A Simplified Model of Iterative Compound Optimization

John S. Delaney

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

john.delaney@syngenta.com

KEYWORDS

Compound development, iterative improvement, simplified recognition model.

ABSTRACT

This paper presents a simplified model of iterative compound optimization in drug/agrochemical discovery. Compounds are represented as binary strings, with project evolution simulated through random bit changes. The model reproduces key statistical features of real projects, including activity distributions and time-series characteristics. This framework enables statistical simulation of compound optimization, potentially aiding project planning and resource estimation.

INTRODUCTION

This work describes a simple model of iterative compound optimization as observed in a typical drug discovery project with a design-make-test-analyze (DMTA) cycle[1] and is a continuation of work described in this earlier paper[2]. That paper attempted to describe how a drug optimization project evolved as a self-avoiding walk through chemical space but provided no underlying mechanism for how the series of compounds changed step-to-step. The hope is that by filling in these gaps the model will become more useful in modelling real projects.

Structured endeavors involving multiple people or resources often employ some form of abstraction[3] to manage different levels of the operation, software development[4] being a conspicuous example. Drug/agrochemical development can also be viewed at

different levels of abstraction. A project can be defined as a series of related compounds made in time order with associated measured data for the target of interest and other properties that might pertain to a compounds suitability to become a commercial drug. Compound optimization in drug or agrochemical development is usually analyzed as a time-independent process with the focus on the structure-activity relationships within closely related sets of compounds. It is difficult to obtain the time a compound was added to a project outside of the databases of individual drug companies[5]. Even within large companies the fact that each project is an aggregate of hundreds or thousands of compounds means that their number is relatively limited compared to the number of compounds in their collections. A model of the statistical behavior of projects that goes beyond the available data might be useful.

Compound optimization within a project tends to follow a pattern. Projects start with a lead compound (sometimes more than one) found by design or random screening and proceed by making relatively small, incremental changes to the structure of the lead, measuring the response to the changes in an assay. This guides the chemists' choice of the next compound to make. The cycle repeats until either a compound suitable for progression is found or the patience of the project manager is exhausted, at which point the project ends. This paints a rather skeletal picture of a chemical project and does not capture some of the subtleties seen in real projects – series diverging in multiple directions at once, sudden serendipitous discoveries that radically change the projects focus etc. But for the purposes of modelling projects a simpler, reductive approach has been adopted. The aim is to create a "digital twin"[6] of a project that reproduces its statistical behavior rather than the structures of the individual compounds within it.

METHODS

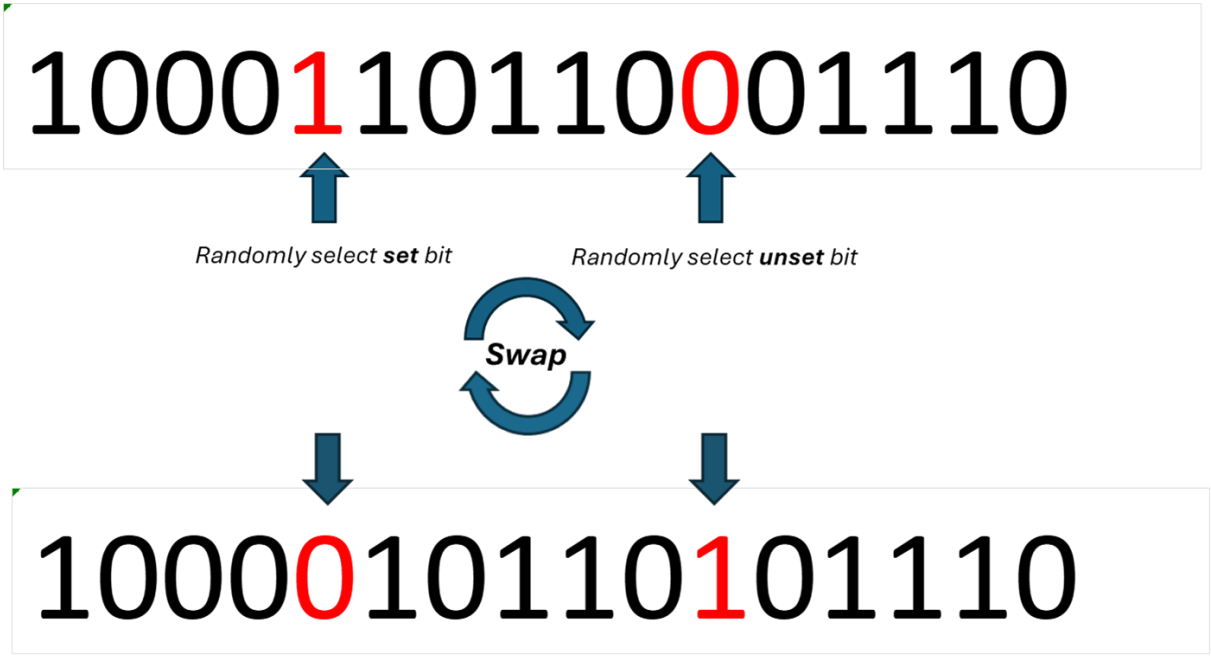
To build a model of a project we need to choose a way to represent molecules, how each compound interacts with the project's biological target to produce an assay signal and a way of capturing the way that compounds change in an ordered sequence as the project progresses. For this work compounds are described as fixed length binary strings (analogous to a substructural fingerprint[7]) and molecular recognition is achieved by matching a subset of set bits at fixed bit positions. This eschews directly using standard chemical structures and protein receptors in favor of a more abstract representation of chemistry. Such a reduced model of molecules and their binding to receptors has precedence in the work of Hann et al[8]. The authors of that paper fashioned a simple model of ligand binding by binary feature matching to draw out general principles of selecting compounds to test. This allowed exhaustive enumeration and simulation techniques to be applied to a population of model compounds interacting with a target.

