Box S2 | Searching chemical space

In the 1970s and 1980s, an iterative search for a new drug might involve, at most, the synthesis and testing of 1,000 molecules. Today, a typical corporate collection of compounds for high-throughput screening (HTS) contains in the order of 1,000,000 molecules. In this text box, we aim to explain, to non-chemists, some of the reasons why an old and labour-intensive search process that involved 1,000 molecules might be more efficient than a highly automated search process that involves 1,000,000 molecules. We use two examples to illustrate the arguments. The first is the 20 questions game. The second uses geometric logic to show how search efficiency changes with the size and dimensionality of a search space. At the end of this supplementary text, we list some references that give an introduction to some of the concepts of "chemical space" that are important for drug discovery[1-4].

20 questions

Imagine that you know all of the 600,000 or so words in the English language and that you are asked to guess an English word written in a sealed envelope. You are offered two search strategies. The first is the familiar '20 questions' game. You can ask a series of questions. You are provided with a "yes" or "no" answer to each, and you win if you guess the word in the envelope having asked 20 questions or fewer. The second strategy is a brute force method. You get 20,000 guesses, but you only get a "yes" or "no" once you have made all 20,000 guesses. So which is more likely to succeed, 20 questions or 20,000 guesses?

A skilled player should usually succeed with 20 questions (since 600,000 is less than than 2^{20}) but would fail nearly 97% of the time with "only" 20,000 guesses.

Our view is that the old iterative method of drug discovery was more like 20 questions, while HTS of a static compound library is more like 20,000 guesses. With the iterative approach, the characteristics of each molecule could be measured on several dimensions (for example, potency, toxicity, ADME). This led to multi-dimensional structure–activity relationships, which in turn meant that each new generation of candidates tended to be better than the previous generation. In conventional HTS, on the other hand, search is focused on a small and pre-defined part of chemical space, with potency alone as the dominant factor for molecular selection.

The "Blackian demon", super-HTS and high dimensional spaces

Imagine we have a large number of "nodes" that occupy a space (for example, 10^{60} drug-like molecules). The space might be low dimensional. This would be true if, for example, there was one class of molecule; simple chain hydrocarbons for example, which went from CH_4, C_2H_6, C_3H_8, and so on up to a chain that was 10^{60} carbon atoms long. Alternatively, the space might be high dimensional, as chemical space really is. In this context, think of the number of dimensions of the space as the number of coordinates that would be required to uniquely specify any point.

Imagine we have two search strategies to find the single best molecule in this space. One is a brute force search, which assays a molecule and then simply steps to the next molecule, and so exhaustively searches the entire space. We call this "super-HTS". The other, which we call the "Blackian demon" (in reference to the "Darwinian demon", which is used sometimes to reflect ideal performance in evolutionary thought experiments, and in tribute to James Black, often acknowledged as one of the most successful drug discoverers), is equivalent to an omniscient drug designer who can assay a molecule, and then make a single chemical modification to step it one position through chemical space, and who can then assay the new molecule, modify it again, and so on. The Blackian demon can make only one step at a time, to a nearest-neighbour molecule, but it always steps in the right direction; towards the best molecule in the space. We can compare the number of steps that super-HTS and the Blackian demon have to make to exhaustively search the space for the best molecule. The number of steps for the Blackian demon follows from simple geometry. If you have a d dimensional space with n nodes in the space, and – for simplicity – these are arranged in a neat line, square, cube, or hypercube, you can traverse the entire space, from corner to corner, with $d \times \left(n^{\frac{1}{d}} - 1\right)$ steps. This is because each vertex is $n^{\frac{1}{d}}$ nodes in length, and there are d vertices. If we have 64 molecules that vary on one dimension, and the Blackian demon starts at one end of the space, it has to take 63 steps to get to the other end of the space. If the 64 molecules vary on 2 dimensions and occupy an 8x8 square, then it takes only 14 steps to get from one corner of the square to the opposite corner. If the molecules vary on 3 dimensions and occupy a 4x4x4 cube, it takes only 9 steps to go from one corner to the opposite corner. The

Blackian demon will always take $d \times \left(n^{\frac{1}{d}} - 1\right)$ steps, or fewer to find the best molecule in the space. In contrast, the number of steps for super-HTS is always equal to *n*, the number of nodes (or, more precisely, *n-1*, if we assume that, like the demon, it is stepping from the first molecule).

