ELSEVIER

How well do medicinal chemists learn from experience?

David R. Cheshire

Department of Chemistry, AstraZeneca Charnwood, Bakewell Road, Loughborough, LE11 5RH, UK

To an outsider, the exploration of thousands of molecules to find a small number of potential candidate drugs must appear enormously wasteful, but many medicinal chemists would defend this waste as unavoidable. Here, I provide evidence that suggests that modern medicinal chemists are overproductive in that they synthesise many more compounds than are required to achieve the objectives of the project. The difficulties encountered in finding the data for the analysis presented here prompted the design and implementation of a more rigorous approach to capture the essence of a medicinal chemistry program. The result, medicinal chemistry knowledge sharing (MeCKS), was designed to capture and communicate emerging issues and their solutions to the medicinal chemistry community.

Introduction

The science of medicinal chemistry has been conducted in earnest for over 60 years. Have we made the most of this experience? To an outsider, the exploration of thousands of molecules to find a small number of potential candidate drugs (CDs; compounds suitable for toxicological assessment before human dosing) must appear enormously wasteful, but many medicinal chemists would defend this waste as necessary and unavoidable.

There is no doubt that the work of Lipinski, Leeson [1] and others has greatly increased the awareness in the medicinal chemistry community of the importance of physicochemical properties to early drug discovery. These reports suggest that if the search for a CD starts from a lead molecule with drug-like properties and continues in an optimum property space, the chances of discovering a successful and marketable drug will increase. It is hoped that, by adhering to these principles, the number of needless compounds made in a chemical program will be reduced.

However, others have argued [2] that these principles need to be followed with care, as many successful drugs fail the proposed criteria. Efficiency savings in medicinal chemistry and screening have also been reported [3]; however, although claims are made that time is saved and costs reduced, candidate drugs remain challenging to discover. In addition, it has been suggested that measuring efficiency savings and process performance can lead to staff demotivation and a reduction in innovation [4]. Surprisingly little has been reported about the way in which medicinal chemists take the lead and develop it into a CD [5]; that is, the practice known as lead optimisation.

Analysis of the lead optimisation process

To understand more about lead optimisation, I undertook a closer examination of ten projects conducted at one research site of AstraZeneca over a 10-year period. The data collected represents 32 of the 39 CDs delivered from all projects during that period. The unique compounds* made for each project during the lead optimisation phase were identified by discussions with the medicinal chemists, database searches and reading project reports. Medicinal chemists are all aware that, during the cycles of design, synthesis and testing, they will discover structural modifications that make significant progress towards achieving the project goal. An example of a significant step change (Fig. 1) is a structural change resulting in an improvement in bioavailability.

During the course of the work, it quickly became apparent that a 'series' is often poorly defined and subject to casual redefinition. Medicinal chemists use series names to communicate their plans and results, and to organise the synthesis of compounds. A series might be a group of prospective compounds, a collection of structurally similar compounds, or identified retrospectively for use in presentations or publications. At best, the name given to a

E-mail address: dave.cheshire@astrazeneca.com.

1359-6446/06/$ - see front matter © 2011 Published by Elsevier Ltd. doi:10.1016/j.drudis.2011.06.005

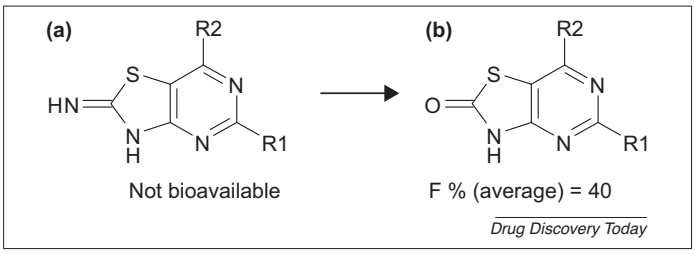


FIGURE 1

A structural change in a non-bioavailable series **(a)** resulting in a consistent bioavailability (F %) within the series **(b)**; 29 examples were measured.

