ELSEVIER



# Modelling iterative compound optimisation using a self-avoiding walk

**John Delaney**

Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire RG42 6EY, United Kingdom

**The optimisation phase is a crucial step in the process of drug development, yet the mechanics of the projects that make it up are poorly understood. Weak documentation of failed projects makes statistical analysis of the factors affecting project performance challenging, so a better approach may be the development of an underlying theory of how projects work. We present a model based on a modified random walk in a relevant chemical space and use it to produce simulations of projects and portfolios of projects. Simulation is used to explore parameters that might affect the performance of a project and shows that they fall into two groups – target and process – that affect the overall performance in distinct ways.**

## Introduction

The problems facing pharmaceutical research have been well documented over the past few years [1]; in essence, while R&D spending has gone up substantially [2], the number of new chemical entities being submitted for regulatory approval has not increased. One response to this has been an increased interest in the actual process of drug research and ways in which it could be made more efficient. This has ranged from quantifying the causes of late stage failures [3] to attempting to apply modern manufacturing process thinking (e.g. six-sigma [4]) to research activities. Most of this work has treated the various stages of research (screening, lead finding, optimisation, pre-clinical and so on) as 'black boxes' [5,6], the inner workings of which are of little concern. The aim of this article is to pry open one of these boxes, optimisation (approximately 14% of the cost of drug development [7,8]) and take a look inside.

The chemical project is a key component of drug and agrochemical research [9] because it is the very stuff of which the optimisation phase is made. Once a lead compound has been identified by screening, patent searching or rational design, a process of development is applied where compounds similar to the lead are synthesised and tested in an iterative cycle [10]. The hope is that the project learns more about the biological target with successive iterations, enabling better compounds to be designed for the next round of testing. This learning process has been extensively studied, because it is at the heart of our ideas about the relationship between structure and activity (SAR) [11], and at times there is a feeling of 'manifest destiny' [12] about the whole practice. This sense is not born out in reality, where many projects never reach a satisfactory conclusion, such as promotion to a development candidate [13]. Experienced synthetic chemists are all too familiar with series that never capture the level of activity of the lead, or which seem to hit a plateau of activity that defies improvement. The nature of the target changes as the project evolves because more factors come into play beyond raw *in vitro* activity – *in vivo* availability, pharmacokinetics, toxicity, product formulation and clinical efficacy; all potentially stand in the way of a project yielding a marketable product.

The disjunction between SAR vision and practical project outcomes is best understood by considering what scientific criteria terminate a project – simplistically, a project halts when a sufficiently good compound is found for promotion or when hope for producing a better compound evaporates. An active project remains active precisely because none of its members has met the criteria for progression to development. SAR works best when interpolating within its own data set [14,15], and extrapolation (implied by the need for an improved molecule) is much harder. Because the nature of projects emphasises extrapolation, SAR necessarily struggles [16,17] as

*E-mail address:* john.delaney@syngenta.com.



1359-6446/06/$ - see front matter  doi:10.1016/j.drudis.2008.10.007

it explains where a series has been rather than where it should go next.

We have attempted to analyse in-house projects to investigate whether there were any differences between successful and unsuccessful projects, but this has highlighted some problems. A study of 50 randomly selected early stage Syngenta projects, conducted in 2003, found that nearly 30% of them had never been formally closed – failure could only be inferred from scrutiny of the date the project document was last accessed (12 months of inactivity was assumed to indicate failure). This illustrates a general point – we are better at documenting successes than failures [18]. It is also difficult to ascribe simple, numerical project parameters to series. 'Project parameters' could include how strong a lead needed to be to initiate a project, how strong the SAR was around the lead or how long the project was allowed to run-on for without getting closer to its target. Knowing which values are important and getting a quantitative idea of what they are is essentially impossible given the way the project information is currently recorded.

Because the empirical data for real projects are rather sketchy, a better approach might be to model them using an underlying theory of their behaviour. If such a description could be devised, the door would open to producing simulations of projects, potentially allowing project parameters to be explored in a controlled way. A model would also provide a context, allowing a better definition of the data needed to allow meaningful statistical studies of the projects to be carried out.

What kind of properties does a project (as opposed to a set of compounds) have anyway? The crucial distinction is temporal – a project is a time-ordered series of compounds, with properties that evolve over the course of the project [19]. The idea that compound properties change as a project advances is not new; for example, recent ideas about pharmaceutical 'lead-likeness' are based on the observation that drugs become, on average, larger (increase in molecular weight) as they move from lead to product [20,21]. This might be regarded as a simple project trajectory – a path through a descriptor space that varies with time. Molecular weight is a simple molecular property and it would be interesting if the idea of temporal change could be usefully applied to more complex descriptors.

This article will propose that the idea of a project trajectory through a complex descriptor space is reasonable [22] and useful given the right representation of chemical space, that this trajectory can be modelled using an elaboration of a standard random walk (RW - see box 3 for glossary of abbreviations) and that these walks can be used as the basis for simulating groups of projects in a chemistry department. Some simple experiments with these project models show that they can exhibit realistic behaviour and provide some insight into the fundamentals of project management.

## Representing compound series for projects

An ideal chemical descriptor for this work needs to be sufficiently complex to distinguish between closely related analogues in a typical project, produce chemically sensible, quantitative similarities between compounds and have some relevance to biological activity (the main driver for most projects). Substructural fingerprints [23–26] provide a rich way of describing compounds for applications that depend on calculating the pairwise similarity between molecules and these similarities have been shown to correlate with biological activity [27–30]. Their high dimensional nature (typically 1024 bits) precludes their direct use as a way of visualising chemical space, but dimension reduction methods such as Sammon mapping [31] can be used to produce a projection suitable for display. Sammon mapping is designed to reproduce inter-compound similarities with little distortion in the final map [32], preserving the correlation with activity – active compounds tend to cluster together in a Sammon map produced from fingerprint similarities [33]. The temporal dimension of a project can be introduced into a static map by time-ordering the points and applying a moving average [34] to the coordinates in each dimension. A moving average takes a fixed length window on a sequence of numbers, averages the values in that window, moves the window on by one step and repeats.

