**Griffen, EJ, Dossetter, AG, Leach, AG and Montague, S**

**Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence?**

# Can we accelerate medicinal chemistry by augmenting the chemist with Big Data and artificial intelligence?

Edward J Griffen*, Alexander G Dossetter*, Andrew G Leach+ and Shane Montague*

* MedChemica Ltd, Biohub, Alderley Park, Macclesfield, Chesire SK10 4TG United Kingdom
+ Liverpool John Moores University, School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF , United Kingdom.

Corresponding author: Griffen, E. J. (ed.griffen@medchemica.com, +44 1625 238843)

ORCID ids:

| | |
|---|---|
| Edward J Griffen | 0000-0003-0859-554X |
| Alexander G Dossetter | 0000-0002-4181-3193 |
| Andrew G Leach | 0000-0003-1325-8273 |
| Shane Montague | 0000-0003-3211-6422 |

Keywords: Medicinal chemistry, lead optimization , AI, big data, MMPA

## Teaser(25 words)

AI comes to lead optimization: medicinal chemistry in all disease areas can be accelerated by exploiting our pre-competitive knowledge in an unbiased way.

## Abstract(100 words)

It is both the best of times and the worst of times to be a medicinal chemist. Massive amounts of data combined with machine learning / artificial intelligence (AI) tools to analyse it can increase our capabilities. However, drug discovery faces severe economic pressure and a high level of societal need against very challenging targets. We show how improving medicinal chemistry by better curating and exchanging knowledge can contribute to improving drug hunting in all disease areas. Securing intellectual property is a critical task for medicinal chemists, however it impedes the sharing of generic medicinal chemistry knowledge. Recent developments enable sharing knowledge both within and between organizations while securing intellectual property. Finally the effects of the structure of drug discovery's corporate ecosystem on knowledge sharing is explored.

## Critical Glossary

| | |
|---|---|
| Matched Molecular Pair Analysis (MMPA) | a technique for chemical structure activity or property analysis where single precise changes to molecule structures are identified and the effect of the same change (described as a *transformation*) is studied across multiple pairs of molecules to identify transformations that are transferable between chemical series[1] |
| Quantitative Structure Activity Relationship models | Mathematical models built from descriptors of individual molecule structures that aim to predict biological or physicochemical properties of new proposed molecules[2,3] |
| Econometrics | mathematical economics[4] |
| Publication bias | bias in research where the outcome of the research biases the public disclosure, a particular problem in using public data sets to infer knowledge[5,6] |
| Hierarchy of evidence | a ranking of different types of study when making decisions, with the highest level being least prone to systematic bias and the lowest level being anecdotes or expert opinion[7] |

The goal of this paper is to show how it is possible to accelerate drug discovery by analyzing, systematizing and sharing medicinal chemistry knowledge in an unbiased and coherent way. This is at the core of AI approaches based on supervised learning. The fall in productivity of drug discovery is reviewed and the role medicinal chemistry has to play addressing this issue is highlighted. We discuss how the human aspects of working within a drug discovery team impact on the practice of medicinal chemistry and how using more evidence-driven approaches can counter-balance natural human cognitive biases. The central part of this work shows how application of modern approaches to systematizing knowledge in an unbiased way can both extract new knowledge and circumvent the confidentiality issues created by the need to generate intellectual property(IP). Finally, we discuss the corporate challenges and benefits to global drug discovery in sharing medicinal chemistry knowledge broadly between large Pharmaceutical companies and more widely with the academic, not-for-profit and biotech sectors.

## The Central Challenge for Chemists in Drug Discovery

Drug discovery is facing severe economic stress against a background of increasing societal need. The output of global drug discovery has held surprisingly constant with a median of 16 NMEs launched per year between 1950 - 2014. Taking a straightforward definition of productivity in drug discovery as the number of NCE's brought to market divided by cost, longitudinal analysis shows an average annual increase in cost of 8% per annum – christened "Erooms Law" [8]. A drop in productivity that few budgets in any industry can tolerate. Using 2016 prices: in 1950 a billion dollars would deliver over 30 drugs to market, today: less than one. Against this background the repeated organizational "efficiency drives" in drug discovery have clearly failed to deliver the promised radical improvements in productivity. Even rigorous application of good decision making practice and focusing research resources as recently discussed by Pangalos et al[9] although useful are nowhere near returning us to the heights of 1950's productivity

Econometric analysis of the areas of cost in drug discovery challenges the traditional view that Phase III trials represent the highest cost section in the process. Accounting for attrition, portfolio modeling and the cost of capital, the true area for maximum process improvement is in the lead optimization phase, as it occurs relatively early in the research and development cycle, is long, expensive and has a significant attrition rate[4]. A recent analysis[10] of the contrast between the technological and organizational advances made over the last 60 years makes the key point that accessing compounds active in a disease-relevant *in vivo* model remains the critical turning point in a drug discovery program from a scientific and investment perspective. The responsibility of the biologists in drug hunting teams is ensuring that the targets, assays and experimental models are aligned to the disease state. Here understanding the historic probability of success in different target classes may be of value[11], however the ingenuity of chemists has often overcome what were perceived as "undruggable" classes as demonstrated by recent progress in protein-protein interactions enabled by structural biology, fragment based lead generation and DNA encoded libraries[12]. There has been significant progress in this area using automated methods to identify better drug targets. The key roles of the chemists are to find and optimize compounds to the point where they can be dosed *in vivo* to the disease model and generate the critical data to select the right compound to take into the clinic.

Reducing the cost of lead optimization campaigns by accelerating them would allow more lead series to be studied per project and more biological approaches targeted per disease state. This has the potential to significantly increase the probability of developing new therapeutic approaches for an aging global population and address emerging threats from drug resistant pathogens.

## How can we accelerate Medicinal Chemistry?

Accelerating lead optimization entails addressing the central technical challenge of medicinal chemistry, which is the need to optimize potency at the biological target while simultaneously maintaining bioavailability through the appropriate therapeutic route of administration and avoiding toxicity; a highly challenging multifactorial design problem. Unsurprisingly, the design-make-test-analyze cycle most commonly used as a cyclic prototyping process usually requires a

large number of cycles for success, with the chemists using experience, chemical knowledge and simple general rules for guidance. The challenge implicit in these approaches is the build up of mental models that may become rigid and a practitioner or team may become "stuck" in a chemical series unable to deliver compounds with adequate properties to use *in vivo*.

Medicinal chemistry has gathered huge amounts of *in vitro* testing data in the last three decades particularly addressing ADME and toxicology issues. The vast majority of this is held in data vaults that are in danger of becoming data tombs if the knowledge contained there cannot be exploited[13]. The key *non-technical* goal in medicinal chemistry is the requirement to generate intellectual property in order to generate a return on the investment in research. This constraint means inactive or compromised structures are usually not publicly disclosed to avoid weakening IP positions. This generates a very significant publication bias[5] of chemical matter as spectacularly shown by Kramer et al [14] analyzing the distribution of potency data in ChEMBL where the modal pKi is 8.5. No practicing medicinal chemist would expect "average" compounds to have nanomolar potency. This lack of balanced publications generates a conflict between the societal need to lower the cost of drug discovery by improving medicinal chemistry practice and the commercial imperative to secure patents.

Across all fields of science, the cycling between systematizing knowledge leading to experiments, analysis of the causes of relationships and the rationalization of exceptions has driven progress. Though all codification and classifications are partial and flawed they create a framework for exploration and dialogue. Although a map is not a full description of a territory, it may be sufficient for navigation. The exceptions to apparent "rules" are fertile locations for exploring the underlying drivers of phenomena. Historically, the Chemical Abstract Service, Beilstein, Gmelin and more recently the ChEMBL, PDB and CCDC databases have all made a significant difference to the progression of chemistry, although it is hard to quantify their exact value.

Systematization of medicinal chemistry has therefore three main benefits: 1) the immediate understanding of what effects a particular chemical modification is most likely to have on an ADMET property to solve problems in a drug hunting program, 2) to create a corpus of knowledge that enables trends and meta-rules to be extracted and finally, 3) to generate hypotheses to test mechanistic understanding. All of these should improve the discipline of medicinal chemistry and therefore drug hunting. This is a particularly pressing need, as a decade of corporate reorganization and consolidation has moved much medicinal chemistry into contract research organizations and off-shored suppliers without access to the historic corporate knowledge of large Pharma. Contract research organizations are constrained in the use of their clients' data, so that knowledge mining across multiple projects is essentially impossible within these companies.