The activity of a compound in this representation is defined by two counts. Firstly, the number of target bit positions matched sets the potential level of activity – the more positions matched, the greater the activity. Secondly, the target defines a few "kill" bits – matching one or more of these reduces the activity of that compound to zero, regardless of how many target bit positions are matched. The reasoning behind the "kill" bit is to mimic the effect of activity cliffs[9] that are frequently observed in real biological data[10]. In the example below (figure 1) the first four bit-positions (blue) are recognized by the assay, conferring activity as the sum of the set blue positions, in this case 1. The last bit-position (red) is a "kill" bit - if it is set the overall activity of the compound is set to zero regardless of the activity sum of the blue positions. The bit-positions in-between (black) can vary without affecting the assay result.

1000110110001110

(Figure 1)

Sequential changes that produce the compound sequence are produced by randomly swapping a set bit in the current fingerprint with an unset one. This produces the next fingerprint in the series with the same number of set bits and a Hamming distance[11] of 2 from the previous fingerprint. This is the smallest change between fingerprints that preserves the number of set bits – keeping the compound size constant (figure 2).



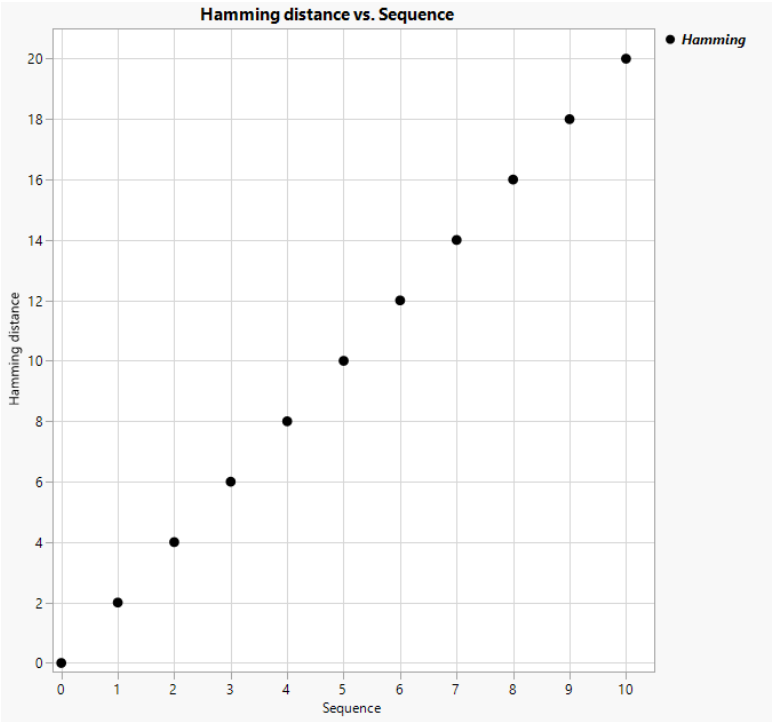
(Figure 2)

An example of a series of changes is shown below (figure 3), one set bit is unset, and an unset bit is set at each step, keeping the same total number of set bits. The changes from step 1 to step 2 are in red, 2 to 3 in green and 3 to 4 in blue.

Step	Bit String
1	10001101
2	00101101
3	00111001
4	00111010

(Figure 3)

The steps can be widened (i.e. more bit positions changed at each step) by sampling this generator at regular or random intervals. In the above example, sampling every fourth step leads to a Hamming distance of 6 between successive members of the series. The plot below (figure 4) shows the increase in Hamming distance (relative to compound 1 in the sequence) for a longer bit-string (256 bits, 40 bits set) as the sequence evolves away from its first member. The Hamming distance increases linearly with the number of steps away from the first string, at least for the first few steps.



(Figure 4)

For these simulations sampling was achieved by only outputting a string if a uniformly distributed random variable exceeded a censor value. For example, sampling every tenth member (on average) was achieved by setting a censor value of 0.1.

To initialize a sequence, we start with a weakly active compound (1 active position set, no kill bits set, the rest of the bit positions randomly assigned (set/unset) so that the overall string has the required bit density (number of set bits)). The series advances one step at a time as described above, an assay result being produced for each string. The evolving sequence and assay results constitutes the simulation of a chemical project. The method outline above has been implemented as a Python[12] program and can be found in the supplementary material (iterated_project_evolution.py). The current program outputs the calculated assay activities based on hard coded parameters and the individual bit-string values for a fixed number of compounds as a comma separated text file.