When the search space is high dimensional (as is chemical space) and there is a very large number of nodes (as is the case for drug-like molecule*s)*, the Blackian demon is many orders of magnitude more efficient than super-HTS. For example, in a 10 dimensional space with 10^{40} molecules, the Blackian demon can search the entire space in 10^5 steps and steps (or less), while the brute force method requires 10^{40} steps. In this case, super-HTS is 10^{35} times less efficient, in terms of the number of search steps. The relationships between the size and dimensionality of the search space, and the relative efficiency of the Blackian demon versus super-HTS are illustrated in the figure (below). Lower values reflect fewer steps and higher efficiency (note the logarithmic axis).

Our view is that the old iterative method of drug discovery, at its best, approximates the Blackian demon, while conventional HTS, at its worst, approximates a sub-scale version of super-HTS. If there is a practical message here for drug discovery, it is to use iterative chemistry early and often, and to use a range of functional assays (for example, proxies for ADMET), rather than potency alone.

In format provided by Scannell *et al.* (MARCH 2012)

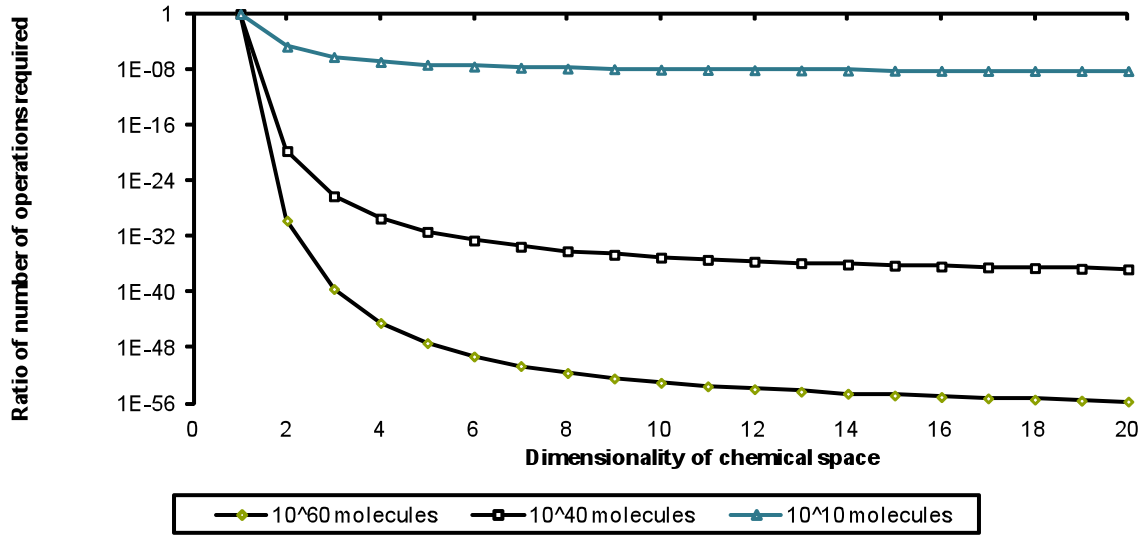

Figure | **Ratio of the number of steps required by the Blackian demon versus super-HTS to exhaustively search a space**. | The lines show the ratio of the number of steps required by the Blackian demon versus the number of steps required by super-HTS to exhaustively search for the best molecule in spaces of varying dimensionality (horizontal axis) in molecule universes of varying size (lines). Note the logarithmic vertical axis. If chemical space contains a very large number of molecules, and if the space is high dimensional, the Blackian demon requires many orders of magnitude fewer steps to find the best molecule.

References

1. Eckert, H. & Bajorath, J. Molecular similarity analysis for virtual screening: foundations, limitations, and novel approaches. *Drug Discov. Today* **12**, 225– 233 (2007).
2. Cramer, R. D. *et al*. "Lead hopping". Validation of topomer similarity as a superior predictor of similar biological activities. *J. Med. Chem*. **47**, 6777– 6791 (2004).
3. Dobson, C. M. Chemical space and biology. *Nature* 432, 824–828 (2004).
4. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).