series is descriptive either of the structures being made, for example 'the pyridine series', or the structural feature that marks it out as interesting, for example 'the methyl amide series'. At worst, it means nothing to anyone outside of the synthesis team. I offer a definition of a series as 'a distinct set of compounds that contain a common structural motif that consistently provides the analogues with the same unique advantage'. The advantage is simply an activity or property of the compound that is required to meet the objectives of the project. Importantly, the progress provided by the structural change should be seen in most of the compounds within the newly defined series. Of course, poor design, for example increasing lipophilicity, might remove the advantage from a minority of the series members.

Each CD is a collection of structural changes, usually exemplified in several series, which together provide the project with the desired profile for the CD. This analysis looks only at the last structural change to be discovered where the series defined by this structural change contained the CD. Table 1 shows data from ten projects: the first entry for GPCR1 reveals that the first CD was contained in Series 1 and that the total number of compounds made to discover all the CDs against this biological target (GPCR1) was 2250. The series with the first CD contained 197 compounds upon completion of synthesis; every compound containing the structural change was counted as a series member. In this example, the CD was the second compound to be synthesised in this series. Examining all the projects reveals that, once a structural change is made, the first CD (CD1) in the series is synthesised after a median of 19 analogues. The median number of analogues in a series is 147. Subtraction of the position of the CD1 from the total number of analogues in the series reveals that over 100 additional compounds (median of 128) are synthesised. Of the 23 discrete CD series identified in this work, eight produced further CDs (Table 2).

A feature of this analysis is how quickly after the step change in structure the first CD appears; for example, for biological target GPCR1 Series 1 (Table 1), the CD was the second member of the series. In addition, an average of 12% of the entire number of compounds synthesised are within the CD series†.

Several factors might explain the overproduction of analogues in a series:

(i) Extra analogues are made while waiting for the CD to be identified; however, in these projects, the synthesis of analogues continued after CD1 had been identified, and large-scale synthesis commenced. Inevitably, upon discovering CD1 some synthesis is already underway. Multidisciplinary project teams, and broader and quicker screening until the compound fails, can do much to remedy this [5].

(ii) The project team might have decided to wait for a better compound (e.g. in GCPR1 Series 1). Experience at AstraZeneca reveals that the early compound is usually the best unless the CD profile of the project is amended.

(iii) The project team might revise the required profile of the desired CD, which will then require re-appraisal of the existing chemical assets. However, in the analysis presented here, a significant structural change and a new series (according to the definition above) was required to provide the new profile.

(iv) Is it surprising that the CD is found early in the series? The CD is the product of hard-won learning acquired from previous efforts. This previous learning should enable the best array of substituents to be used immediately in the final CD series; the chemists simply need to click together the new structural change with the best substituents and produce the CD. How often can chemists simply stick together several desirable features of different molecules? Why then still synthesise the extra analogues in the CD series? In addition, further examination of several series, which contain at least one of the structural features embedded in the final CD, suggests that the structural change (a preferred substituent or core change) is also

TABLE 1

The position of the first CD in a series

Project[a]	Series name	Total number of compounds made in the Project	Number of compounds in 'CD series'	CD is n'th in series
GPCR1	Series 1	2250	197	2
	Series 2	2250	167	51
	Series 3	2250	64	1
GPCR2	Series 4	1600	336	10
	Series 5	1600	8	1
GPCR3	Series 6	2120	206	59
	Series 7	2120	486	300
	Series 8	2120	155	30
GPCR4	Series 9	1660	111	21
	Series 10	1660	228	14
	Series 11	1660	111	6
GPCR5	Series 12	820	77	36
	Series 13	820	23	11
	Series 14	820	63	1
	Series 15	820	145	47
GPCR6	Series 16	740	252	74
	Series 17	740	141	57
Enz2	Series 18	970	34	9
Enz1	Series 19	750	222	20
IC1	Series 20	3240	76	2
	Series 21	3240	47	23
	Series 22	3240	149	82
Mem1	Series 23	1300	237	10
	Series 23	1300	237	17
Median		1630	147	19

[a] Abbreviations: Enz, enzyme; GPCR, G protein-coupled receptor; IC, ion channel; Mem, membrane-bound receptor.