RESULTS

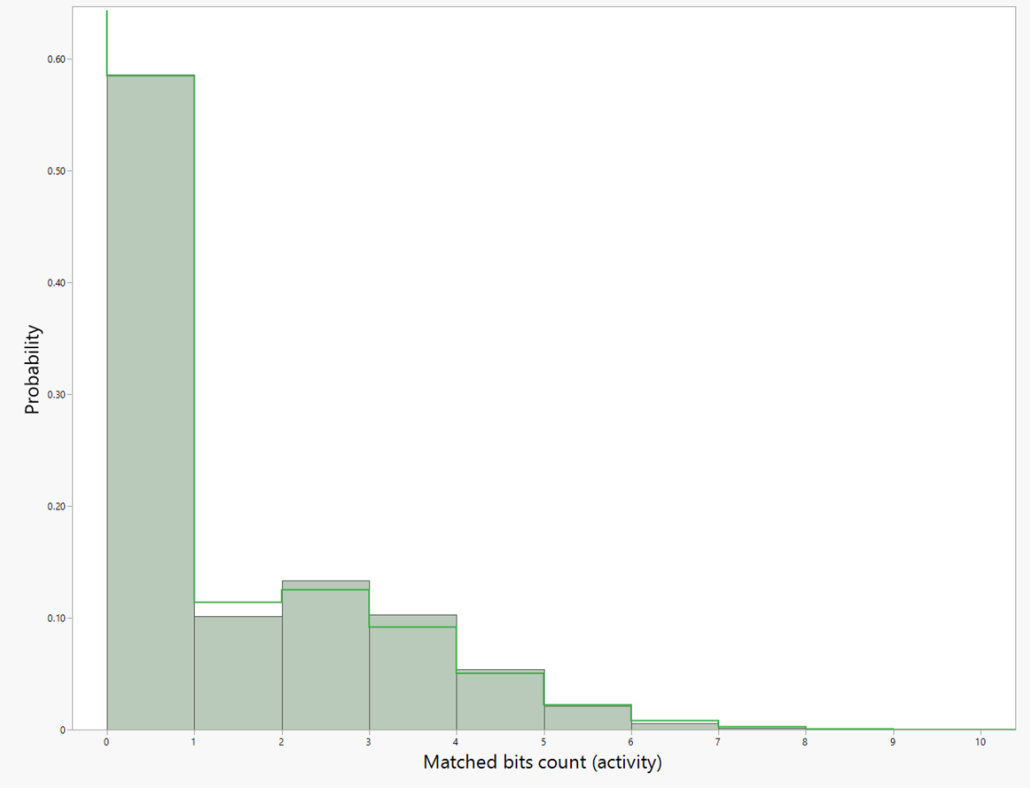
The output from the simulation is a stream of bit strings defining a path through chemical space and a list of assay results. To get a qualitative feel for whether the program was producing a path similar to that seen in real herbicide projects, an example project from Syngenta's internal database was profiled as follows. The time ordered list of compounds from the project were represented as 256-bit ECFP6 Morgan fingerprints[13] (Pipeline Pilot), a 2D UMAP generated (R[14] function tumap from the uwot package[15]) and a moving average applied (with a window size of 25 steps) to the UMAP coordinates[2]. A similar 2D UMAP was generated for the simulation output, allowing the plots to be compared (figure 5).

(Figure 5)

The purpose of these diagrams is to illustarte the broad similarity between the trajectories produced by real compound series (represented by fixed length binary fingerprints/strings) and a synthetic trajectory produced by the random tweak process.

To check whether the simulations produced good quantitative facsimilies of real project behaviours, we have focused on a couple of measures that can be applied to both real and synthetic projects. The first was to look at the overall distribution of assay results produced by projects, the second to examine the time evolution of projects using the Hurst exponent[16,17].

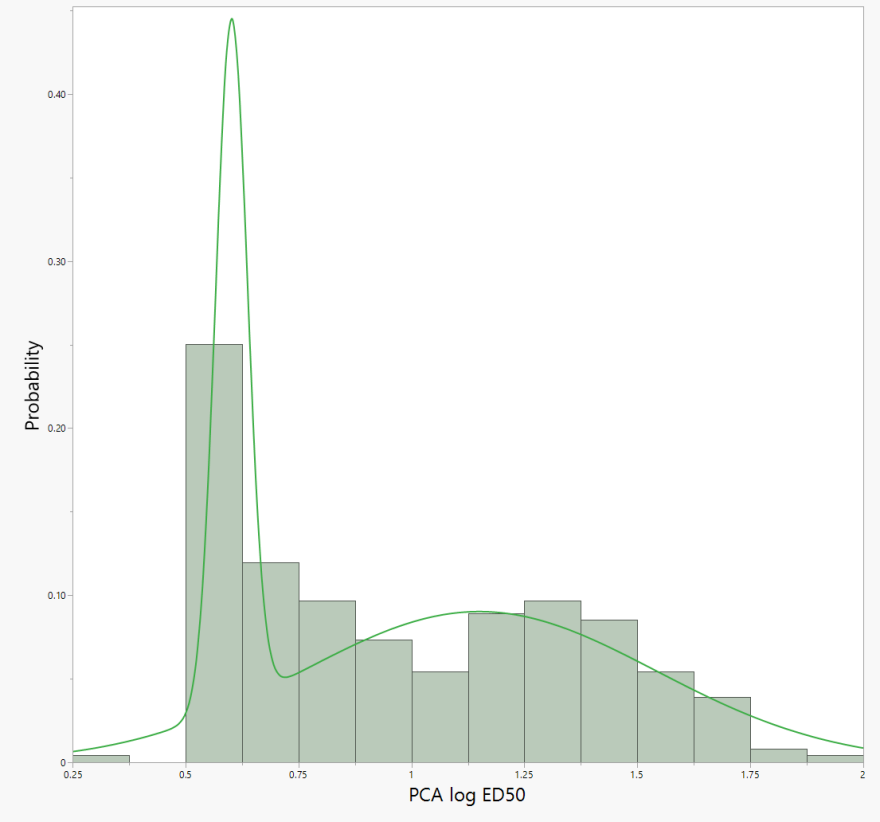
A typical distribution of simulated activities is shown below (figure 6 - 256 bit-string, 40 bits set, 15 active bits, 5 kill bits, sampled every 10th generation – 2000 simulation steps). The distribution was modelled using the distribution fitting function in JMP[18], the fitted Zero-Inflated (ZI) Poisson[19] is shown below (green).



