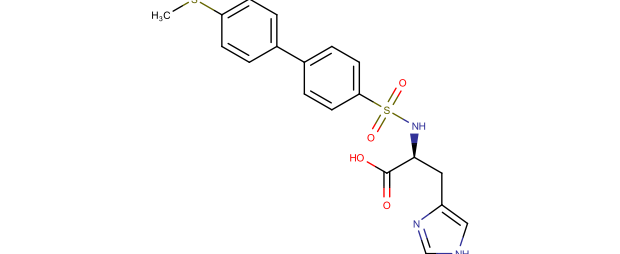
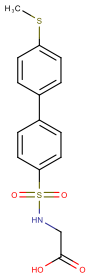
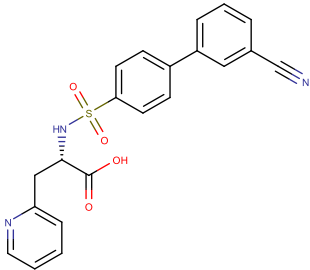
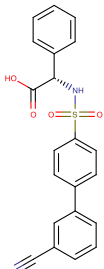
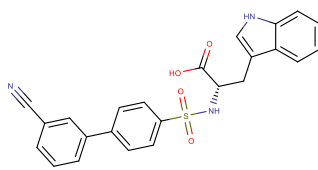
Compound Structure	ID	MMP12_IC50	Selection
	MOL111	8	Exploit
	MOL112	7.8	Exploit
	MOL113	7.2	Exploit
	MOL114	6.9	Exploit
	MOL115	6.5	Exploit

Table 3 (continued)

	MOL116	6.3	Exploit
	MOL117	5	Exploit
	MOL118	4.5	Exploit
	MOL119	Could not be synthesized	Explore

The future

In the current implementation, BRADSHAW is limited to cheminformatics and machine learning models. There are no 3D or docking methods, partly because it is difficult to produce a standard Task that can cater for the more bespoke models typically required in that domain, and partly because the predictivity and uncertainty quantification is not yet at level that we are comfortable including these as scoring functions. Those physics-based methods that are more rigorous, for example FEP+ [75], require an element of specialist control that, again, is not yet suitable for inclusion in the BRADSHAW framework, but may be included in a design workflow as an additional step [76].

As we hope to have demonstrated, BRADSHAW is at a respectable level of competency in the combined processes that comprise molecule design. The aforementioned rapid pace of innovation and improvement in the field of molecule generation, model building and optimisation algorithms means that systems like BRADSHAW will improve from this level. The potential level of capability that can be reached is difficult to predict. However, even the current level of performance raises questions about how the system should best be integrated into the work practice of a lead optimisation team e.g. is it as a complement to the creativity of the team, or should the automated system become the fundamental workhorse for the team which is complemented by suggestions and decisions by the human supervisors? We feel these answers will become clearer as the platform is

Table 4 Molecules generated by BRADSHAW and selected by the Active Learning Task

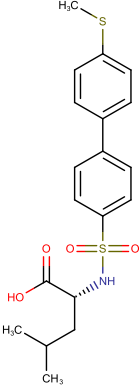
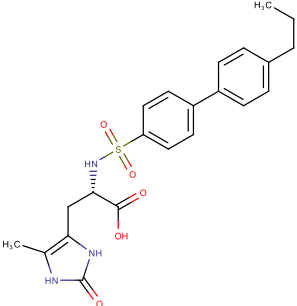
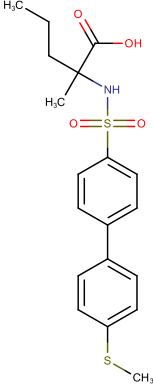
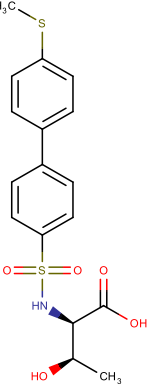
Compound Structure	ID	Selection Method
	MOL119	Exploit
	MOL120	Exploit
	MOL121	Exploit
	MOL122	Exploit

Table 4 (continued)

	MOL123	Exploit
	MOL124	Exploit
	MOL125	Exploit
	MOL126	Exploit
	MOL127	Exploit

further tested in prospective applied scenarios, which will be the subject of a future publication.