TABLE 2

The position of CD2 for series that contained a second CD

Project[a]	Series name	Total compounds made in Project	No. of compounds in 'CD series'	CD is n'th in series
GPCR1	Series 2	2250	167	113
	Series 2	2250	167	139
GPCR3	Series 7	2120	486	371
	Series 7	2120	486	479
GPCR4	Series 9	1660	111	87
	Series 11	1660	111	62
IC1	Series 20	3240	76	34
Mem1	Series 23	1300	237	130
Median		2120	167	122

[a] Abbreviations: GPCR, G protein-coupled receptor; IC, ion channel; Mem, membrane-bound receptor.

revealed early. This reinforces the observation that more analogues are made than necessary; not only within each series, but also within the entire project.

(v) Another CD might be needed to provide a back-up for the first. It might be thought that the obvious place to look is within the series that yielded the first CD. However, in most projects, synthesis continued before an issue was identified with CD1. Where a further CD (CD2) was found (Table 2), it is appears later within the series; a median 122nd place compared with 19th place for CD1. The median number of analogues made before discovering CD2 is more than that for CD1. My experience is that the issues arising in CD1 are rarely addressed by a second compound from the same series. The difficulty in finding CD2 is also an indication of a lack of clarity within the project on what issues in CD1 really need to be solved. A new series is required and, by the definition of a series given above, this means a real issue has been identified and addressed by a structural change. For example, in the project GPCR2 (Table 1), there are two series (series 4 & 5) in which series 5 contains a structural modification removing a toxicological issue that stopped CD1.

(vi) It might be considered safer to retain a project chemistry team in readiness and making compounds 'just in case'. However, the inability of chemistry teams to disband and regroup in a timely manner suggests that medicinal chemists are not able to pass on knowledge retained by the individual team members effectively.

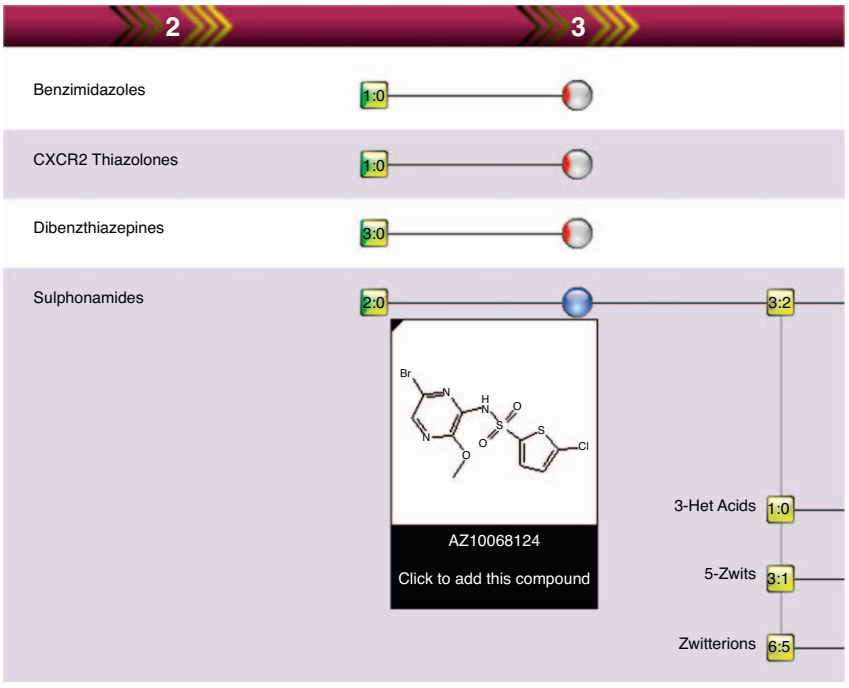
(vii) Some targets were relatively easy to synthesise; for example, Series 7 in Table 1. Here, a Suzuki coupling reaction facilitated the synthesis of numerous analogues. This resulted in a higher number of very close analogues being synthesised before the CD was discovered. The eagerness of chemists to synthesise and submit large numbers of analogues needs to be tempered with the enormous number of possible analogues, and the ruggedness of the structure–activity landscape [6]. It is doubtful whether making targets because synthesis is straightforward is a winning approach.