(Figure 6)

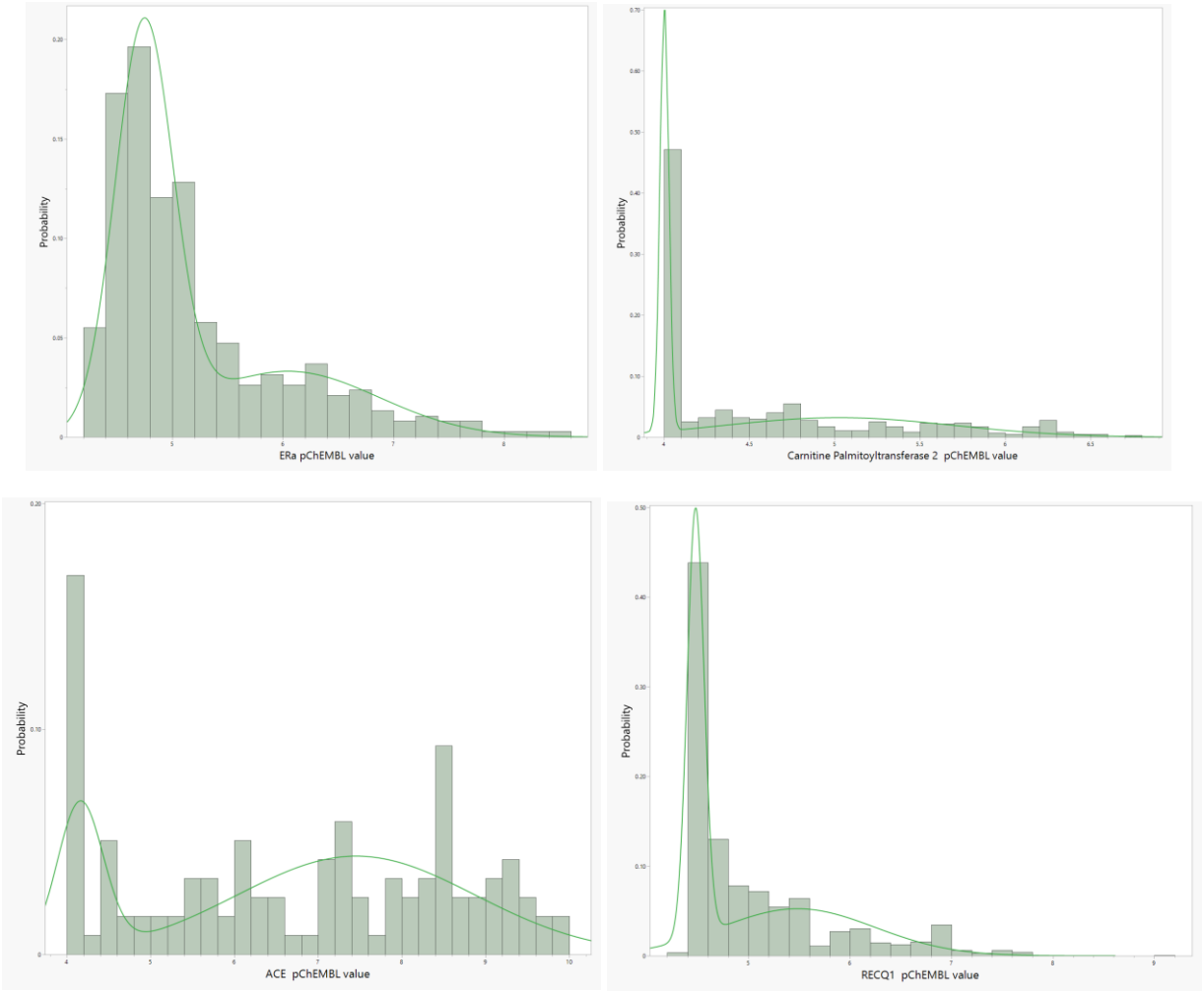
The bimodal ZI Poisson fitted the observed distribution better than a simple Poisson, illustrating the effect of the kill bits which produced a larger proportion of zero results than a unimodal distribution would have predicted.

For comparison, the activity distribution for a real herbicide project[20] was generated. The plot below (figure 7) has a bimodal 2 normal mixture distribution[21] (green) fitted.



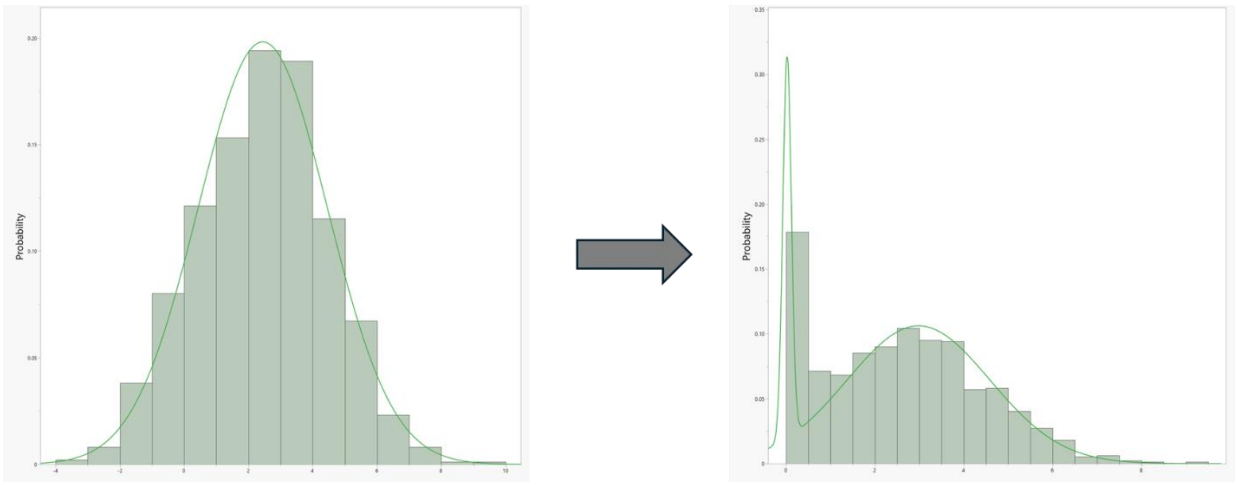
(Figure 7)