The trajectory is this series of averages for each dimension (a window size of 50 was used for this work). The number of dimensions needed to capture the behaviour of a series of related compounds varies depending on the chemistry and the nature of the target. The work on a small number of in-house projects has shown that an *in vivo*, herbicidal biological response can be adequately captured using two- to four-dimensional maps (i.e. the clustering of compounds with similar biology does not increase by adding more dimensions to the Sammon map). We have settled on 3D maps as a reasonable compromise, accurate enough in most cases and suitable for visual display.

The plots for real projects resemble protein chains (Fig. 1) – extended, writhing patterns in space. A closer examination, including a detailed statistical analysis of distributions of internal coordinates [35] for 63 real project trajectories, reveals a structure similar to a RW, an analogy we shall explore in the next section.

## Self-avoiding walks

A RW is created by moving a point in discrete jumps in a random direction [36], creating a trajectory from successive random steps. RWs have been applied to many physical, chemical and economic processes that involve some element of chance – Brownian motion [37], polymer chain dynamics [38] and share price fluctuations [39,40] being the notable examples. Indeed, the simplest reasonable model of project trajectories in a descriptor space would be a RW. This makes no assumptions about the project knowing where it is going (if it did it would go straight to that point), just that it has to move on at each step. Each new compound is usually closely related to the last (projects do not hop randomly around the entirety of chemical space) but the position of the target is unknown – the project only knows it has hit its target after the successful compound has been made and tested (Fig. 2).

A self-avoiding walk [41,42] (SAW) is a specific example of the general RW, adding the constraint that the walk cannot intersect with itself. This is reasonable as pharmaceutical companies spend a lot of money on IT infrastructure to ensure that this constraint is met. Newly synthesised compounds are registered in a database [43] to make sure that precious chemistry resource is not wasted making the same compound twice. Self-avoidance makes the walks search for a solution more efficient. The difference between a RW and a SAW is shown in Fig. 3 (for the remainder of this article we shall consider walks with fixed length jumps on a square or on a cubic grid because this simplifies the computer generation of

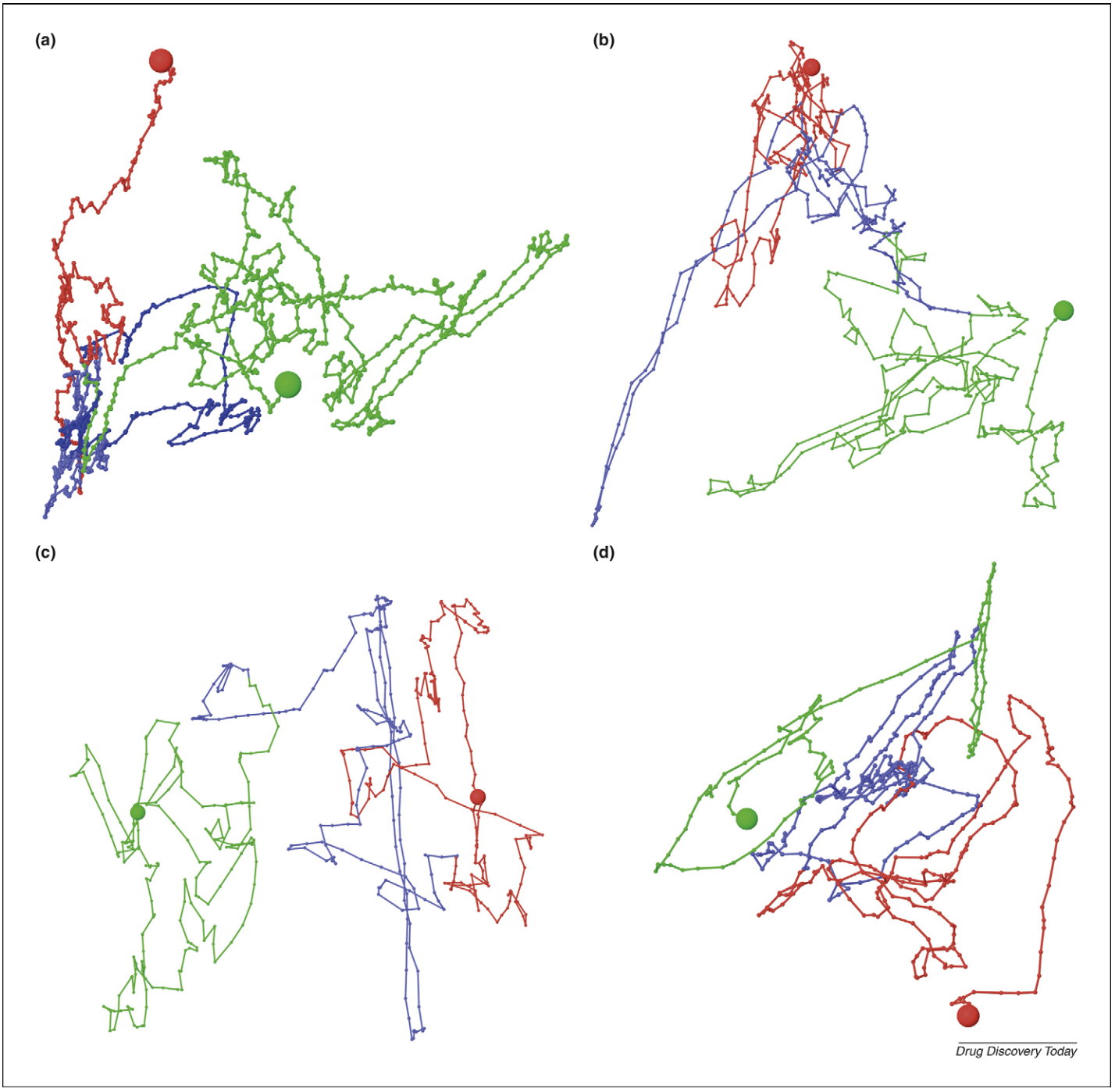


**FIGURE 1**

Four 3D moving average trajectories derived from real optimisation projects (two herbicide (a and b), one insecticide (c) and one fungicide (d)). The earliest compounds in each sequence are coloured red, mid-sequence blue and end-sequence green. The start and end of each trajectory are marked by an enlarged ball.

walks) – both walks have 21 steps, but in the pure RW the same sites are hit repeatedly (RW only visits 9 sites in 21 steps), whereas all 21 steps of the SAW are unique.

The SAW has a more thread-like structure with more widely separated end points. A synthetic three-dimensional SAW is shown in Fig. 3 and makes an interesting comparison to the real project trajectories as shown in Fig. 1.