## Human aspects

Discussions of technology often avoid the critical discussion that all methods are mediated through people. As in most professions, the key value of the human element in medicinal chemistry is to make the critical assessment of what situation a discovery project is currently in, and then to choose an appropriate strategy to respond with. Line and project managers therefore have to be aware that chemists are often making these critical decisions in the context of poor data with immature theory and their role is to support and challenge them appropriately. All parties need to take responsibility to be on guard against the broad range of natural cognitive biases particularly the in-group behaviors that can be expressed when working under pressure and this is where a more data driven approach can assist. Across many fields, where an area is poorly quantified and tacit knowledge and experience are key, the Highest Paid Person's Opinion (HiPPO) may hold sway; this can obviously influence the uptake of new methods[15]. Some medicinal chemists may see a large knowledge base of potential solutions as a threat to their professional practice, and this may be the case if they only have a limited repertoire of tactics – or as a biologist colleague once described it "the chemists make the same compounds *whatever* the project". We must acknowledge that part of the development of professional medicinal chemistry is acceptance at a personal level that moving to more evidence led practice can be uncomfortable if it challenges our professional standing. However we can choose to treat evidence led practice as a framework to develop our expertise as a medicinal chemist[16]. Finally, the learning from other disciplines is that increased access to automated knowledge enables an "augmentation strategy" where the person + machine

combination is more effective than either on their own[17]. These are described in the artificial intelligence literature as "centaur teams", with the possibility being held out that partnership between human and machine should allow the speed, thoroughness and lack of bias of machines to be complimented by the creativity, intuition and situational awareness of the human.

In the absence of an instruction manual of medicinal chemistry, which would describe how to change the structure of a compound to transform it into a drug, the practitioner must adhere to broad "rules" and theoretical principles. These necessitate playing safe and although no "transgressions against medicinal chemistry rules" might be made such as "never make an aniline" [18], by avoid challenging assumptions, they may also result in avoiding innovative solutions[19]. In an alternative extreme strategy, the practitioner may choose to believe in the "exceptionalism of the drug", the "special" molecule that balances the conflicting properties; once this can be found, all will be well. The latter is a case of optimism bias, which can lead to an obsessive exploration of the end of a cul-de-sac when the overall picture may have been clear for some time. Under these circumstances, the argument may be made that "if only we find just the right compound it will be worth it when it gets to market". Although true, the continued expenditure at the most expensive part of the drug discovery process, may represent a severe opportunity cost in that there may well be better series or projects to work on.

Within different organizations the process of capturing medicinal chemistry knowledge varies depending on an organization's history, local culture and needs. Historically, the core process was essentially a scientific "apprenticeship" with experiential learning supported by older colleagues. The introduction of "wikis" – user maintained encyclopedias and storyboards of drug hunting projects has been discussed by others[20–22]. The offshoring and consolidation of medicinal chemistry has led to the move of significant numbers of experienced drug hunting medicinal chemists into universities. This clearly undermines the industrial "apprenticeship" approach. In both Europe and the United States the concern that medicinal chemistry education in industry is in trouble has been voiced with a view that industry-academia partnerships may present a solution[23–25]. This has enabled a more applied approach to the field being taught in universities but at the risk that those teaching are no longer practitioners and that the tools at their disposal are behind the current state of the art. Both of these approaches will still deliver a biased transfer of medicinal chemistry knowledge.

## Approaches to the systematization of medicinal chemistry

If we seek the best guidance possible, we must develop a means to systematize medicinal chemistry knowledge. As for many disciplines, there are three main approaches to this: 1) case series - gathering together stories to identify common themes in success and failure, 2) a reductionist approach - attempting to define the underlying factors for compound efficacy, metabolism and toxicity, and 3) data mining and supervised learning/AI - identifying tactics which have a statistically robust evidence base. The requirement to gain intellectual property makes the sharing of compound data and structures (required for all three approaches) difficult as weak compounds within its scope may undermine a patent, and compounds outside a patent scope may represent an opportunity for others to exploit. Although any of the approaches could be adopted either within private companies or on the censored public data, the broadest and most robust learning would be derived from access across both public and private data sets.

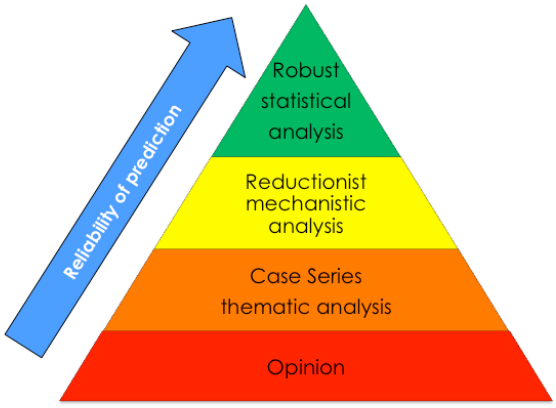


**Figure 1 Hierarchy of evidence for medicinal chemistry tactics**

In medicine, there is an acknowledged "hierarchy of evidence"[7] whereby case studies are seen as valuable in rapidly alerting practitioners to potential emergent issues such as safety issues and in generating hypotheses to test either mechanistically, via retrospective studies or randomized controlled trials. Retrospective studies in medicine carry more weight, with mechanistic studies and finally meta-analyses give most support for decision making. In the same way, it is possible to structure future attempts at better systematizing medicinal chemistry with a broad base of case studies feeding into both experimental and retrospective analyses to collate evidence to support a hypothesis and explore. A mechanistic chain can then be developed from evidence to hypothesis [26].

### Case Studies and Series in Medicinal Chemistry

The majority of medicinal chemistry teaching has been based around case studies of individual drug discovery stories[27–29]. Though educational, these run the risk common to anecdotes that the specific case is extrapolated and believed to be generally applicable. Time and effort can be wasted following an approach that only worked once - an example of the "base rate fallacy"[30]. Recently a number of publications have gathered together data in a more thematic approach, looking at particular functional groups and their effects[31–33]. These provide a more reliable view than overemphasis of particular solutions such as "fluorine blocks metabolism" [34] and that tetrazole is a general replacement for an acid [28]. However they still suffer the intrinsic limit of case study approaches in being vulnerable to the quality of the reviewers' literature searching and the limits on published data. The emerging challenge to the case study approach is that as the size of the literature for review grows exponentially larger, curating case studies in a comprehensive and unbiased way will inevitably become increasingly untenable.

## Can Medicinal Chemistry be codified without sharing IP-sensitive structures?

Part of the challenge in medicinal chemistry is in describing exactly what is done. There is a conflict here between the physics model of molecules as collections of sub-atomic particles and the structural representations used as shorthand by synthetic organic chemists. One challenge for practicing medicinal chemists is that although certain properties may appear continuous, it is hard to change them precisely when the tools to modulate properties are adding and removing atoms. More prosaically: finding an atom with half the volume of a chlorine but the same electronic properties may not be feasible. Relating a desired change in biological properties to a corresponding change in structure is therefore more useful than relating it to a change in an alternative property such a lipohilicity. This requires a language for the description of changes in chemical structure to be developed.