From these findings, it is reasonable to infer that once approximately 50 compounds have been made and tested, a CD will have been found and, if not, then further synthesis in that series is risky as the probability of finding a CD is reduced. This would have resulted in some of the CDs in this analysis not being made, but it is not possible to say how many other CDs might have been made with the redeployed resources. In addition, not all CDs are equally valuable and too many can cause delays downstream in the development process. The importance of gathering all the data on each compound before starting the next design–make–test cycle will mean that this approach is not necessarily faster, but rather requires fewer chemists, biologists and resources. An advantage is that the savings could be invested in developing more innovative chemistry, running a parallel project or undertaking a more thorough analysis of the screening results before the next round of synthesis.

While collecting the information for this analysis, it quickly became apparent that, after a period of time, medicinal chemists were unsure of all the step changes in structure and their chronological order. This is not surprising given the shortcomings of human memory, but it is also a result of the lack of a mechanism or structure for capturing project chemistry experience. In addition, most biological targets are worked on by multiple chemists over long periods of time, perhaps at different locations within the same company. Most of the relevant information (e.g. plans, structure activity analysis and project reports) about the project and its chemistry effort resides in a paper trail consisting of eRooms (http://en.wikipedia.org/wiki/ERoom; for a recent innovative solution, see [7]), e-mails and meeting presentations; the latter two are particularly poor ways of capturing knowledge and retrieving learning. A few projects are published or reported at conferences, but years can pass before public disclosures are made.