The analogy between a SAW and an optimisation project becomes clearer if we consider how features of both map onto each other. The start of the walk corresponds to the lead compound in a project, the target (a set distance from the start of the walk) represents the state the project needs to reach to progress and the distance from the end of the walk to the target is the current state of the project as measured by its biological activity or activity in combination with other experimental data (e.g. solubility). A synthetic SAW has a limited number of well-defined control parameters which have clear analogies to optimisation project decisions – for example 'stop an optimisation project if a compound with better biology is not found in the next 50 synthesised molecules' becomes 'terminate the walk if it does not come

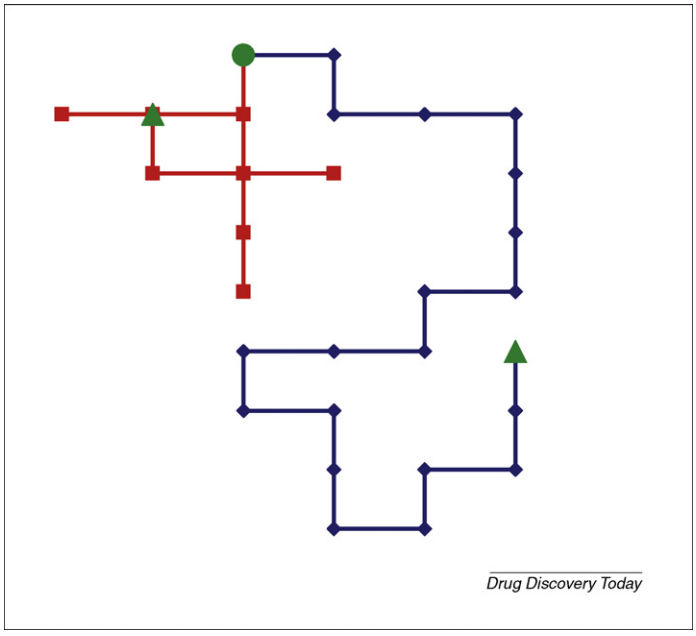


**FIGURE 2**

The simple random walk is coloured red, the self-avoiding walk blue. The start of the walks is marked with a green circle, and the end with a green triangle.

closer to the target in the next 50 steps'. A SAW has only two ways of ending – a successful project/walk is one that finds its target, while a failure is a walk that fails to get closer to its target quickly enough and is terminated.

## Simulating compound series

If one accepts SAWs as model for project behaviour, then it becomes reasonable to produce computational simulations of individual projects on a simple cubic grid. Project simulation can be done by generating multiple SAWs, applying control parameters to them and counting up the number of successes and failures.

**BOX 1**

1. Distance from starting point to target: How strong is the initial lead? Varies between 1 (a strong lead) and 15 (a weak lead).
2. Target size: How easy is the target to hit? Size corresponds to the radius of the target (size = 0 indicates a single point in space as a target, ranges up to 10).
3. SAR strength around target: Can the project exploit SAR information? The closest next possible step to the target is included twice in the random draw on a proportion of the steps. This gives a bias to the walk ranging between 0 and +67%.
4. Project run-on: How long can a project run without producing an improved compound? Values range between 1 (a very aggressive cut-off) and 400 (projects can run on for a long time without improvement).

A simple simulation was set up with the four control parameters shown in Box 1. Ten thousand walks were initiated with three of the four parameters held constant while the other was systematically varied. The number of successful projects (SAW reaches target) and failures (termination of SAW before reaching target) was recorded together with the total number of SAW steps (each step being a unit of synthetic chemical effort). These figures can be combined in different ways, the most sensible being the number of successes *per* step (SPS), the total number of projects (successes + - failures) *per* step (PPS) and the number of successes *per* project (SPP). Each figure highlights the different aspects of the process (discussed below), but it is SPS that seems to capture the most