For instance, "add a fluorine para to the substituent on a monosubstituted phenyl ring" is a clear, actionable instruction to a medicinal chemist. It can also be transformed into a query to identify all

previous examples of such a change to be identified. This is a first step in moving to an evidence based approach in medicinal chemistry.

| Level | Description | Comments |
|---|---|---|
| 0 | C-H→C- electron withdrawing group | Carbon could be aromatic or aliphatic and a specific definition of "electron withdrawing group" is required. |
| 1 | Ar-H→Ar-electron withdrawing group | Ar = heteroaromatic rings of any size or composition, further substation allowed and not constrained |
| 2 | Ar-H→Ar-Halo | Halo defined as F,Cl,Br, I. |
| 3 | Ar-H→Ar-F | Only Ar is now variable |
| 4 | Ph-H→Ph-F | Specific to Ar = phenyl |
| 5 | R-p-phenyl-H → R-p-phenyl-F | Specific to substitution pattern on phenyl |
| 6 | Whole-molecule-p-phenyl- H → Whole-molecule-p-phenyl-F | *Single pair of compounds, an anecdote* |

Table 1 Describing structural changes with increasing levels of detail

Although the example described is a simple replacement of a substituent, the structural change could equally be a cyclisation, the shielding of a hydrogen bond acceptor by a vicinal methyl or the exchange of linking chains or core ring systems for isosteres. It is important to be sure that the structural change has been described and encoded in an appropriate fashion. The objective is to codify knowledge in a specific way so that the aggregated set of examples reveals a signal above the noise. For example, differing levels of specificity are shown in Table 1. Level 1 is the level at which a medicinal chemistry textbook might describe an approach, operating from the theoretical argument that because the predominant route of metabolism in aromatic rings is oxidative, destabilizing an incipient cation will reduce the rate of oxidation. As the transformation becomes more structurally specific, the number of examples will decrease and the variance of the effect will become smaller, indicating that the representation is capturing some correct feature of the chemical structure that influences the metabolic process. There is some evidence that aggregation at level 2 may offer some benefits as "fuzzy matched pairs"[35]. As soon as the transformations are aggregated, the details of the method of aggregation become an issue. For example, if we were to group "aryl-hydrogen bond acceptor" as a component, the definition of what constitutes a hydrogen bond acceptor becomes important; the case of methoxy being a good acceptor on an aliphatic ring but poor on an aromatic ring is a simple example. At the other extreme, if just a single pair of compounds is specified, it completely describes the transformation, but is just a scientific anecdote. In the development of matched molecular pair methodology, two critical papers suggest that the level 3, Ar-H→Ar-F can be too unspecific and give a "smear" of outcomes, whereas the inclusion of more chemical context (equivalent to level 5 R-p-phenyl-H → R-p-phenyl-F) can give a result with a lower variance so increasing confidence that this could be a useful change[36][37]. The possibility of using context-encoded matched molecular pair analysis to share medicinal chemistry knowledge was proposed by Dossetter et al[38] in 2013, with a confirmation of the data security of such an approach provided by Swamidass[39] in 2014.

## Mining Unbiased Medicinal Chemistry Knowledge from Data

The same root in physical chemistry that led to the study of model systems also drove the quantitative analysis of structure activity relationships (QSAR). The hope for broad scale models has been challenged by the shear size of accessible medicinal chemistry space[40]. Over the last two decades, three themes have been clear: the development of large scale QSAR models, the attempt to extract "simple rule models" for complex properties and the development of matched molecular pair analysis (MMPA) to analyze success frequencies for biological or physical properties[41].

Large-scale QSAR models have been developed mainly within Pharma as the only organizations with enough data to make adequately precise predictions. Their influence on broader practice has been limited by the constraint that it has been impossible to share the underlying structures between organizations or publicly. Even compounds anonymized by an identifier and the descriptors or fingerprints cannot be shared due to the information content in the descriptors presenting the opportunity to infer the probable structures[39]. The nature of the descriptors and the statistical approaches used also makes interpreting the models without access to the underlying structures difficult.

The "first wave" of simpler models were based on attempting to draw constraints around what may be "acceptable" chemical space and developed from the early and highly influential efforts of Lipinski [42]. Clearly the impact of poor solubility and permeability will have a negative effect throughout the drug discovery process from the validation of hits in cellular assays to in vivo studies. Further elaboration of the drug-likeness approach has been described by Congreve[43], Lovering[44] and Gleeson[45] and the PAINS style filters[46]. Recently, these have been critiqued for their methodology and more particularly for the underlying belief that "Given that drug discovery would appear to be anything but simple, the simplicity of a drug-likeness model could actually be taken as evidence for its irrelevance to drug discovery."[47,48]. It is not that focusing on simple molecular weight, logP hydrogen bond donor and acceptor counts is wrong, it is just usually not sufficient to solve medicinal chemistry problems.

In the early years of the 21st century, the formalization of matched molecular pair analysis (MMPA) was developed[1,49]. This is an approach that medicinal chemists had been using informally and with a rich statistical heritage in analogous medical matched cohort studies. The evolution of this approach was driven by its apparent simplicity and clarity of interpretation. A number of large scale MMPA have been carried out for a range of physical and biological endpoints[38]. A variety of methods have been developed to automate the MMPA process. One difference between MMPA and QSAR that is often overlooked is that QSAR generates declarative knowledge – a QSAR model is presented with a molecule and it makes an estimate for the modeled property. MMPA creates procedural knowledge – "if you change this substituent/linker/scaffold to the suggested group it will give a certain change in the property". MMPA presents new molecules to chemists as potential solutions.

In the last four years an approach has been tested to address the goal of sharing medicinal chemistry knowledge operating within the IP constraints around sharing primary structural data. It was recognized that matched pair relationships could be used as a "one way mapping" or "trapdoor" function. Once compounds have been assigned as members of a pair by a transformation, the original structures can no longer be inferred from the transformation. This has allowed aggregation of data across three large Pharma and consequently testing the question of whether particular medicinal chemistry methods are general or specific within a chemical class or project context. This addresses the critique that like QSAR models, the inferences depend on the data sets. Previously, either the results have been drawn from individual large Pharma datasets, where the risk is that particular individual projects or chemical series are generating an inference particular to that series, or from published data with the concomitant severe sample selection biases and small compound sets for some critically important biological endpoints.

The different approaches to automatic detection of matched molecular pairs have been reviewed[50]. Three critical features of a method for the capture of medicinal chemistry knowledge are: the ability of an algorithm to capture as many of the matched pairs a medicinal chemist would identify as possible, avoiding inclusion of "false pairs" and the transparent encoding of chemical environment as described above. A detailed analysis of the synergy between two complementary methods, "Fragment and Index" and "Multiple Common Sub Structure" (MCSS) has recently been published[51] showing on average a third of the pairs are found exclusively by one or other method, and a third are found by both.

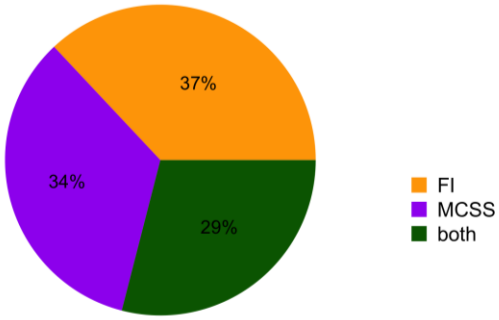


Figure 2 Mean percent of matched pairs found over 6 public data sets by method (VEGF, Dopamine transporter, GABA-A receptor, human D2 receptor, acetylcholine esterase, monoamine oxidase)

## Examples of Medicinal chemistry Knowledge gained by sharing MMPA analysis:

In the database created by merging medicinal chemistry knowledge from AstraZeneca, Genentech and Roche, a huge number of statistically significant medicinal chemistry rules were extracted as summarized in Table **2**.

| Datasets(s) | Number of increase / decrease/ neutral rules |
|---|---|
| logD7.4 | 153,449 |
| Solubility | 46,655 |
| *In vitro* microsomal clearance:<br>human, rat ,mouse, cyno, dog | 88,423 |
| *In vitro* hepatocyte clearance :<br>human, rat, mouse, cynomolgus monkey, dog | 26,627 |
| MDCK permeability A-B / B–A efflux | 1,852 |
| Cytochrome P450 inhibition:<br>2C9, 2D6 , 3A4 , 2C19 , 1A2 | 40,605 |
| Cardiac ion channels<br>NaV 1.5 , hERG ion channel inhibition | 15,636 |
| Glutathione Stability | 116 |
| Plasma protein binding<br>human, rat ,mouse, cynomolgus monkey, dog | 64,622 |

Table 2 Number of Statistically significant rules found from merging AstraZeneca, Genentech and Roche *in vitro* ADMET knowledge

### Definition of a rule:

There are several challenges in creating a definition for what constitutes a "rule". These are: managing out of range data, the contrast between the amount of evidence and the strength of the signal, avoiding the assumption that the data is normally distributed, and a metric simple enough to explain quickly to a non specialist. We therefore use a simple "coin flipping model" [51]. A given matched pair will either show an increase or a decrease in the measured property (where both members of a pair are out of range,

that pair is excluded). The number of increases and decreases are then treated as the equivalent of "heads" and "tails" in a binomial test. The results are tested to see if the distribution of increases and decreases would be outside what would be expected for a random distribution 95% of the time. Using this method, the more evidence (number of examples) that a rule has, the lower the frequency of a given direction is needed for it to pass statistical significance. Four worked examples show some effects of this:
32 matched pair examples of a rule are found: 23 (72%) lead to an increase this just passes the binomial test at 95% confidence with a p-value of 0.02;
16 matched pair examples: 13 lead to an increase just passing the binomial test with a p-value of 0.02, but now 81% of the examples need to increase to pass significance;
8 matched pair examples: now all 8/8 examples must increase reporting a p-value of 0.007, however if only 7/8 examples show an increase the p-value is 0.07, so there is a 7% chance that this distribution could be seen from a random distribution of examples;
5 matched pair examples: 5/5 show an increase in the property – this fails the binomial test with a p-value of 0.06, therefore even if 5 examples all show an increase from this rule, there is a 6% chance this could be due to a random distribution. Only where there are 6 or more examples of a matched pair can the binomial test be passed. This is an important piece of learning for "anecdotal" medicinal chemistry discussions, unless there are 6 or more examples, using a simple binomial test, it is not possible to state with >95% confidence that the medicinal chemistry "rule" proposed is anything other than a random distribution.