Acknowledgements The authors thank Eric Manas for helpful discussions and support of this work. Also Constantine Kreatsoulas, Jacob Bush, Peter Pogany, Ian Wall, Sandeep Pal, Guanglei Cui, Jen Elwood for their contributions to our automated design community.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Adam M (2005) Integrating research and development: the emergence of rational drug design in the pharmaceutical industry. *Stud Hist Philos Biol Biomed Sci* 36(3):513–537. <https://doi.org/10.1016/j.shpsc.2005.07.003>
- Elion GB (1969) Actions of furine analogs: enzyme specificity studies as a basis for interpretation and design. *Cancer Res* 29:2448–2453
- Kresge N, Simoni RD, Hill RL (2008) The rational design of nucleic acid inhibitors to treat leukemia: the work of George H. Hitchings. *J Biol Chem* 283:e10
- Hansch C, Maloney CP, Fujita T (1962) Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194:178–180
- Maliski EG, Bradshaw J (1993) QSAR and the role of computers in drug design. In: Ganellin CR, Roberts SM (eds) *Medicinal chemistry: the role of organic chemistry in drug discovery*. Academic, London
- Martin YC (1978) *Quantitative drug design*. Marcel Dekker, New York
- Darvas F (1973) Application of the sequential simplex method in designing drug analogs. *J Med Chem* 17(8):799–804
- von Itzstein M, Wu W-Y, Kok G, Pegg MS, Dyason JC, Jin B, Phan TV, Smythe M, White HF, Oliver SW, Colman PM, Varghese JN, Ryan DM, Woods JM, Bethell RC, Hotham VJ, Cameron JM, Penn CR (1993) Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* 363:418–423
- Ghosh AK, Gemma S (2014) *Structure-based design of drugs and other bioactive molecules: tools and strategies*. Wiley-VCH Verlag GmbH, Weinheim
- Glen RC, Payne AWR (1995) A genetic algorithm for the automated generation of molecules within constraints. *J Comput Aided Mol Des* 9:181–202
- Nishibata Y, Itai A (1991) Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* 47(43):8985–8990
- Kier LB, Hall LH (1993) The generation of molecular structures from a graph-based QSAR equation. *Quant Struct Act Relatsh* 12:383–388
- Gordeeva EV, Molchanova MS, Zefirov NS (1991) General methodology and computer program for the exhaustive restoring of chemical structures by molecular connectivity indexes. Solution of the inverse problem in QSAR/QSPR. *Tetrahedron Comput Methodol* 3(6):389–415
- Hodgkin EE (1994) The Castlemaine project: development of an AI-based drug design support system. In: Vinter JG, Gardner M (eds) *Molecular modelling and design. Topics in molecular and structural biology*. The Macmillan Press Ltd, Basingstoke, pp 137–167
- Manas ES, Green DV (2017) CADD medicine: design is the potion that can cure my disease. *J Comput Aided Mol Des* 31(3):249–253. <https://doi.org/10.1007/s10822-016-0004-3>
- Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguez RM, Huang XP, Norval S, Sassano MF, Shin AI, Webster LA, Simeons FR, Stojanovski L, Prat A, Seidah NG, Constam DB, Bickerton GR, Read KD, Wetsel WC, Gilbert IH, Roth BL, Hopkins AL (2012) Automated design of ligands to polypharmacological profiles. *Nature* 492(7428):215–220. <https://doi.org/10.1038/nature11691>
- Merk D, Grisoni F, Friedrich L, Gelzinyte E, Schneider G (2018) Computer-assisted discovery of retinoid X receptor modulating natural products and isofunctional mimetics. *J Med Chem* 61(12):5442–5447. <https://doi.org/10.1021/acs.jmedchem.8b00494>
- Pickett SD, Green DV, Hunt DL, Pardoe DA, Hughes I (2011) Automated lead optimization of MMP-12 inhibitors using a genetic algorithm. *ACS Med Chem Lett* 2(1):28–33. <https://doi.org/10.1021/ml100191f>
- Pant SM, Mukonoweshuro A, Desai B, Ramjee MK, Selway CN, Tarver GJ, Wright AG, Birchall K, Chapman TM, Tervonen TA, Klefstrom J (2018) Design, synthesis, and testing of potent, selective Hepsin inhibitors via application of an automated closed-loop optimization platform. *J Med Chem* 61(10):4335–4347. <https://doi.org/10.1021/acs.jmedchem.7b01698>
- Schneider G (2018) Automating drug discovery. *Nat Rev Drug Discov* 17:97–113
- Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V (2017) Demystifying MultiTask deep neural networks for quantitative structure–activity relationships. *J Chem Inf Model* 57(10):2490–2504. <https://doi.org/10.1021/acs.jcim.7b00087>
- Ramsundar B, Liu B, Wu Z, Verras A, Tudor M, Sheridan RP, Pande V (2017) Is MultiTask deep learning practical for pharma? *J Chem Inf Model* 57(8):2068–2076. <https://doi.org/10.1021/acs.jcim.7b00146>
- Altae-Tran H, Ramsundar B, Pappu AS, Pande V (2017) Low data drug discovery with one-shot learning. *ACS Cent Sci* 3(4):283–293. <https://doi.org/10.1021/acscentsci.6b00367>
- Wenzel J, Matter H, Schmidt F (2019) Predictive MultiTask deep neural network models for ADME-Tox properties: learning from large data sets. *J Chem Inf Model* 59(3):1253–1268. <https://doi.org/10.1021/acs.jcim.8b00785>
- Feinberg EN, Sheridan R, Joshi E, Pande VS, Cheng AC (2019) Step change improvement in ADMET prediction with Potential-Net deep featurization. arXiv: 190311789v1
- Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Mol Pharm* 14(9):3098–3104. <https://doi.org/10.1021/acs.molpharmaceut.7b00346>
- Pogany P, Arad N, Genway S, Pickett SD (2018) De novo molecule design by translating from reduced graphs to SMILES. *J Chem Inf Model*. <https://doi.org/10.1021/acs.jcim.8b00626>
- Segler MHS, Kogej T, Tyrchan C, Waller MP (2017) Generating focussed molecule libraries for drug discovery with recurrent neural networks. arXiv: 170101329v1
- Lim J, Ryu S, Kim JW, Kim WY (2018) Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J Cheminform* 10(1):31. <https://doi.org/10.1186/s13321-018-0286-7>
- You J, Liu B, Ying R, Pande V, Leskovec J (2019) Graph convolutional policy network for goal-directed molecular graph generation. arXiv: 180602473v3

31. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminform*. <https://doi.org/10.1186/s13321-017-0235-x>
32. Segler MHS, Waller MP (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* 23(25):5966–5971. <https://doi.org/10.1002/chem.201605499>
33. Coley CW, Rogers L, Green WH, Jensen KF (2018) SCScore: synthetic complexity learned from a reaction corpus. *J Chem Inf Model* 58(2):252–261. <https://doi.org/10.1021/acs.jcim.7b00622>
34. Polykovskiy D, Tatanov O, Zhebrak A, Belyaev S, Sanchez-Lengeling B, Kurbanov R, Zhavoronkov A (2018) Molecular sets (MOSES): a benchmarking platform for molecular generation models. arXiv: 1811.12823
35. Brown N, Fiscato M, Segler MHS, Vaucher AC (2019) GuacaMol: benchmarking models for de novo molecular design. *J Chem Inf Model* 59(3):1096–1108. <https://doi.org/10.1021/acs.jcim.8b00839>
36. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, Wood JM, Colclough N, Law B (2006) Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J Med Chem* 49:6672–6682
37. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J Chem Inf Model* 50:339–348
38. Papadatos G, Alkarouri M, Gillet VJ, Willett P, Kadiramanathan V, Luscombe CN, Bravi G, Richmond NJ, Pickett SD, Hussain J, Pritchard JM, Cooper AJ, Macdonald SJF (2010) Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J Chem Inf Model* 50:1872–1886
39. Wassermann AM, Bajorath J (2011) Large-scale exploration of bioisosteric replacements on the basis of matched molecular pairs. *Future Med Chem* 3(4):425–436
40. Warner DJ, Griffen EJ, St-Gallay SA (2010) WizePairZ: a novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J Chem Inf Model* 50:1350–1357
41. Griffen E, Leach AG, Robb GR, Warner DJ (2011) Matched molecular pairs as a medicinal chemistry tool. *J Med Chem* 54(22):7739–7750. <https://doi.org/10.1021/jm200452d>
42. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J Chem Inf Model* 52(5):1138–1145. <https://doi.org/10.1021/ci3001138>
43. O’Boyle NM, Bostrom J, Sayle RA, Gill A (2014) Using matched molecular series as a predictive tool to optimize biological activity. *J Med Chem* 57(6):2704–2713. <https://doi.org/10.1021/jm500022q>
44. Keefer CE, Chang G (2017) The use of matched molecular series networks for cross target structure–activity relationship translation and potency prediction. *MedChemComm* 8(11):2067–2078. <https://doi.org/10.1039/c7md00465f>
45. Ehmki ESR, Kramer C (2017) Matched molecular series: measuring SAR similarity. *J Chem Inf Model* 57(5):1187–1196. <https://doi.org/10.1021/acs.jcim.6b00709>
46. Schultz TW, Diderich R, Kuseva CD, Mekenyan OG (2018) The OECD QSAR toolbox starts its second decade. *Methods Mol Biol* 1800:55–77
47. Sushko I, Salmina E, Potemkin VA, Poda G, Tetko IV (2012) ToxAlerts: a Web server of structural alerts for toxic chemicals and compounds with potential adverse reactions. *J Chem Inf Model* 52(8):2310–2316. <https://doi.org/10.1021/ci300245q>
48. Matthews EJ, Kruhlak NL, Benz RD, Contrera JF, Marchant CA, Yang C (2008) Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscape PDM, and Derek for Windows software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol Mech Methods* 18(3):189–206