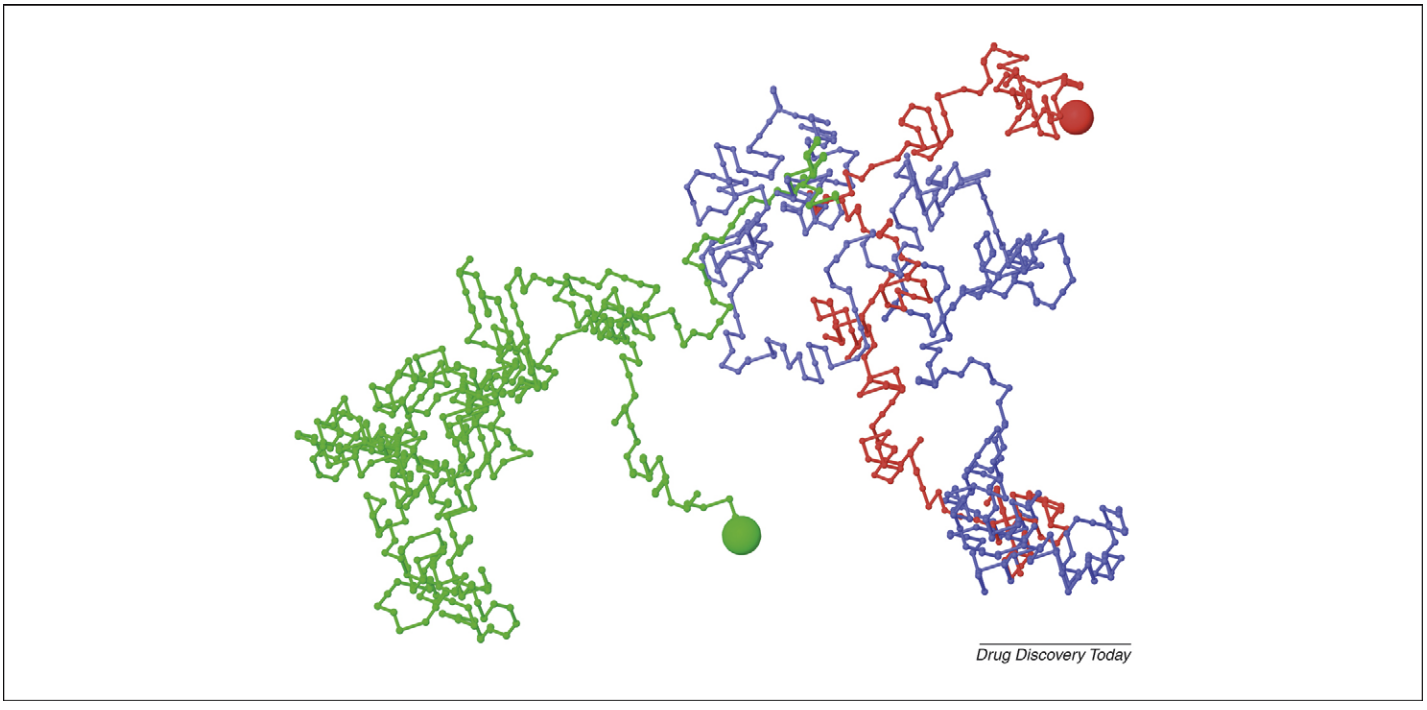


**FIGURE 3**

A SAW generated computationally on a 3D grid using a pseudo-random number generator. The walk was set to generate 1000 points before halting. The points are marked and colour-coded as in Fig. 2.

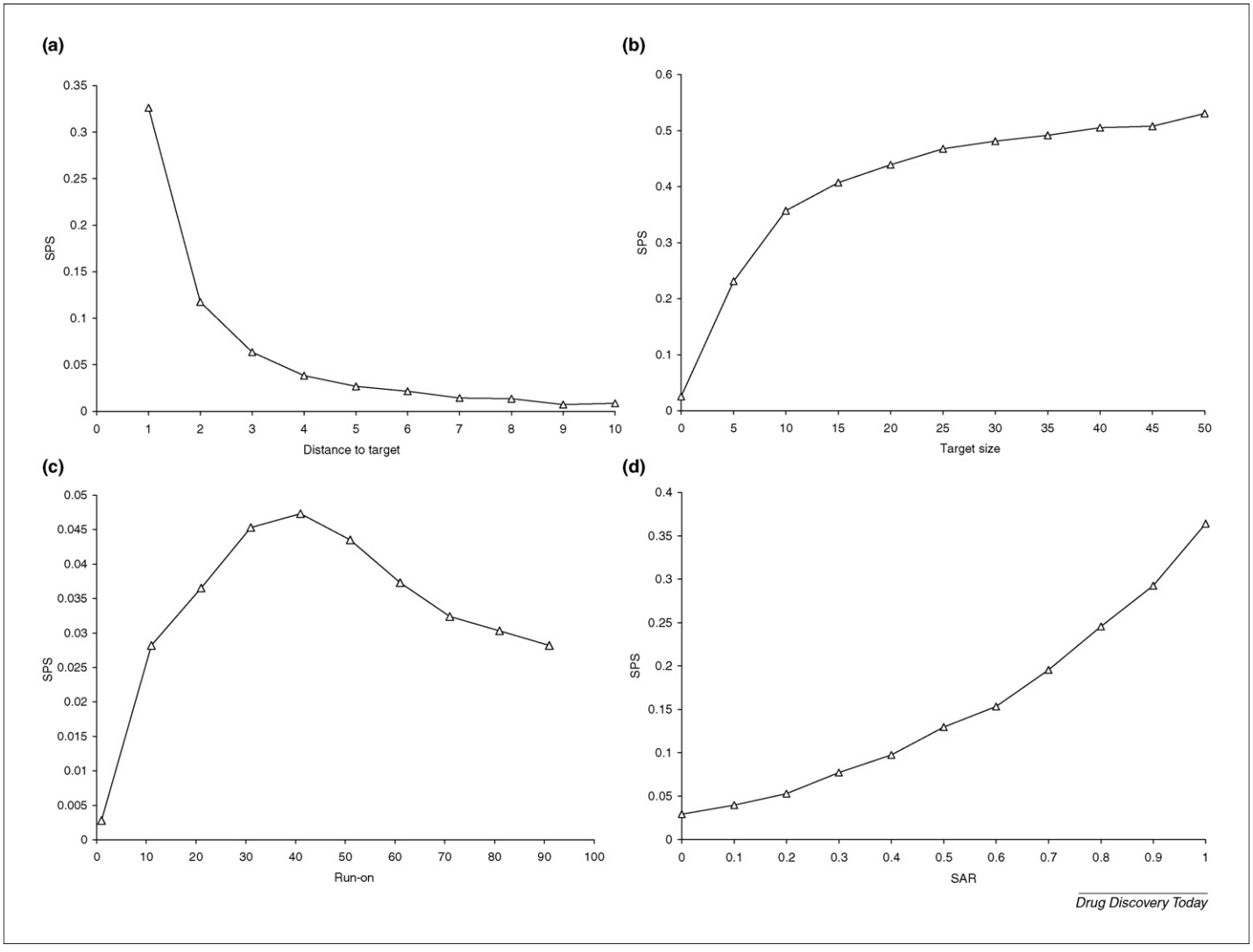


**FIGURE 4**

Graphs showing the changes in the number of successful projects per SAW step (SPS) plotted against (a) distance to target (lead strength), (b) target size (easy/hard lead), (c) run-on (how long the project is allowed to continue without getting closer to its target) and (d) SAR (degree of structure activity bias, i.e. applied to the walk).

crucial aspect of performance in projects – successes for a given amount of synthetic effort.

The effects of each of these parameters in isolation on SPS are shown in Fig. 4.

SPS behaves monotonically with respect to the three target based parameters. The number of successful projects drops away exponentially with target distance, shows a more linear dependence on SAR and a saturating effect with increased target size (radius) – the target effectively changes from a point to a plane with respect to the starting point (set at distance = 5), and further increases have less effect.

The process-based parameter (run-on) is more interesting because it shows a clear peak, indicating a trade-off between two effects. What seems to be happening is that, initially, increasing run-on allows a greater proportion of projects to succeed and this offsets the effect of allowing projects that eventually fail to use up more steps/effort. Because run-on is increased further the second effect overwhelms the first and the SPS declines.

These parameters can also, potentially, interact with each other and, because chemistry departments usually run several projects concurrently, with parameters connected with the structure of the department. The individual simulations can be combined to create a 'virtual' chemistry department using discrete event simulation [44], which allows several projects to be run in parallel with resource constraints. A simulation runs as a series of parallel jobs that acquire, hold and release resources, events being recorded against an internal clock. Each job competes for resources, waiting if necessary. In this case the discrete event is the execution of a project – the project holds on to its resource (chemistry man hours) until it reaches a conclusion (success or failure). The time taken for the project to do this (the number of steps in the SAW divided by the amount of chemistry resource assigned to the project) is fed back to the simulation which handles the book-keeping aspects of the simulation (Fig. 5).