Multimodal activity distributions can also be observed in in-vitro assays, four examples from ChEMBL[22] are shown below (figure 8 - CHEMBL829540, CHEMBL1614199, CHEMBL3431931, CHEMBL648337, CHEMBL1613829), bimodal 2 normal mixtures (green) fitted[18]).



(Figure 8)

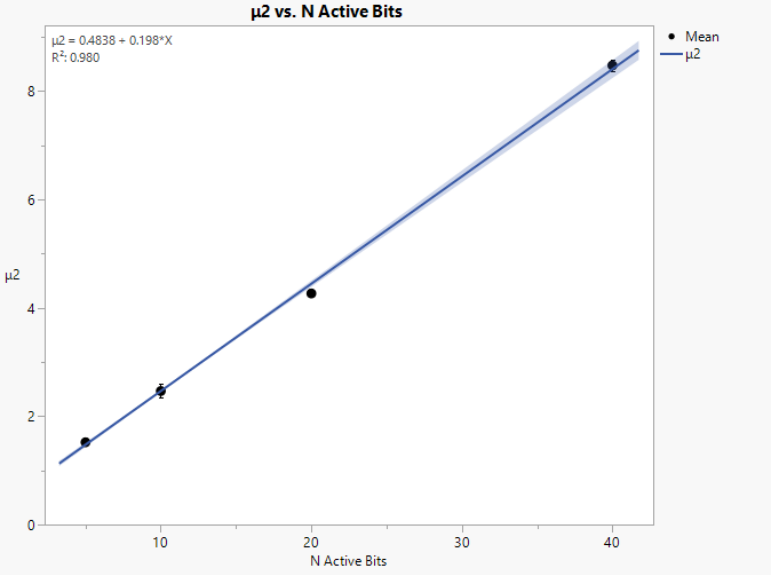
The bimodal nature of these observed distributions can be reasonably well explained by considering that real assays tend to have a limit in how low the activity they detect can be characterized. This may be caused by practical constraints (e.g. compound solubility) or simply that we're not interested in characterising degrees of inactivity (activity levels that have no commercial interest). The effect of applying this sort of censoring to a normal distribution is shown below (figure 9), with values below zero set to zero and a 2 normal distribution fitted.

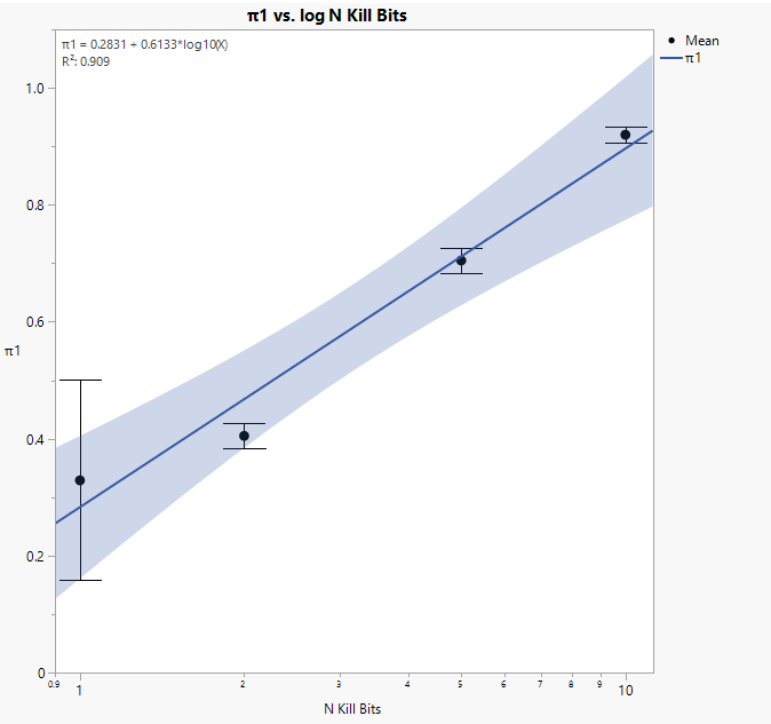


(Figure 9)

The simulations reproduce the multi-modality of assay results quite well, giving some confidence in the overall description of the process in terms of active and kill bits.This leads to a natural interpretation of the simulation's number of active (nactive) and kill (nkill) bits – nkill controls the proportion of the distribution that is censored by the assay lower limit, nactive determines the average activity in the second fitted normal distribution.