49. Waring MJ (2010) Lipophilicity in drug discovery. *Expert Opin Drug Discov* 5(3):235–248
50. Hill AP, Young RJ (2010) Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discov Today* 15(15–16):648–655. <https://doi.org/10.1016/j.drudis.2010.05.016>
51. Young RJ, Green DV, Luscombe CN, Hill AP (2011) Getting physical in drug discovery II: the impact of chromatographic hydrophobicity measurements and aromaticity. *Drug Discov Today* 16(17–18):822–830. <https://doi.org/10.1016/j.drudis.2011.06.001>
52. Hoene M (2016) Spend wisely, not more, on IT. McKinsey. Accessed 8/4/2019
53. Maliski EG, Latour K, Bradshaw J (1992) The whole molecule design approach to drug discovery. *Drug Des Discov* 9:1–9
54. <https://github.com/angular/angular-cli>. Accessed 18th Oct 2019
55. Dassault Systèmes (2019) Dassault Systèmes BIOVIA PP. Dassault Systèmes, San Diego
56. Schrödinger LLC (2019) LiveDesign. Schrödinger LLC, New York
57. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) On the art of compiling and using ‘drug-like’ chemical fragment spaces. *ChemMedChem* 3(10):1503–1507. <https://doi.org/10.1002/cmdc.200800178>
58. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) RECAPs-Retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J Chem Inf Comput Sci* 38:511–522
59. Fan Z, Casey FXM (2008) Estimating solute transport parameters using stochastic ranking evolutionary strategy. *Vadose Zone J* 7(1):124. <https://doi.org/10.2136/vzj2007.0021>
60. Sanz F, Pognan F, Steger-Hartmann T, Diaz C, eTox, Cases M, Pastor M, Marc P, Wichard J, Briggs K, Watson DK, Kleinoder T, Yang C, Amberg A, Beaumont M, Brookes AJ, Brunak S, Cronin MTD, Ecker GF, Escher S, Greene N, Guzman A, Hersey A, Jacques P, Lammens L, Mestres J, Muster W, Northeved H, Pinches M, Saiz J, Sajot N, Valencia A, van der Lei J, Vermeulen NPE, Vock E, Wolber G, Zamora I, (2017) Legacy data sharing to improve drug safety assessment: the eTOX project. *Nat Rev Drug Discov* 16(12):811–812. <https://doi.org/10.1038/nrd.2017.177>
61. Shivanyuk LN, Bogolyubsky AV, Mykytenko DM, Chupryna AA, Heilman W, Kostyuk AN, Tolmachev AA (2007) Enamine real database: making chemical diversity real. *Chem Today* 25(6):58–59
62. Briggs K, Cases M, Heard DJ, Pastor M, Pognan F, Sanz F, Schwab CH, Steger-Hartmann T, Sutter A, Watson DK, Wichard JD (2012) Inroads to predict in vivo toxicology—an introduction to the eTOX Project. *Int J Mol Sci* 13(3):3820–3846. <https://doi.org/10.3390/ijms13033820>
63. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. *J Chem Inf Model* 50:742–754
64. Harper G, Bravi GS, Pickett SD, Hussain J, Green DVS (2004) The reduced graph descriptor in virtual screening and data-driven clustering of high-throughput screening data. *J Chem Inf Comput Sci* 44:2145–2156
65. Cox R, Green DVS, Luscombe CN, Malcolm N, Pickett SD (2013) QSAR workbench: automating QSAR modeling to drive compound design. *J Comput Aided Mol Des* 27(4):321–336. <https://doi.org/10.1007/s10822-013-9648-4>
66. Chakravorty SJ, Chan J, Greenwood MN, Popa-Burke I, Remlinger KS, Pickett SD, Green DVS, Fillmore MC, Dean TW,