The large amount of unique knowledge found as shown in the filled donut diagram (Figure 2) mirrored the expected lack of overlap between different Pharmaceutical company collections[52].

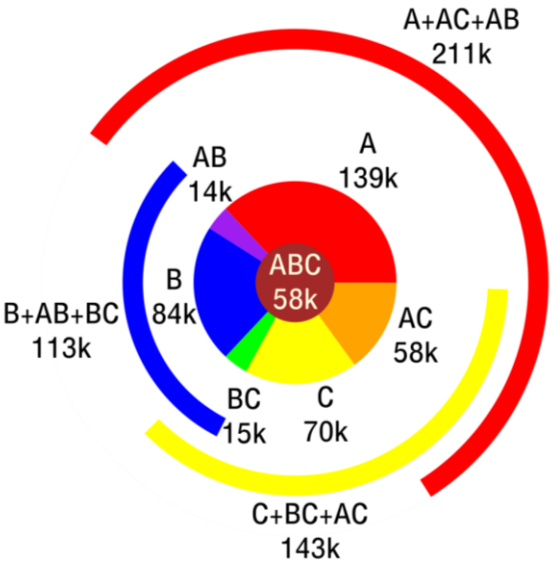


Figure 3 Origin of rules by company. The overlaps indicate the rules that contain examples from multiple companies, *i.e.* 58,000 rules had examples from all three companies, 139,000 rules were derived from company A data only. The total gain in rules by company was A:62%, B:156%, C:118%.

On inspecting this cross company database, one noticeable feature is the fine structural detail now available to direct medicinal chemistry. For example, previously the understanding of the probability of success of using a fluorine as a metabolic block could be summarized as "sometimes it works, sometimes it doesn't" with a number of case studies available[32]. Given the frustration and waste when a hard to make or expensive building block fails to deliver the reduction

in metabolism hoped for, knowing the circumstances where a "fluorine block" is likely to work is a significant benefit.

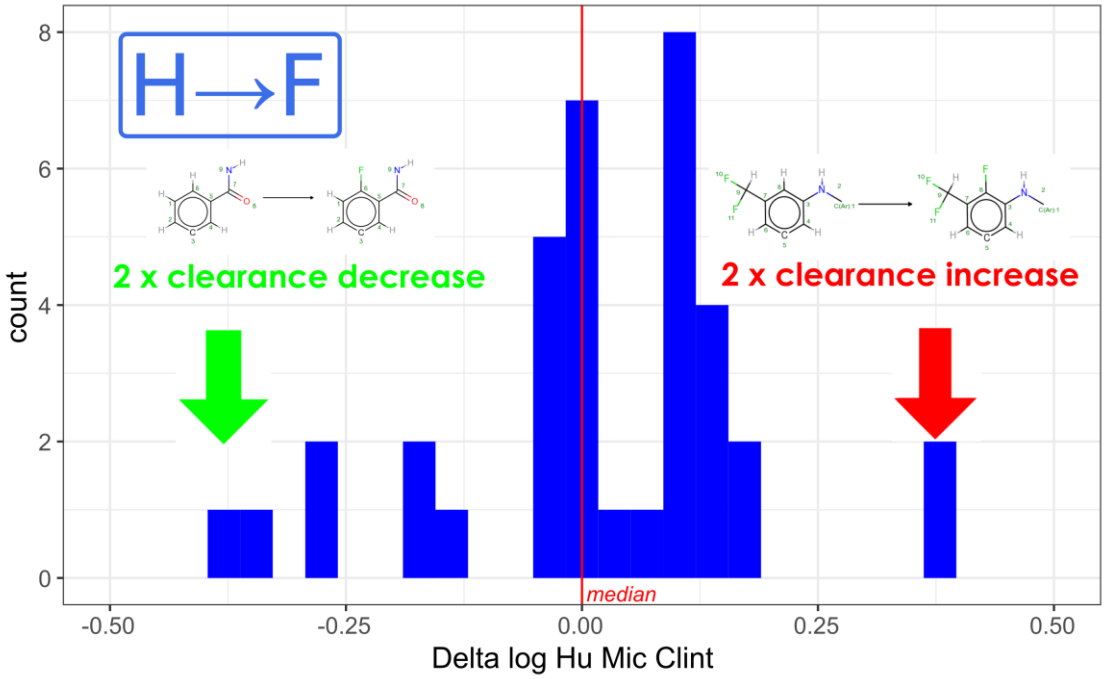


Figure 4 H→F substitution effects on human microsomal clearance in different environments and extreme examples

If the effect of H→F is split for different chemical environments around the point of change (Figure **4**), the historical distribution of successful transformations can be seen. The summary of this is that for the vast majority of chemical environments, the H→ F transformation has made less than 2 fold difference in either direction in human microsomal clearance (-0.3 < Δlog(Mic Clint) < 0.3). There are, however, precise situations where the H→ F transformation has been a good strategy, a 2 fold or more increase in metabolic stability would be expected (Δlog(Mic Clint) <= -0.3), and similarly in a few precise environments, a H→F transformation has significantly increased the rate of metabolism (Δlog(Mic Clint) >= +0.3). Two of these extreme examples are shown in Figure **4**.

As recognized by Hussein and Rea[53], the vast majority of transformations have very few example pairs supporting them and overall form a Zipfian distribution. This is an unsurprising consequence of the vast size of chemical space. The effect of this is that in merging data between companies, for the very few, very commonly observed transformations little new information is gained, however for the vast majority of transformations there is the opportunity to learn more by pooling examples. For the human liver microsome set as shown Figure **5**, 99% of the transformations in the data set had been observed less than 6 times (the minimum criteria for statistical testing). All of which represent an opportunity for increasing knowledge by pooling data.

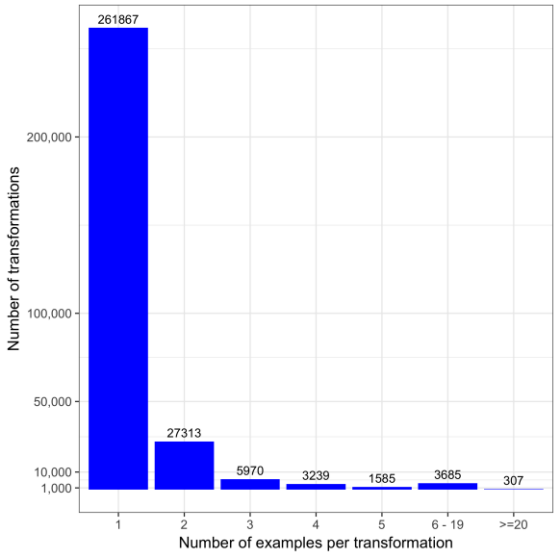


Figure 5 Distribution of number of pairs per transformation for a single company"s human microsomal clearance data

In looking at the knowledge in the database, further questions can immediately be addressed:

- Which medicinal chemistry changes are highly reproducible?
- What should I usually expect for a given structural change?
- What is the range I expect for that change (and therefore when to be suspicious?)
- How do they relate to "theory" and "experience"

One example of this is in looking at the relationships between properties for the same chemical transformation. For example (Figure **6**) when the large number of transformations for which there is both measured logD and measured solubility data is examined, a familiar broad trend is clear. As experience and theory suggest, overall solubility is negatively correlated with logD, as ΔlogD increases, Δsolubility decreases – however several more inferences are possible from this data. First: overall a drop of 1 unit logD gave, on average, a increase in solubility of approximately 0.6 log units (4 fold), second – the "lipophilic efficiency" of different transformations has a huge range. For example for isolipophilic changes (median ΔlogD = 0), the effect on median Δsolubility could range from -1.5 to +1.5, a 30 fold change in either direction. The colouring of regions of the Δsolubility/ΔlogD transformation plot shows that the majority of transformations that increase solubility are inefficient for the amount of solubility gained with respect to change in logD – ie Δsolubility < -Δ□□□□ where Δsolubility >0. There are a small number of transformations that are unexpectedly good as outliers where logD can be increased and solubility increases as well, these represent very high value transformations to medicinal chemists.