Determining the number of active and kill bits to set to mimic a given project is more involved. Earlier we fitted a 2 normal mixture to the activity distribution of a real project which yields 2 means (locations), 2 standard deviations (spreads) and a mixing parameter that gives the contributions of the two normal distributions to the overall mixed distribution. Applying the same technique to the outputs from eight simulations with a range of parameters (numbers of active bits varied between 5 and 40, kill bits between 1 and 10) and fitting the mixed normal distributions yields the following relationships (figure 10): -

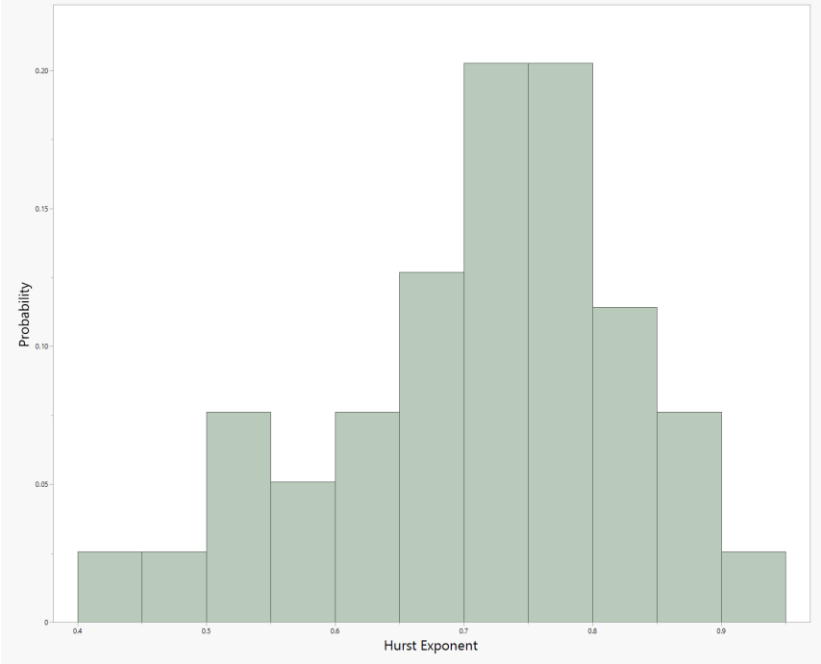




(Figure 10)

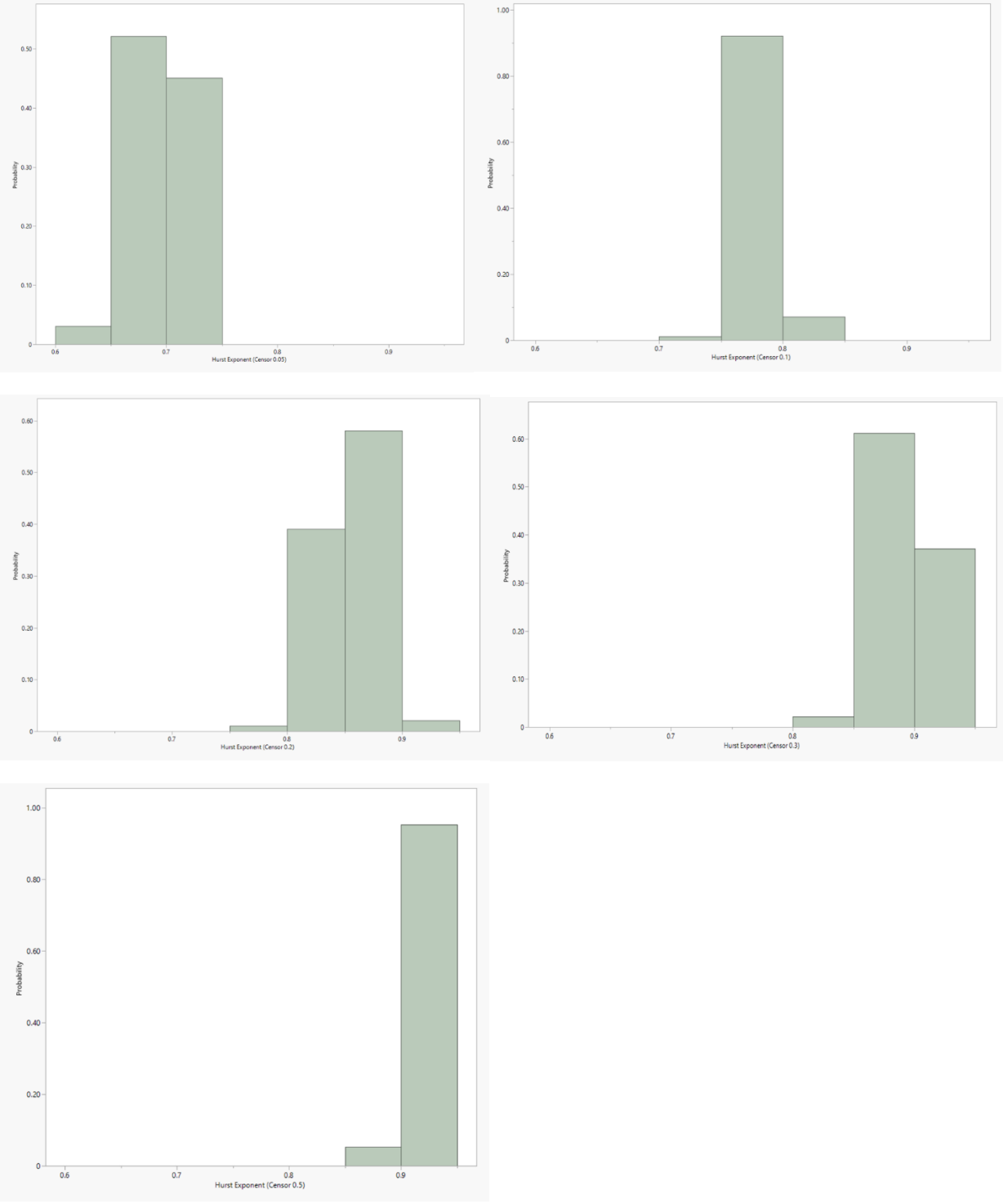
μ2 is the mean of the larger normal distribution, π1 is the proportion of the overall distribution derived from the smaller normal distribution.