- Luengo JI, Macarron R (2018) Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discov* 23(6):532–545. <https://doi.org/10.1177/2472555218768497>
67. Reker D, Schneider G (2015) Active-learning strategies in computer-assisted drug discovery. *Drug Discov Today* 20(4):458–465. <https://doi.org/10.1016/j.drudis.2014.12.004>
68. <https://github.com/modAL-python/modAL>. Accessed 18th Oct 2019
69. Free SM, Wilson JW (1964) A mathematical contribution to structure–activity studies. *J Med Chem* 7(4):395–399
70. Leach ARL, Green DVS, Hann MM, Harper G, Whittington AR (2001) SIV: a synergistic approach to the analysis of high-throughput screening data. In: Paper presented at the 221st national meeting of the American Chemical Society, San Diego, CA
71. Congreve M, Brown GA, Borodovsky A, Lamb ML (2018) Targeting adenosine A2A receptor antagonism for treatment of cancer. *Expert Opin Drug Discov* 13(11):997–1003. <https://doi.org/10.1080/17460441.2018.1534825>
72. Congreve M, Andrews SP, Dore AS, Hollenstein K, Hurrell E, Langmead CJ, Mason JS, Ng IW, Tehan B, Zhukov A, Weir M, Marshall FH (2012) Discovery of 1,2,4-triazine derivatives as adenosine A(2A) antagonists using structure based drug design. *J Med Chem* 55(5):1898–1903. <https://doi.org/10.1021/jm201376w>
73. Congreve M, Andrews SP, Mason JS, Richardson CM, Brown GA (2019) 1,2,4-Triazine-4-amine derivatives. United States Patent US20170291888A1
74. <https://www.surechembl.org/search/>. Accessed 18th Oct 2019
75. Wang L, Wu Y, Deng Y, Kim B, Pierce L, Krilov G, Lupyan D, Robinson S, Dahlgren MK, Greenwood J, Romero DL, Masse C, Knight JL, Steinbrecher T, Beuming T, Damm W, Harder E, Sherman W, Brewer M, Wester R, Murcko M, Frye L, Farid R, Lin T, Mobley DL, Jorgensen WL, Berne BJ, Friesner RA, Abel R (2015) Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J Am Chem Soc* 137(7):2695–2703. <https://doi.org/10.1021/ja512751q>
76. Konze K, Bos P, Dahlgren M, Leswing K, Tubert-Brohman I, Bortolato A, Robbason B, Abel R, Bhat S (2019) Reaction-based enumeration, active learning, and free energy calculations to rapidly explore synthetically tractable chemical space and optimize potency of cyclin dependent kinase 2 inhibitors. *J Chem Inf Model* 59(9):3782–3793

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.