A way to record medicinal chemistry experience

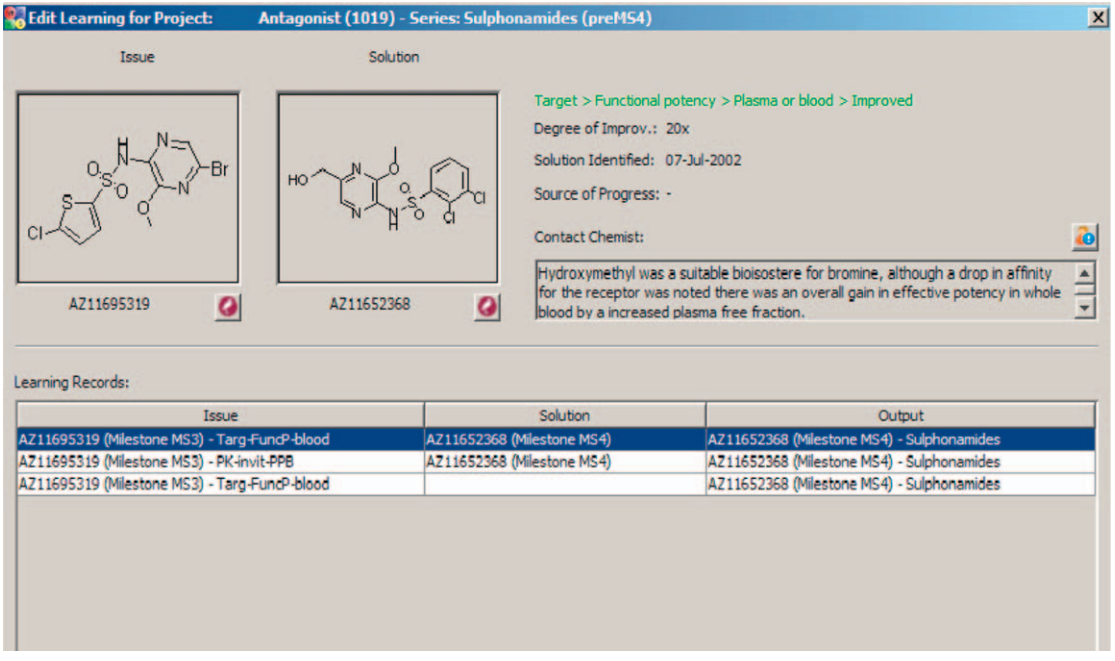
Reanalysing projects for this report took a considerable amount of effort. Perhaps the most striking point learnt from this work is that the CD research was not recorded in a way that enabled retrospective analysis and learning. To try to address the issue of recording the journey of a drug discovery project, an application was commissioned: MeCKS‡. MeCKS enables medicinal chemists to record and share their work by charting the step changes in structure discovered and indexing them against the biological target and project phase. MeCKS does not capture design ideas and hypotheses§; neither does it report screening results [8]; instead, it is a searchable repository for the distilled learning of the chemical programme (compound origins, failures, problems met and their solutions). A comprehensive and hierarchical list of issues met during the drug discovery process enables searching so that the project community can quickly access pertinent information. This carefully structured list of issues is comprised at the highest level from; *in vitro* and *in vivo* absorption, distribution, metabolism, and excretion (ADME), *in vivo* efficacy, molecular and series issues, material (or pharmaceutical) issues, toxicological (*in silico*, *in vitro* and *in vivo*) issues, portfolio and, finally, biological target issues. The application also acts as a directory to facilitate contact between scientists who have solved or faced similar issues; their current contact information is one click away. Changes in key project people, for example lead chemists or project leaders, are also retained.



Drug Discovery Today

FIGURE 2

Screenshot of the project map, also showing a pop-up structure. The bifurcation (bottom right) shows a series (sulphonamides) that provided three new sub-series. The numbers contained within the yellow learning modes indicate the amount of learning; '3:2' indicates three issues, of which two have been solved.



Drug Discovery Today

FIGURE 3

Screenshot of a learning node entry. This learning node contains two solved issues and one unsolved issue.

*Compounds included in the analysis were purified and submitted as solid samples; at least one test result was reported.

†On average, one third of the CD series was published as examples in patents.

‡Known as MeCKS (medicinal chemistry knowledge sharing). The work was a venture with Tessella plc (Technology and Consulting; http://www.tessella.com/).

The screenshot in Fig. 2 shows how medicinal chemistry for a biological target is represented as an interactive map. To my knowledge, this is the first time that such a map has been attempted. Every series investigated against a unique biological target, or combination of biological targets, is listed down the screen in order of data entry; the most recently edited series is displayed at the top. The progress of the project is indicated from left to right across the screen with nodes alternately containing either a collection of 'learning' (yellow squares) encountered en-route or a milestone (blue circles). A 'learning node' is shown in Fig. 3: in this case, a step change is shown that solved an issue, the abbreviated description of which is shown in green.

Computational methods have been reported that examine databases of screening results looking for structural motifs that consistently improve potency ([9] and references therein). This approach could be used to find solutions for other troublesome activities, for example human Ether-à-go-go related gene (hERG) and cytochrome P450 (CYP) inhibition. This could be a great boon to drug hunters. However, a validated solution is revealed only after considerable screening data has been deposited in a database. Also, the approach only deals with a limited number of potential issues and requires consistent data sets. The wide range of issues, for example toxicological studies, material properties and drug metabolism, captured in MeCKS enables the identification of emerging issues. MeCKS then captures emerging solutions that can be tested, thus driving the learning process. Experts could use MeCKS to forewarn a project team of potential issues and gather information for the rapid analysis of emerging issues and solutions. Hyperlinks to the external world can be stored within the application making it a useful organiser for project documentation. Exporting search results in Excel™ compatible formats ensures data transferability. A weakness of MeCKS is the need to encourage medicinal chemists to enter information. However, the majority of that effort is in finding the required information! Some medicinal chemists, who are always looking toward the next compound, are reluctant chroniclers of their own science and encouragement is needed.