SimPy (http://simpy.sourceforge.net – a Python programming language extension) allows discrete event models to be easily

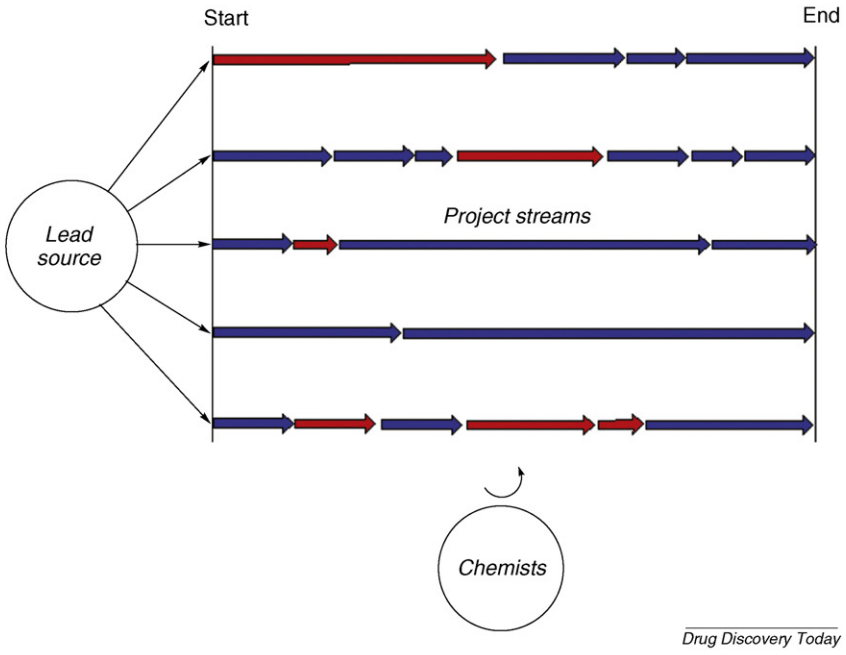


**FIGURE 5**

A schematic diagram of a discrete event simulation of a chemistry department. Leads are assigned to one of the five project streams, initiating a project that runs until it hits its targets or fails (red = success and blue = failure). The simulator shares out chemists to each stream. In this case, the lead source is assumed to be prolific and no stream is ever left waiting for a fresh lead when its current project completes.

created by handling synchronisation and resource allocation issues allowing more complex systems to be built up – for example, a portfolio of projects with different numbers of chemists working on each.

The departmental simulation introduced three additional parameters shown in Box 2.

To fully explore a reasonable range of combinations, the parameters were assigned random values at the start of each simulation run (Monte Carlo) and 3000 runs were performed.

The simulations highlighted some interesting effects. There was interaction between some of the parameters, particularly at extreme values. The most noticeable was between target distance and project run-on – allowing projects to continue for longer in the face of no improvement tends to increase the SPP (projects that would have been chopped manage to turn around and hit their target), while reducing the overall throughput of projects, PPS. The magnitude of this effect depends on the target distance – for example, combining a weak lead with an aggressive run-on cut-off dramatically reduces the SPS because it becomes impossible for any walk to generate improvement at a sufficient rate to avoid being axed.

**BOX 2**

1. Number of chemists in the department: more chemists allow more compounds synthesised in a given time. Values range between 10 and 100.
2. Number of project streams: the number of independent projects that a department runs in parallel. Ranges between 1 and 10.
3. Fixed cost of project initiation: setting up a project is not instantaneous – delays can be caused waiting for reagents to arrive, personnel to be redeployed, among others. Ranges between 0 (no delay) and 10 (a substantial wait to start each project).

The departmental structure parameters showed little effect on SPP, but a marked influence on PPS – essentially, the project throughput – could be improved by, unsurprisingly, adding more chemists or more project streams as well as by reducing the fixed cost of setting up a project. None of this affected how probably any one project was to succeed, but increased throughput did boost the overall SPS figure, illustrating the important point that SPS is the product of SPP and PPS. One interesting interaction between the number of streams and fixed cost was also noted, because for small number of streams the effect of changes to fixed cost became amplified – a larger number of streams seemed to dilute the impact of increased fixed costs. This is reasonable because, *in extremis*, a single project stream has to halt completely as each project starts up, whereas in multiple streams some work can continue most of the time.

The most interesting result from aggregating all of the simulations is the separation of the parameters into two roughly disjoint groups. This becomes clear when SPP is plotted against PPS and the points are colour-coded by a parameter value. Some parameters show a strong variance with SPP, and others with PPS. Plots for SAR and project run-on are shown in Fig. 6.

The effects of the control parameters are summarised in Fig. 7.

The crucial point to bear in mind is that the SPS is the key figure of merit and it is the product of these two effects. This suggests that there are at least two broad paths to improved performance – focus on operational excellence to improve the process side of things, or