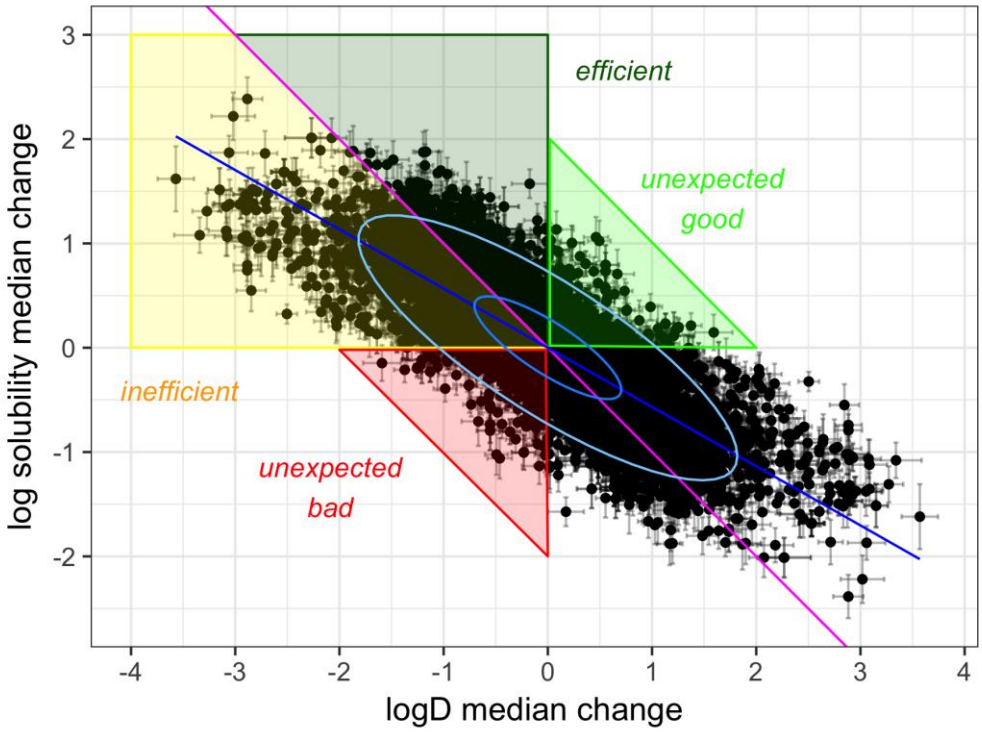


Figure 6 Solubility vs logD effects, >=20 examples per rule, n=13453. $R^2$ = 0.66, slope = -0.57, intercept = 0. Magenta line: line of slope -1, intercept 0, dark blue line linear best fit, pale blue density ellipse contains 99% and the mid blue ellipse contains 50% of the transformations.

A very similar picture is seen when we look at the relationship between metabolic stability and logD. The same form of analysis holds true for the other *in vitro* endpoints studied and interesting outliers have been identified [54].

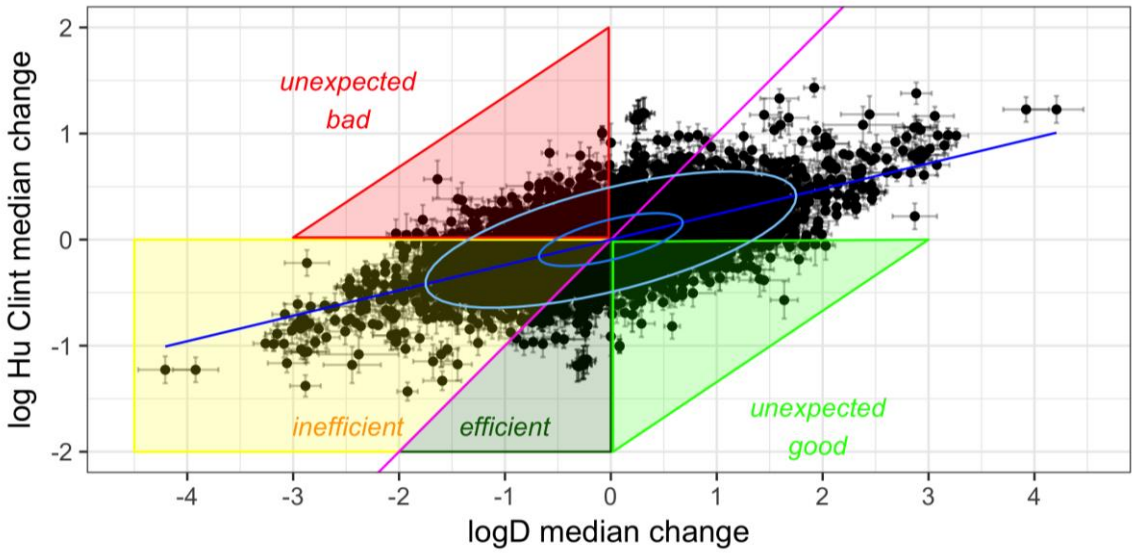


Figure 7 Human microsomal clearance vs logD, >=20 examples per rule, n=11,572. $R^2$ = 0.40, slope 0.23, intercept = 0. Magenta line: line of slope 1, intercept 0, dark blue line linear best fit, pale blue density ellipse contains 99% and the mid blue ellipse contains 50% of the transformations.

Beyond analysis of the efficiency of medicinal chemistry transformations with respect to lipophilicity, other SAR insights can be gained. Comparing the effects of sets of transformations on different ion channels can be important; a transformation that reduces hERG binding but increases NaV1.5 ion channel binding would be a pyrrhic victory. Similarly, how the same transformations

have different effects on hepatic metabolism according to species can help to avoid moving compounds forward that will have a beneficial effect on rodent metabolism but a poorer effect on human or dog. These insights have the potential to make very considerable savings in the progression of compounds in late lead optimization.

### Potential Risks of collating Medicinal Chemistry Knowledge

One risk that has been voiced in generating a large knowledge base is that "everyone will just do the same thing" in particular that it could lead to limiting medicinal chemistry to what has been well explored in the past. There are several responses to this. The first is to acknowledge that actually given the evidence of multiple published analyses[55,56] and above (Figure **5** Distribution of number of pairs per transformation for a single company"s human microsomal clearance data) this is where medicinal chemistry is currently. The evidence is that medicinal chemists apply a limited set of tactics very frequently. Understanding the historic success rates for these at least allows the best of the "common tactics" to be tried first. With a broader collection of medicinal chemistry knowledge the chemist may be prompted to explore outside their preferred set of tactics and exploit the learning of others more effectively. Chemists may also then take on the higher level analysis grouping successful solutions into classes and generating more strategic hypotheses which then can be used to create new solutions to test.

## Medicinal Chemistry Knowledge Exploitation

A database of validated medicinal chemistry transformations can be exploited in a number of ways. The most obvious is as a searchable resource to suggest alternatives and to benchmark proposed ideas. There are additional benefits in encoding the transformations as reactions in that enumeration software can be used to convert "problem" molecules into potential solutions. This was originally recognized by Fujita with the development of EMIL[57] and then again in Abbott's Drug Guru software[58] where manually encoded reactions were used to capture medicinal chemistry experience. Extension of this type of approach via cyclic enumeration and scoring can create "evolutionary optimization"[59],[60]. Integration of a knowledge base of transformations with an enumerator creates an expert system that can propose solutions to medicinal chemistry problems – one class of artificial intelligence.