The time evolution of a series project activity measures can be characterised by the the Hurst exponent. A random walk produces a time-series with a Hurst exponent close to 0.5 – for a process with long-term memory the value is higher, and lower for an antipersistent process. Harold Hurst was a hydrologist working in Egypt from 1906. He was interested in the yearly variation in the levels of the Nile and he noticed that the system, rather than being simply periodic or random, showed clustering of flood years and drought years. He devised a way of numerically characterising a time-series of observations as a single number, the Hurst exponent which takes values between 0 and 1. The Hurst exponent must be estimated from a time-series and the estimate is generally considered unreliable for series with fewer than a hundred points[23]. The Hurst exponents for time ordered series of PCA log ED50 activity values were calculated by the rescaled range method[24]. The distribution of Hurst exponents for 100 Syngenta herbicide projects is shown below (figure 11), the mean value is 0.72 and the standard deviation 0.11.



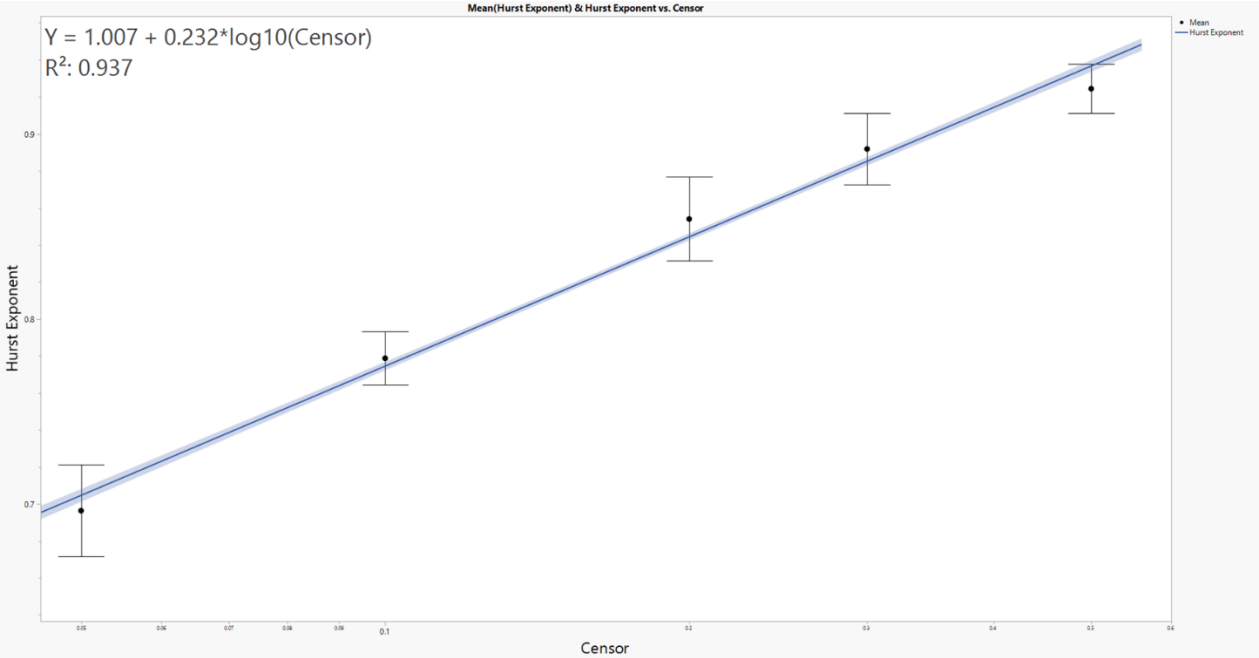
(Figure 11)

For comparison, the following plots (figure 12) show the distribution of Hurst exponents for 100 project simulations (256 bit-string, 40 bits set, 15 active bits, 5 kill bits) at five different levels of sampling (censor).


Probability
Hurst Exponent (Censor 0.05)
Hurst Exponent (Censor 0.1)
Hurst Exponent (Censor 0.2)
Hurst Exponent (Censor 0.3)
Hurst Exponent (Censor 0.5)

(Figure 12)

The Hurst exponent of each simulation can be tuned by changing the effective diveristy of the stream – increasing the gap (as measured by the Hamming distance between successive fingerprints) decreases the average Hurst exponent in a predictable way (figure 13).



(Figure 13)

The overall project simulation is thus controlled by three main parameters – numbers of active and kill bits, and the sampling rate – as well as the number of set bits and the bit-string length. The number of set bits and bit-string length for real projects can be derived from analysis of their Morgan fingerprints. The number of bits set in the simulation bit-string can be directly related to the median number of bits set in a real project – for the set of Syngenta herbicide projects this was 61 bits in a 256-bit Morgan fingerprint (ECFP6, Pipeline Pilot).

The three linear relationships above were used to map the measured values for real projects (Hurst exponent, μ2 and π1 – all derived from the time-ordered list of assay activities) to the simulation control parameters (censor value, nactive and nkill bit counts) allowing simulations that mimicked any given project to be run.