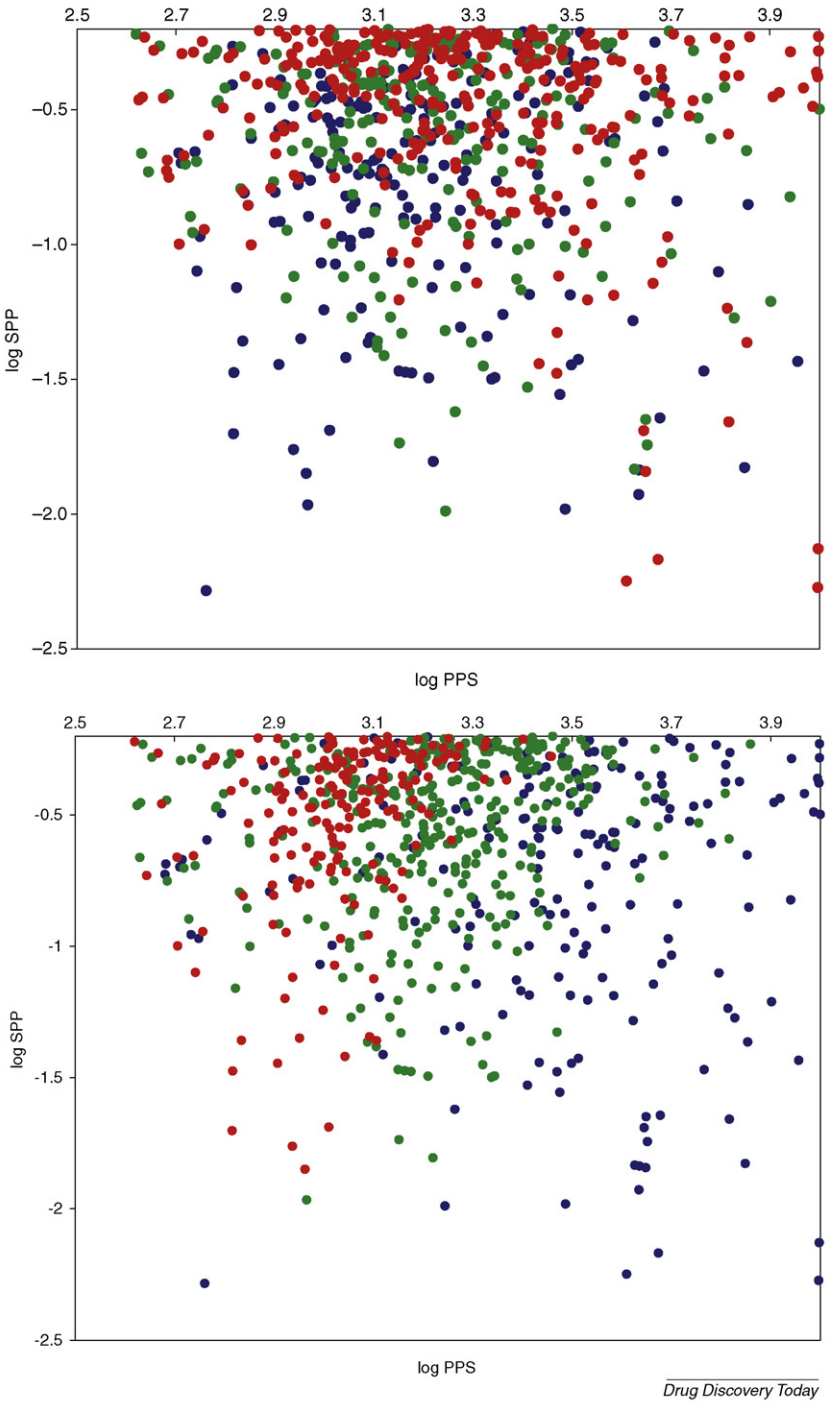


**FIGURE 6**

Results for 3000 simulations with randomly assigned parameters. The logged values of SPP and PPS are plotted against each other and the points colour-coded by SAR (left) and run-on (right). Red points denote a high parameter value, moving through the spectrum of orange, yellow, green and blue/purple for low values. SPP seems to vary more strongly with SAR, PPS with run-on.

pick and choose your starting points to make the target easier to hit. The results shown in Fig. 4 indicate that target distance has the most drastic effect on SPS which, in turn, suggests that focusing on the strength of the initial lead will bring the most dramatic improvements in the overall process. Indeed, the extent to which target/lead combinations are given in an organisation may limit the overall scope for improvement. There are also some potentially nasty combinations of parameters, particularly long target distance/short run-on and small number of project streams/high fixed cost, which should be avoided. On a more positive note, the simulations show that there is some reward for improving fixed costs and obtaining better SAR information, even if neither constitutes a knockout blow to the problem of attrition.