Automated encoding of transformations between molecules also enables further possibilities such as using a set of known measured molecules to predict the properties of a proposed new molecule by matching the transformations from the known to the proposed generating a *de facto* "ratings" service. Further possibilities include a broader scale exploration of the matched series concept[61–63]. Here, a set of compounds with a common core but a range of different substituents generates an SAR "fingerprint". Biological targets demonstrating the same SAR can be aligned and useful questions like "given that the SARs of these two series align; what is the potency of this new substituent likely to be?", ""what is the most potent substituent that can be transferred from one aligned series to another?" and "what biological targets have the same SAR as my target?" Clearly this can address one of the other key weak points in the chemical arm of drug hunting: generating potent chemical leads. Even more generally, the network of transformations can be explored to see if knowledge can be inferred: if A→B and B→ C can A→ C be estimated or will errors propagate too strongly? Alternatively if A→B, A→C, A→D all lead to a decrease in toxicity, can A be designated a toxophore?

## What role for artificial intelligence to improve Medicinal Chemistry

The last decade has seen a resurgence in artificial intelligence (AI). In its current form, the predominant AI paradigm is the analysis of very large data sets with statistical machine learning methods. The most ubiquitous uses have been in image recognition and natural language processing of textual data. These have developed from experimental identification of cat pictures and film review classification to ubiquitous face recognition and sentiment analysis in marketing. [64,65] This has been enabled by three factors: massive free training data sets harvested from the

internet, hardware acceleration driven by the gaming industry and these two factors in turn enabling the application of complex neural net architectures such as deep, convoluted and adversarial networks. The application of these methods in drug hunting has been demonstrated in high profile cases such as the Merck Kaggle challenge[66], the NIH Tox21 challenge[67] and the use of generative neural networks to identify novel RXR agonists[68]. More recent successes in the test arena of game playing, Alpha Go[69] and Alpha Go Zero[70] have demonstrated that extending neural net methodologies continues to create new opportunities. The potential to integrate AI systems with automated synthesis and testing synthesis "closing the loop" of cyclic prototyping has recently been reviewed[71] . This would have the potential of replacing the medicinal chemist, however as we discuss below, the machine learning methods and critically, data sets they are built on will have to significantly improve before "self-driving drug hunting" becomes a reality. Currently, AI appears to be at the peak of its latest hype cycle with the "Trough of Disillusionment" beckoning, hopefully this will followed by more realistic integration into drug hunting and the "Plateau of Productivity".[72]

One simple question however is, will the current class of AI approaches be able to address the complex multi parameter optimization problem of drug discovery? A trite summary of the neural net based systems appears to be that "neural nets are good at tasks humans are good at". Asking the question: "what tasks in drug design are humans good at?" does not bode well for AI's based on this technology. Indeed, multi-objective design remains an intrinsically hard problem because as the number of objectives rises the training data becomes increasingly sparse; this is the "curse of dimensionality". A further critical issue not often explicitly addressed is that the non-chemical successes in AI have been built on vastly larger training data sets than currently available within any given Pharma or publicly for drug hunting. Machine learning approaches built on data sets that are too small tend to be "brittle": they appear reasonable until challenged with situations where they are undertrained at which point the predictions become poor. The "human backstop" in a human-AI mixed team should mitigate this risk. Lack of interpretability is a further challenge in the application of neural net based and other "black box" machine learning algorithms. Unlike regression-based machine learning it's very hard to expose which factors (in chemical terms, which substructures) are driving the estimate the algorithm is generating. At a human level this disempowers the user as they are faced with "doing what the machine says" or not, but without a method to assess the validity of the prediction. This may decelerate adoption of AI methods in the scientific arena. Auditable AI is of such significant interest as to have been recognized by the US National Science and Technology Council as a key component to building trust in AI systems in their 2016 R&D Strategic Plan.[73] At a technical level, the algorithm may be making a prediction on a very small subset of the data, or may have effectively encoded biases in the data. This algorithmic bias is an area of significant current concern and research [74]. Without being able to understand the drivers of predictions, the medicinal chemist may be just perpetuating existing organizational preferences but with the "fig leaf" of computational support. Interpretability is a harder area to research than predictive accuracy since it requires a subjective assessment rather than a numerical score. One can imagine uncritical overreliance on AI methods exposing drug hunting teams to significant risks. Without experienced medicinal chemists to audit the rationale behind suggestions, black box models operating outside their domain of applicability could be making wild suggestions and a project could waste significant resource in making unlikely compounds. This is particularly relevant as novel biological target classes are explored, where although chemists may not have specific knowledge of an area, they have the skills to construct sets of experiments to explore the parameters of the medicinal chemistry search space.

The success of AI and machine learning approaches in medicinal chemistry is critically dependent on access to large enough data sets to train on and methods that enable interrogation and interpretation. It seems most likely though, that to misquote McAfee and Brynjolfsson: "AI won't replace medicinal chemists, but medicinal chemists who use AI will replace those who don't".[75]

**Artificial Intelligence and Chemistry**

Artificial intelligence has addressed chemical problems throughout its history. .
The "expert system" era which used rules encoded by specialists produced DENDRAL [76] which could identify compounds from their mass spectra, CASE [77] and DEREK [78] for

identifying potential toxicities, EROS [79] and LHASA [80] the systems for proposing synthetic routes. All these were based on human experts encoding a set of "rules" in a format that could be used computationally to rank potential solutions. As discussed above, the more modern "Big Data" approach is to use large data sets and statistical methods to infer either rules, or statistical algorithms (such as the variety of neural net architectures) that when presented with a problem will provide the most likely solution(s). The question "is this intelligence in chemistry?" can be addressed by a Feigenbaum test [81]: if you present a system with a chemistry based challenge and the response is indistinguishable from that which a trained chemist or group of chemists would provide, then it's indistinguishable from intelligence.

## Medicinal Chemistry Knowledge in the Drug Discovery Ecosystem

As discussed at the start of this article, medicinal chemistry is an applied science. The majority of therapeutic agents have been discovered in the private sector and in the vast majority of cases developed there. Drug discovery and development is highly regulated and expensive so organizations that carry this out need large capital reserves to develop and market new agents. Therefore the sharing of knowledge of how to discover and develop drugs better has an implicit tension. For a large Pharma, which has invested significantly in large scale compound synthesis and testing to generate data, the value of sharing knowledge with an equivalent sized organization is relatively straightforward to assess. Both parties are expecting an approximately equal benefit. This can be described as a transactional relationship, although each party may not be able to estimate the exact return on investment, there is the assumption that for each party the knowledge will be equally useful. The value that can be extracted from the new knowledge gained depends on the efficiency with which research is undertaken in each organization.

Taking a more strategic view: Pharma could improve the quality of in-licensing candidates by sharing knowledge more widely with the academic, not-for-profit and biotech sectors. This would result in faster drug discovery in the non-large Pharma sector and hence cheaper and better in-licensing candidates which in turn gives a better long term return on investment. However, this "rising tide lifts all boats" argument is harder to quantify as the benefits are more distant, and therefore it is more difficult for Pharma managers to make the case that it is valid. This is an example of the cognitive bias of hyperbolic discounting – longer term, larger rewards being underestimated. A short term counter argument for internal Pharma research teams is that generating more external competition to their endeavors is a counterproductive to their own survival.

Within Pharma discovery teams, the argument is put forward that SAR knowledge represents key intellectual capital for their company. This appears to be rarely supported in practice as large Pharma frequently relocate research centres or outsource programmes which inevitably leads to loss of tacit knowledge. Within the not-for-profit sector it is proposed that "open source drug discovery" is analogous to "open source software". Though there are parallels in that both fields are involved in generation of intellectual property, there are very significant differences in the regulatory regimes for software and pharmaceuticals, consequences of errors and the product life cycles. Counterbalanced against this argument is the widely held visceral public view that drug discovery is a "public good", which leads to the Pharma sector demonstrating "corporate social responsibility" (CSR) in funding neglected disease research and providing low cost critical medicines to poorer parts of the world in for example in the treatment of HIV and parasitic worm infections. It is reasonable to expect that, as in the case of CSR, different companies will take different views on the strategic benefit of making their knowledge more widely available and then some may see this as another arena in which to compete for reputational gain.