CONCLUSIONS

Studying the statistical behaviour of optimisation projects is hard because of the paucity of well-curated examples, even within large research organisations. A simplified model that characterizes the essential features of a project's evolution offers an

alternative. This work attempts to show that project characteristics such as step-to-step diversity and overall activity distribution can simulated by a simple model. The main features of the simulation output are: -

- A bimodal activity distribution similar to observed project assay data
- Parameters that can be fitted to project assay data distribution (nactive and nkill)
- Tuneable diversity, controlled by the censor value and adjusted to match the Hurst value seen in real projects

There seem to be two main uses for a model like this. The fitted assay activity distribution for a project can be used to produce a model with similar statistical properties, allowing multiple simulations to be run from randomly generated starting points. The aim would be to assess the range of possible project outcomes (e.g. the maximum activity observed in a given number of steps). This is somewhat analogous to equilibrium pricing of financial futures contracts ("fair price"[25]) – another way of framing the question could be "starting from this set of compounds (in a project with assay results), how many compounds would one need to make before finding one with a ten-fold improvement in assay result?". The fair price in this case might be the median number of compounds required over repeated randomized simulations. Another avenue would be to examine the effect of different project management strategies (e.g. early stopping) for a realistic range of model parameters – this is analogous to the work described in reference 2 but with a model more directly related to real world parameters.

The current simulation framework is very bare bones with no consideration of SAR or chemical design. As such it offers opportunities for improvement – e.g. some form of reinforcement learning (even a very basic, "MENACE"-like[26] system), applications of different stop policies (currently the simulation runs for a fixed number of steps) and improved parameter fitting.

ACKNOWLEDGMENT

The author would like to thank Dr Chris Baker for his helpful comments and encouragement in preparing this paper.

AUTHOR INFORMATION

Corresponding Author

*John S. Delaney - Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire. RG42 6EY. United Kingdom.

https://orcid.org/0000-0003-2218-1167

Present Addresses

†435A Woodham Lane, Woodham, Surrey. KT15 3QE. United Kingdom. Email: john.delaney.1998@gmail.com

REFERENCES

[1] Plowright, Alleyn T. Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discovery Today* **2012**, 17(1-2), 56-62. DOI: 10.1016/j.drudis.2011.09.012

[2] Delaney, John S. Modelling iterative compound optimisation using a self-avoiding walk. *Drug Discovery Today* **2009**, 14(3-4), 198-207. DOI: 10.1016/j.drudis.2008.10.007

[3] Hofstadter, Douglas. *Gödel, Escher, Bach.* **1979** Basic Books. ISBN 978-0-465-02656-2.

[4] Colburn, Timothy and Shute, Gary. Abstraction in Computer Science. *Minds and Machines* **2007**, 17 (2), 169–184

[5] Landrum et al. SIMPD: an algorithm for generating time splits for validating machine learning approaches. *Journal of Cheminformatics* **2023**, 15, 119. DOI: 10.1186/s13321-023-00787-9

[6] The Economist "Digital twins are fast becoming part of everyday life", 29 August 2024

[7] Chen et al. Concepts and applications of chemical fingerprint for hit and lead screening **2022**, 27(11), 103356

[8] Hann et al. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001,** *41,* 856-864. DOI: 10.1021/ci000403i

[9] Bajorath et al. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, 4, 11, 14360–14368. DOI: 10.1021/acsomega.9b02221

[10] Maggiora, GM. On outliers and activity cliffs—why QSAR often disappoints. *J Chem Inf Model*. **2006**, 46,1535. DOI: 10.1021/ci060117s

[11] R. W. Hamming, Error detecting and error correcting codes, *The Bell System Technical Journal* **1950**, vol. 29, no. 2, pp. 147-160. DOI: 10.1002/j.1538-7305.1950.tb00463.x

[12] Van Rossum, G. & Drake, F.L., 2009. Python 3 Reference Manual, Scotts Valley, CA: CreateSpace.

[13] Rogers, D and Hann, M. Extended-Connectivity Fingerprints, *J Chem Inf Model*. **2010**, 50, 5, 742-754. DOI: 10.1021/ci100050t

[14] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (accessed 2024-10-09)

[15] James Melville (2019). uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. R package version 0.2.2. https://CRAN.R-project.org/package=uwot (accessed 2024-10-09)

[16] Hurst, H. "Long Term Storage Capacity of Reservoirs" *Transactions of the American Society of Civil Engineers* **1951**, 116, 770-799. DOI: 10.1061/TACEAT.0006518

[17] Mandelbrot, BB and Hudson, RL "The (Mis)Behaviour of Markets" **2004** Profile Books.

[18] JMP®, Version 17.2.0. SAS Institute Inc., Cary, NC, 1989–2023.

[19] Lachin, John M. (2011), Biostatistical Methods: The Assessment of Relative Risks (Second ed.), Wiley, ISBN 978-0470508220

[20] Activity from a principal components analysis of log effective dose for 50% plant kill (log ED50) across a standard mix of weed species – more positive means more herbicidally active, i.e. lower log ED50.

[21] Robertson, CA and Fryer, JG. "Some descriptive properties of normal mixtures" **1969** *Skandinavisk Aktuarietidskrift*. **69** (3–4): 137–146. DOI: 10.1080/03461238.1969.10404590

[22] Overington, JP et al. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery **2012** *Nucleic Acids Res.*, 40, 1100-1107. DOI: 10.1093/nar/gkr777

[23] Karagiannis, T et al, Long-Range Dependence: Now you see it, now you don't! **2002**, *IEEE GLOBECOM - Global Internet Symposium*, IEEE Communications Society

[24] Di Matteo, T et al. Scaling behaviors in differently developed markets. Physica A 2003, 324,183-188. DOI: 10.1016/S0378-4371(02)01996-9

[25] Nasdaq Glossary https://www.nasdaq.com/glossary/f/fair-price (accessed 2024-10-09)

[26] Michie, Donald (November 1963). "Experiments on the Mechanization of Game-Learning Part I: Characterization of the Model and its Parameters". The Computer Journal. 6 (3): 232–236. DOI: 10.1093/comjnl/6.3.232