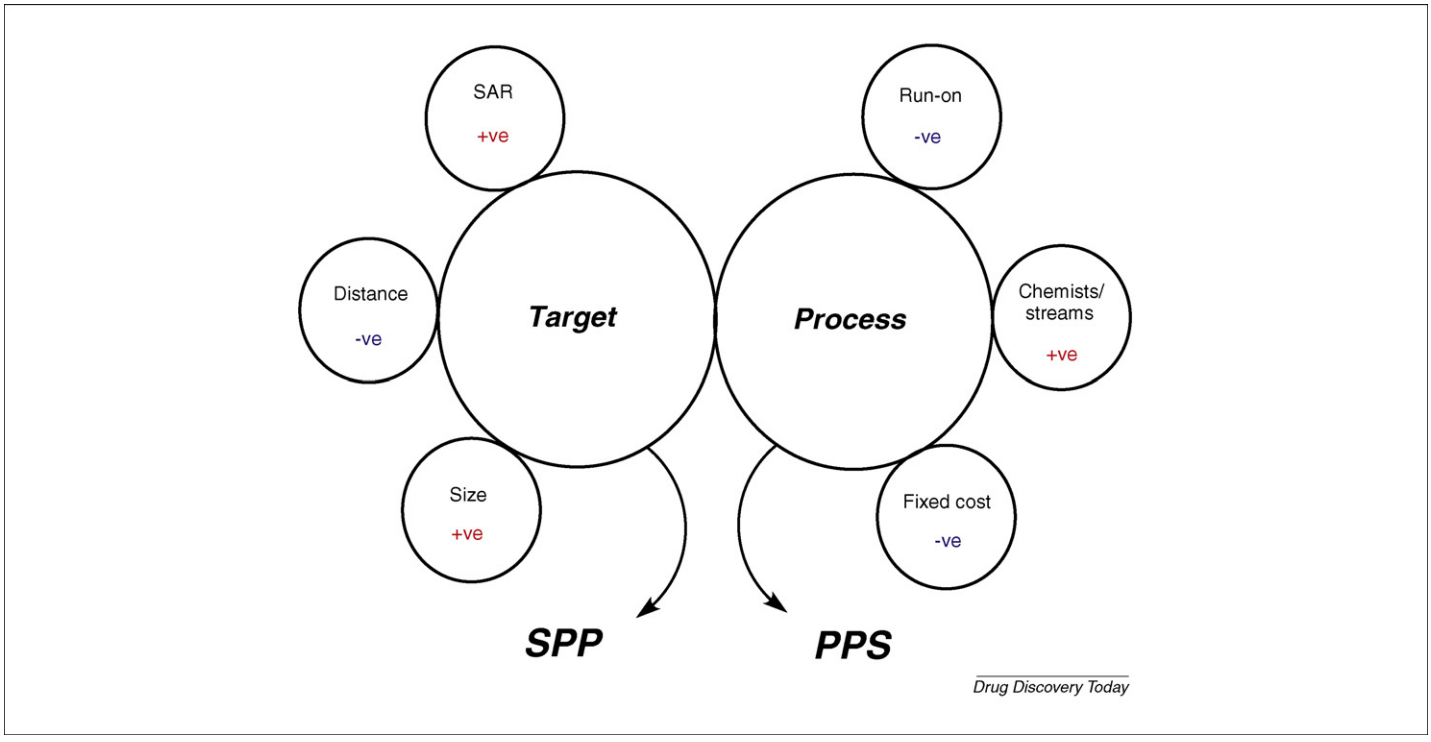


**FIGURE 7**

The simulation parameters fall into two groups – those broadly related to the properties of the target such as lead strength (distance) and those related to decisions taken by the chemistry department (process) such as how long to allow projects to continue in the face of no improvement. The target group mainly affect SPP while the process parameters are more important for PPS. The red and blue + ve/−ve signs denote whether the parameter positively or negatively correlates with SPP/PPS.

## Conclusions

This article proposes that projects can be modelled using a SAW as it seems to be the simplest model that manages to capture the essential behaviour of an optimisation project. None of the evidence gathered from real projects suggests that anything more complex is required. Accepting this opens the way to simulating projects and collections of concurrent projects run within a chemistry department. The simple simulations presented here indicate that the factors controlling the success of such a department fall into two broad categories, target and process.

The visualisations of real projects may be of use in themselves, particularly if the points along the path were colour-coded by activity. This might allow hotspots along the timeline to be identified and those areas of chemistry revisited, particularly when a project seems to have ended up in a cul-de-sac. At the very least it provides an interesting summary of a whole project because it evolves through time which is novel.

More elaborate RWs, incorporating variations in step length, might provide a more realistic picture of how projects travel through chemical space. A particularly interesting option would be a Lévy flight [36] which allows occasional large jumps through space, sometimes seen in projects when a larger piece of chemical functionality is introduced into a series. Another aspect of the model which could be altered is to allow the walk to branch. Projects, finding themselves at a dead-end, could return to an earlier, more active compound and use this to branch off in a different direction.

How should one interpret a parameter like 'distance to target' in the context of real projects? The change in biological response with chemical distance can be readily determined by plotting activity against distance from a point (distance seems to be proportional to pIC50 and other bio-relevant physical properties [45]). Target really means the right state (or states) which a project must attain to progress to the next stage of development. In the context of a drug this might mean identifying a compound suitable for a pre-clinical trial. The strength of leads within a department is usually characterised by a simple measure of biological response. In principle, it should be possible to use the number of progressions from leads of various strengths to fit the model to the data. This raises the problem mentioned in the introduction of having poorly documented project outcomes, though having a well-defined model should cut down the amount of data required to produce a fitted model and may help to clarify the questions we need to ask to improve this.

In the absence of a truly quantitative model, the trends themselves are useful. The sensitivity of the proportion of projects that succeed to lead strength is striking and should give pause to anyone working on weak leads, as should the interaction of this parameter with very short run-ons. The different effects of target and process are also interesting. One avenue that could be readily explored through simulation is how best to assign chemists to projects – given several projects running in parallel and in different states with respect to their targets, should one place larger numbers of chemists on weak projects, or pile them into the stronger ones? This problem has some striking similarities to optimal betting strategies in gambling, particularly the Kelly criterion [46] which relates to how much a gambler should wager on an event based on his bankroll, probability of winning and the odds

being offered. The Kelly criterion attempts to maximise the long-term return over a long sequence of bets and it would be interesting to see if maximising the Kelly return across parallel projects is more efficient than assigning fixed numbers of chemists to each project. Another potentially interesting simulation would be a 'patent race' where two SAWs are allowed to interfere with each other in a race to the target – each has to avoid the others path as well as their own. A reasonable question might be how much more chemistry resource is needed to overcome starting behind a competitor relative to operating in clear IP?

It is interesting to contrast this work with the molecular modelling/QSAR techniques that have been applied to chemical series in the pharmaceutical industry over the past 20 years [47]. Molecular modelling attempts to reduce the randomness in the development of bio-active compounds by predicting which molecules should be made next to get the project closer to its target. This approach attempts something fundamentally different – it takes randomness as given and tries to transform 'uncertainty' (randomness with unknown probabilities) into 'risk' (randomness with known probabilities) [48]. A properly parameterised version of the model would allow the value of a project to be continually assessed and compared. This creates an opportunity to exploit financial engineering [49] techniques such as project insurance or real options [50]. Large pharmaceutical companies tend to diversify their portfolio of projects, spreading the risk. Could smaller companies do the same by pooling their less diversified risks using some form of project insurance, with well-defined risk? Real options could be incorporated into the decision to allow a project to continue – again, a good estimate of risk allows a better estimate of the balance between abandonment and the chances that the next compound will move the project nearer its goal. Ultimately a better understanding of the underlying mechanics of the optimisation project should allow us to run portfolios of projects more successfully (Box 3).

**BOX 3**

Glossary

**RW** – Random walk, a random process consisting of a sequence of discrete steps of fixed length.

**SAW** – Self-avoiding walk, a variant of the standard random walk on a lattice that does not allow the walk to return to any previously visited points.

**SPS** – Success per step. The number of walks that reached their target during a simulation run divided by the total number of steps performed during the whole simulation. Analogous to the number of successful LO chemistry projects produced for a given amount of synthetic resource (chemist hours).

**SPP** – Success per project. The number of walks that reached their target divided by the total projects initiated during the simulation. Of all the LO projects started, how many were successful in progressing to the next stage of development?

**PPS** – Projects per step. The total number of walks initiated divided by the total number of simulation steps. The total number of LO projects started for a given amount of resource.

## Acknowledgements

The author would like to thank Steven Chalk, Berengere Esparcieux, Natalia Wieckowska, Chris Wood, Ann Stainforth and Tom Sheldon for their contributions to this work.