## Conclusion

From a technical perspective, a sufficiently large medicinal chemistry database of transformations may provide novel approaches to improving drug discovery. A record of historical successes can spur the development of novel solutions by combining old approaches or seeing a conceptual link between multiple previous successes. The question is asked perennially "what if in the end we all make the same compounds" which appears to ignore the evidence of the vast size of chemical

space, it's huge diversity and its under exploration by chemists to date. Huge numbers of ring systems are unsynthesised and the current exploration of macrocyclic molecules [82] towards creating synthetic natural-product-like structures extends the reach of medicinal chemistry space still further. Yet more knowledge is undoubtedly there for the elucidating, but it will only become clear when we have sufficient data.

Conceptually, the use of a sufficiently large corpus of knowledge may be considered analogous to the effect that massive datasets have in automated language translation, where above a certain threshold, prediction becomes highly effective[83]. Treatment of emerging pathogens and the diseases of an aging population may require new chemistries and the exploring of multiple biological targets. To bring these "within range" of the investment available needs the application of all the knowledge we have. Enhancing our medicinal chemistry knowledge seems a central component in this task. The value of a very large scale systematized medicinal chemistry knowledge base appears to be hard to dispute. The technical and legal challenges have been overcome. However, the strategies for sharing such knowledge are corporate issues to be addressed by the leaders in our industry.

[1] Kenny PW, Sadowski J. Structure Modification in Chemical Databases. In: Oprea TI, editor. Methods Princ. Med. Chem., Weinheim, FRG: Wiley-VCH Verlag GmbH & Co. KGaA; 2005, p. 271–85.
[2] Hansch C, Leo A, Hoekman DH, editors. Exploring QSAR. Washington, DC: American Chemical Society; 1995.
[3] Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, et al. QSAR Modeling: Where Have You Been? Where Are You Going To? J Med Chem 2014;57:4977–5010. doi:10.1021/jm4004285.
[4] Paul SM. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010;9:203–14. doi:10.1038/nrd3078.
[5] Rothstein HR, Sutton AJ, Borenstein M. Publication Bias in Meta-Analysis. In: Rothstein HR, Sutton AJ, Borenstein M, editors. Publ. Bias Meta-Anal., Chichester, UK: John Wiley & Sons, Ltd; 2006, p. 1–7. doi:10.1002/0470870168.ch1.
[6] Kramer C, Kalliokoski T, Gedeck P, Vulpetti A. The Experimental Uncertainty of Heterogeneous Public $K_i$ Data. J Med Chem 2012;55:5165–73. doi:10.1021/jm300131x.
[7] Greenhalgh T. How to read a paper : getting your bearings (deciding what the paper is about). BMJ 1997;315:243–6. doi:10.1136/bmj.315.7102.243.
[8] Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov 2012;11:191–200. doi:10.1038/nrd3681.
[9] Morgan P, Brown DG, Lennard S, Anderton MJ, Barrett JC, Eriksson U, et al. Impact of a five-dimensional framework on R&D productivity at AstraZeneca. Nat Rev Drug Discov 2018. doi:10.1038/nrd.2017.244.
[10] Scannell JW, Bosley J. When Quality Beats Quantity: Decision Theory, Drug Discovery, and the Reproducibility Crisis. PLOS ONE 2016;11:e0147215. doi:10.1371/journal.pone.0147215.
[11] Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19–34. doi:10.1038/nrd.2016.230.
[12] Arkin MR, Tang Y, Wells JA. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. Chem Biol 2014;21:1102–14. doi:10.1016/j.chembiol.2014.09.001.
[13] Fayyad U, Uthurusamy R. Evolving data into mining solutions for insights. Commun ACM 2002;45. doi:10.1145/545151.545174.
[14] Kramer C, Fuchs JE, Whitebread S, Gedeck P, Liedl KR. Matched Molecular Pair Analysis: Significance and the Impact of Experimental Uncertainty. J Med Chem 2014;57:3786–802. doi:10.1021/jm500317a.
[15] McAfee A, Brynjolfsson E. Big Data: The Management Revolution. Harv Bus Rev 2012.
[16] Ericsson KA, editor. The Cambridge handbook of expertise and expert performance. Cambridge ; New York: Cambridge University Press; 2006.
[17] Davenport T, Kirby J. Beyond Automation. Harv Bus Rev 2015.
[18] Kalgutkar AS, Gardner I, Obach RS, Shaffer CL, Callegari E, Henne KR, et al. A comprehensive listing of bioactivation pathways of organic functional groups. Curr Drug Metab 2005;6:161–225.
[19] Birch AM, Groombridge S, Law R, Leach AG, Mee CD, Schramm C. Rationally Designing Safer Anilines: The Challenging Case of 4-Aminobiphenyls. J Med Chem 2012;55:3923–33. doi:10.1021/jm3001295.
[20] Robb GR, McKerrecher D, Newcombe NJ, Waring MJ. A chemistry wiki to facilitate and enhance compound design in drug discovery. Drug Discov Today 2013;18:141–7. doi:10.1016/j.drudis.2012.09.002.
[21] Mayweg A, Hofer U, Schnider P, Agnetti F, Galley G, Mattei P, et al. ROCK: the Roche medicinal chemistry knowledge application ? design, use and impact. Drug Discov Today 2011;16:691–6. doi:10.1016/j.drudis.2011.03.005.

[22] Stahl M, Baier S. How Many Molecules Does It Take to Tell a Story? Case Studies, Language, and an Epistemic View of Medicinal Chemistry. ChemMedChem 2015;10:949–56. doi:10.1002/cmdc.201500091.
[23] Rafferty MF. No Denying It: Medicinal Chemistry Training Is in Big Trouble: Miniperspective. J Med Chem 2016;59:10859–64. doi:10.1021/acs.jmedchem.6b00741.
[24] Macdonald SJF, Fray MJ, McInally T. Passing on the medicinal chemistry baton: training undergraduates to be industry-ready through research projects between the University of Nottingham and GlaxoSmithKline. Drug Discov Today 2016;21:880–7. doi:10.1016/j.drudis.2016.01.015.
[25] Allen D. Where will we get the next generation of medicinal chemists? Drug Discov Today 2016;21:704–6. doi:10.1016/j.drudis.2016.04.012.
[26] Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). SAR QSAR Environ Res 2009;20:241–66. doi:10.1080/10629360902949567.
[27] Brimblecombe RW, Duncan WA, Durant GJ, Emmett JC, Ganellin CR, Leslie GB, et al. Characterization and development of cimetidine as a histamine H2-receptor antagonist. Gastroenterology 1978;74:339–47.
[28] Wexler RR, Carini DJ, Duncia JV, Johnson AL, Wells GJ, Chiu AT, et al. Rationale for the chemical development of angiotensin II receptor antagonists. Am J Hypertens 1992;5:209S–220S.
[29] Dorsey BD, Levin RB, McDaniel SL, Vacca JP, Guare JP, Darke PL, et al. L-735,524: the design of a potent and orally bioavailable HIV protease inhibitor. J Med Chem 1994;37:3443–51.
[30] Kahneman D, Tversky A, editors. Choices, values, and frames. New York : Cambridge, UK: Russell sage Foundation ; Cambridge University Press; 2000.
[31] Meanwell NA. Synopsis of Some Recent Tactical Application of Bioisosteres in Drug Design. J Med Chem 2011;54:2529–91. doi:10.1021/jm1013693.
[32] Gillis EP, Eastman KJ, Hill MD, Donnelly DJ, Meanwell NA. Applications of Fluorine in Medicinal Chemistry. J Med Chem 2015;58:8315–59. doi:10.1021/acs.jmedchem.5b00258.
[33] Beno BR, Yeung K-S, Bartberger MD, Pennington LD, Meanwell NA. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. J Med Chem 2015;58:4383–438. doi:10.1021/jm501853m.
[34] Hagmann WK. The Many Roles for Fluorine in Medicinal Chemistry. J Med Chem 2008;51:4359–69. doi:10.1021/jm800219f.
[35] Geppert T, Beck B. Fuzzy Matched Pairs: A Means To Determine the Pharmacophore Impact on Molecular Interaction. J Chem Inf Model 2014;54:1093–102. doi:10.1021/ci400694q.
[36] Papadatos G, Alkarouri M, Gillet VJ, Willett P, Kadirkamanathan V, Luscombe CN, et al. Lead Optimization Using Matched Molecular Pairs: Inclusion of Contextual Information for Enhanced Prediction of hERG Inhibition, Solubility, and Lipophilicity. J Chem Inf Model 2010;50:1872–86. doi:10.1021/ci100258p.
[37] Hajduk PJ, Sauer DR. Statistical analysis of the effects of common chemical substituents on ligand potency. J Med Chem 2008;51:553–64.
[38] Dossetter AG, Griffen EJ, Leach AG. Matched Molecular Pair Analysis in drug discovery. Drug Discov Today 2013;18:724–31. doi:10.1016/j.drudis.2013.03.003.
[39] Matlock M, Swamidass SJ. Sharing Chemical Relationships Does Not Reveal Structures. J Chem Inf Model 2014;54:37–48. doi:10.1021/ci400399a.
[40] Polishchuk PG, Madzhidov TI, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des 2013;27:675–9. doi:10.1007/s10822-013-9672-4.
[41] Lombardo F, Desai PV, Arimoto R, Desino KE, Fischer H, Keefer CE, et al. *In Silico* Absorption, Distribution, Metabolism, Excretion, and Pharmacokinetics (ADME-PK): Utility and Best Practices. An Industry Perspective from the International Consortium for Innovation through Quality in Pharmaceutical Development: Miniperspective. J Med Chem 2017;60:9097–113. doi:10.1021/acs.jmedchem.7b00487.