Concluding remarks

In conclusion, I offer here a concise definition of a series that makes it useful for organising medicinal chemistry output. A logical name should be assigned, and the series members should be tagged with that logical name in the compound collection. The analysis presented here, in a meaningful number of example projects, shows that continuing synthesis after a CD is discovered is not warranted in a large enough number of projects and that doing so requires careful consideration. I hope that this analysis will give medicinal chemists encouragement to jump more quickly into new chemical space and to resist the temptation to turn over every stone on the beach lest a better compound is underneath. The saving in time can be devoted to design and accessing the synthetically more challenging target molecules that often remain unmade. Considerable effort was involved in collecting the data for this work, and that effort was made more difficult by the absence of an accessible repository of project chemistry experience. I feel that MeCKS will go someway to filling that gap.

How well do medicinal chemists learn from their experiences?

There is no doubt that medicinal chemists have become much more aware of their own failings over the last 10 years; the intense focus on drug-like properties is testament to this. However, in order for medicinal chemistry to move forward, it must fulfil the most basic requirement of good science and record observations in a way that enables others to learn and benefit. When this is done, much more will be revealed, folklore and prejudice [10,11] dispelled, and more progress made. Ideally, applications such as that described here should be made publicly available to enable all medicinal chemists to make available the knowledge gained over countless years of medicinal chemistry research into small molecule drugs.

Acknowledgements

The author would like to acknowledge innovative contributions to MeCKS by Andrew Griffin (AZ Montreal), Dan Fraser, Richard Crags and Giles Turner (Tessella). The author also thanks Andrew Griffin for his helpful comments on the manuscript, and Alan Bell and the staff at Tessella for an enjoyable and fruitful collaboration.

References

1 Leeson, P.D. and Empfield, J.R. (2010) Reducing the risk of drug attrition associated with physicochemical properties. *Annu. Rep. Med. Chem.* 45, 393–407

2 Zhao, H. (2011) Lead optimization in the nondrug-like space. *Drug Discov. Today* 16, 158–163

3 Johnstone, C. *et al.* (2009) Making medicinal chemistry more effective: application of Lean Sigma to improve processes, speed and quality. *Drug Discov. Today* 14, 598–604

4 Hoffmann, T. and Bishop, C. (2010) The future of discovery chemistry: quo vadis? Academic to industrial: the maturation of medicinal chemistry to chemical biology. *Drug Discov. Today* 15, 260–264

5 MacCoss, M. and Baillie, T.A. (2004) Organic chemistry in drug discovery. *Science* 303, 1810–1813;
Macdonald, S.J.F. and Smith, P.W. (2001) Lead optimization in 12 months? True confessions of a chemistry team. *Drug Discov. Today* 6, 947–953

6 Bajorath, J. (2010) Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* 50, 1021–1033

7 Barber, C.G. (2010) 'OnePoint' – combining OneNote and SharePoint to facilitate knowledge transfer. *Drug Discov. Today* 14, 845–850

8 Brosius, A.D. (2009) Project-focused activity and knowledge tracker: a unified data analysis, collaboration, and workflow tool for medicinal chemistry project teams. *J. Chem. Inf. Model.* 49, 2639–2649

9 Warner, J.D. *et al.* (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

10 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896

11 Leeson, P.D. *et al.* (2004) Drug-like properties: guiding principles for design – or chemical prejudice? *Drug Discov. Today: Technol.* 1, 189–195