## References


1 Drews, J. (2003) Strategic trends in the drug industry. *Drug Discov. Today* 8, 411–420
2 DiMasi, J.A. *et al.* (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185
3 DiMasi, J.A. (2001) Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* 69, 297–307
4 Ullman, F. and Boutellier, R. (2008) A case study of lean drug discovery: from project driven research to innovation studios and process factories. *Drug Discovery Today* 13, 543–550
5 Schmid, E.F. and Smith, D.A. (2002) Should scientific innovation be managed? *Drug Discov. Today* 7, 941–945
6 Rooney, K.F. *et al.* (2001) Modelling and simulation in clinical drug development. *Drug Discov. Today* 6, 802–806
7 Zimmerman, Z. and Golden, J.B. (2004) The return on investment for ingenuity pathways analysis within the pharmaceutical value chain. *Life Science Insights White Paper*
8 Rawlins, M.D. (2004) Cutting the cost of drug development? *Nat. Rev.: Drug Discov.* 3, 360–364
9 Delaney, J.S. *et al.* (2006) Modern agrochemical research – a missed opportunity for drug discovery? *Drug Discov. Today* 17–18, 839–845
10 Hol, W.G.J. (1986) Protein crystallography and computer graphics toward rational drug design. *Angew. Chem. Int. Ed.* 25, 767–852
11 Obrezanova, O. *et al.* (2008) Automatic QSAR modeling of ADME properties: blood–brain barrier penetration and aqueous solubility. *J. Comput. Aided Mol. Des.* 22, 431–440
12 John O'Sullivan, (1845) Annexation. *United States Mag. Democratic Rev.* 17, 5–10
13 Garnier, J.-P. (May 2008) Rebuilding the R&D engine in big pharma. *Harvard Bus. Rev.* 86, 69–76
14 Jaworska, J. *et al.* (2003) Summary of the workshop on regulatory acceptance of QSARs. *Environ. Health Perspect.* 111, 1358–1360
15 Weaver, S. and Gleeson, M.P. (2008) The importance of the domain of applicability in QSAR modelling. *J. Mol. Graph. Model.* 26, 1315–1326
16 Johnson, S.R. (2007) The trouble with QSAR (or How I learned to stop worrying and embrace fallacy). *J. Chem. Inform. Model.* 48, 25–26
17 Doweyko, A.M. (2008) QSAR – dead or alive? *J. Comput. Aided Mol. Des.* 22, 81–89
18 Ormerod, P. (2005) *Why Most Things Fail: Evolution, Extinction and Economics*. Faber and Faber
19 Ullman, F. and Boutellier, R. (2008) A case study of lean drug discovery: from project driven research to innovation studios and process factories. *Drug Discov. Today* 13, 543–550
20 Teague, S.J. *et al.* (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* 38, 3743–3748
21 Hann, M.M. and Oprea, T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.* 8, 255–263
22 van Deursen, R. and Reymond, J.-L. (2007) Chemical space travel. *ChemMedChem* 2, 636–640
23 Daylight Chemical Information Systems, Inc., 120 Vantis – Suite 550 – Aliso Viejo, CA 92656
24 Lewis, R.A. *et al.* (2000) Computer-aided molecular diversity analysis and combinatorial library design. In *Reviews in Computational Chemistry* (Vol. 16) (Lipkowitz, K.B. and Boyd, D.B., eds), Chapter 1, pp. 1–51, Wiley–VCH, ISBN 0471386677
25 Flower, D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inform. Comput. Sci.* 38, 379–386
26 Willett, P. *et al.* (1998) Chemical similarity searching. *J. Chem. Inform. Comput. Sci.* 38, 983–996
27 Martin, Y.C. *et al.* (1995) Similarity and cluster-analysis applied to molecular diversity. *Chem. Abstr.* 209 3-COMP

28 Delaney, J.S. (1996) Assessing the ability of chemical similarity measures to discriminate between active and inactive compounds. *Mol. Divers.* 1, 217–222
29 Patterson, D.E. *et al.* (1996) Neighborhood behaviors – a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* 39, 3049–3059
30 Martin, Y.C. *et al.* (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.* 45, 4350–4358
31 Sammon, J.W., Jr (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* C-18, 401–409
32 Kowalski, B.R. and Bender, C.F. (1972) Pattern recognition. A powerful approach to interpreting chemical data. *J. Am. Chem. Soc.* 94, 5632–5639
33 Clark, R.D. *et al.* (2000) Visualizing substructural fingerprints. *J. Mol. Graph. Model.* 18, 404–411
34 Chatfield, C. (2003) *The Analysis of Time Series: An Introduction.* Chapman & Hall/CRC
35 Baker, J. and Hehre, W.J. (1991) Geometry optimisation in Cartesian coordinates: the end of the *Z*-matrix? *J. Comput. Chem.* 12, 606–610
36 Viswanathan, G.M. *et al.* (1996) Lévy flight search patterns of wandering albatrosses. *Nature* 381, 413–415
37 Einstein, A. (1956) *Investigations on the Theory of Brownian Movement.* Dover
38 Flory, P.J. (1969) *Statistical Mechanics of Chain Molecules.* Interscience
39 Malkiel, B.G. (1973) *A Random Walk down Wall Street: The Time-tested Strategy for Successful Investing.* W.W. Norton
40 Paulos, J.A. (2004) *A Mathematician Plays the Market.* Penguin Books Ltd.
41 Hayes, B. (1998) How to avoid yourself. *Am. Sci.* 86, 314–319
42 Madras, N. and Slade, G. (1993) *The Self-avoiding Walk.* Birkhäuser
43 Miller, M.A. (2002) Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* 1, 220–227
44 Banks, J. *et al.* (2005) *Discrete-event System Simulation* (4th edn), Pearson Education
45 Brown, R.D. and Martin, Y.C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inform. Comput. Sci.* 37, 1–9
46 Kelly, J.L. (1956) A new interpretation of information rate. *Bell Syst. Tech. J.* 35, 917–926
47 Richon, A.B. (2008) Current status and future direction of the molecular modeling industry. *Drug Discovery Today* 13, 665–669
48 Knight, F.H. (1921) *Risk, Uncertainty and Profit, Hart, Schaffner & Marx.* Houghton Mifflin Company
49 Luenberger, D.G. (1997) *Investment Science.* Oxford University Press Inc.
50 Villiger, R. and Bogdan, B. (2004) Real options are neither complicated nor irrealistic. *Drug Discov. Today* 9, 552–553