[42] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 1997;23:3–25. doi:10.1016/S0169-409X(96)00423-1.
[43] Congreve M, Carr R, Murray C, Jhoti H. A "rule of three" for fragment-based lead discovery? Drug Discov Today 2003;8:876–7. doi:10.1016/S1359-6446(03)02831-9.
[44] Gleeson MP. Generation of a set of simple, interpretable ADMET rules of thumb. J Med Chem 2008;51:817–34. doi:10.1021/jm701122q.
[45] Lovering F, Bikker J, Humblet C. Escape from flatland: increasing saturation as an approach to improving clinical success. J Med Chem 2009;52:6752–6. doi:10.1021/jm901241e.
[46] Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. J Med Chem 2010;53:2719–40. doi:10.1021/jm901137j.
[47] Kenny PW, Montanari CA. Inflation of correlation in the pursuit of drug-likeness. J Comput Aided Mol Des 2013;27:1–13.
[48] Muthas D, Boyer S, Hasselgren C. A critical assessment of modeling safety-related drug attrition. Med Chem Commun 2013;4:1058–65.
[49] Leach AG. 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 2006;49:6672–82.
[50] Tyrchan C, Evertsson E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. Comput Struct Biotechnol J 2017;15:86–90. doi:10.1016/j.csbj.2016.12.003.
[51] Lukac I, Zarnecka J, Griffen EJ, Dossetter AG, St-Gallay SA, Enoch SJ, et al. Turbocharging Matched Molecular Pair Analysis: Optimizing the Identification and Analysis of Pairs. J Chem Inf Model 2017;57:2424–36. doi:10.1021/acs.jcim.7b00335.
[52] Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, et al. Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. Drug Discov Today 2013;18:1014–24. doi:10.1016/j.drudis.2012.10.011.
[53] Hussain J, Rea C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. J Chem Inf Model 2010;50:339–48. doi:10.1021/ci900450m.
[54] Kramer C, Ting A, Zheng H, Hert J, Schindler T, Stahl M, et al. Learning Medicinal Chemistry Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) Rules from Cross-Company Matched Molecular Pairs Analysis (MMPA): Miniperspective. J Med Chem 2017. doi:10.1021/acs.jmedchem.7b00935.
[55] Walters WP, Green J, Weiss JR, Murcko MA. What Do Medicinal Chemists Actually Make? A 50-Year Retrospective. J Med Chem 2011;54:6405–16. doi:10.1021/jm200504p.
[56] Brown DG, Boström J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone?: Miniperspective. J Med Chem 2015. doi:10.1021/acs.jmedchem.5b01409.
[57] Fujita T, Adachi M, Akamatsu M, Asao M, Fukami H, Inoue Y, et al. Background and features of emil, a system for database-aided bioanalogous structural transformation of bioactive compounds. Pharmacochem. Libr., vol. 23, Elsevier; 1995, p. 235–73.
[58] Stewart KD, Shiroda M, James CA. Drug Guru: A computer software program for drug design using medicinal chemistry rules. Bioorg Med Chem 2006;14:7011–22. doi:10.1016/j.bmc.2006.06.024.
[59] Firth NC, Atrash B, Brown N, Blagg J. MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. J Chem Inf Model 2015;55:1169–80. doi:10.1021/acs.jcim.5b00073.
[60] Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P, et al. Automated design of ligands to polypharmacological profiles. Nature 2012;492:215–20. doi:10.1038/nature11691.
[61] Wawer M, Bajorath J. Local Structural Changes, Global Data Views: Graphical Substructure−Activity Relationship Trailing. J Med Chem 2011;54:2944–51. doi:10.1021/jm200026b.

[62] O'Boyle NM, Boström J, Sayle RA, Gill A. Using Matched Molecular Series as a Predictive Tool To Optimize Biological Activity. J Med Chem 2014;57:2704–13. doi:10.1021/jm500022q.
[63] Keefer CE, Chang G. The use of matched molecular series networks for cross target structure activity relationship translation and potency prediction. MedChemComm 2017;8:2067–78. doi:10.1039/C7MD00465F.
[64] Le QV. Building high-level features using large scale unsupervised learning, IEEE; 2013, p. 8595–8. doi:10.1109/ICASSP.2013.6639343.
[65] Pang B, Lee L, Vaithyanathan S. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. Proc. EMNLP, 2002, p. 79–86.
[66] Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. J Chem Inf Model 2015;55:263–74. doi:10.1021/ci500747n.
[67] Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Environ Sci 2016;3. doi:10.3389/fenvs.2015.00080.
[68] Merk D, Friedrich L, Grisoni F, Schneider G. *De Novo* Design of Bioactive Small Molecules by Artificial Intelligence. Mol Inform 2018;37:1700153. doi:10.1002/minf.201700153.
[69] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9. doi:10.1038/nature16961.
[70] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017;550:354–9. doi:10.1038/nature24270.
[71] Schneider G. Automating drug discovery. Nat Rev Drug Discov 2017;17:97–113. doi:10.1038/nrd.2017.232.
[72] Panetta K. Top Trends in the Gartner Hype Cycle for Emerging Technologies, 2017. Top Trends Gart Hype Cycle Emerg Technol 2017 n.d. https://www.gartner.com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/.
[73] National Science and Technology Council. THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH and DEVELOPMENT STRATEGIC PLAN. CreateSpace Independent Publishing Platform; 2016.
[74] Castelvecchi D. Can we open the black box of AI? Nature 2016;538:20–3. doi:10.1038/538020a.
[75] McAfee A, Brynjolfsson E. The Business of Artificial Intelligence. Harv Bus Rev 2017.
[76] Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J. DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artif Intell 1993;61:209–61. doi:10.1016/0004-3702(93)90068-M.
[77] Klopman G. Artificial intelligence approach to structure-activity studies. Computer automated structure evaluation of biological activity of organic molecules. J Am Chem Soc 1984;106:7315–21. doi:10.1021/ja00336a004.
[78] Marchant CA, Briggs KA, Long A. In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. Toxicol Mech Methods 2008;18:177–87. doi:10.1080/15376510701857320.
[79] Gasteiger J, Hutchings MG, Christoph B, Gann L, Hiller C, Löw P, et al. A new treatment of chemical reactivity: Development of EROS, an expert system for reaction prediction and synthesis design. Org. Synth. React. Mech., vol. 137, Berlin, Heidelberg: Springer Berlin Heidelberg; 1987, p. 19–73. doi:10.1007/3-540-16904-0_14.
[80] Corey EJ, Wipke WT. Computer-Assisted Design of Complex Organic Syntheses. Science 1969;166:178–92. doi:10.1126/science.166.3902.178.
[81] Feigenbaum EA. Some challenges and grand challenges for computational intelligence. J ACM 2003;50:32–40. doi:10.1145/602382.602400.
[82] Whitty A, Viarengo LA, Zhong M. Progress towards the broad use of non-peptide synthetic macrocycles in drug discovery. Org Biomol Chem 2017;15:7729–35. doi:10.1039/C7OB00056A.

[83] Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. IEEE Intell Syst 2009;24:8–12. doi:10.1109/MIS.2009.36.