Predictive genomics: James Peach

AI technology for predicting biological mechanisms

We are living in a golden age of chemistry, enabled by data science and exemplified by the 2024 Nobel Prizes for protein design and structure prediction. However biology remains elusive, as reflected in the words of Nobel laureate David Baker: "Biology is still really hard. What people don't understand is that the problem of making a binder to a target is close to being solved. The real question is: what should you make binders to?[1] The answer to David's question has traditionally been answered through finding gene-disease associations in large genetic databases (Genome-wide association studies or GWAS). This has led to genetically informed drug development, which more than doubles the chance of clinical success.[2] However, GWAS ignores causation, relies on large numbers, and is biased towards common or high penetrance variants.

We have only found a small proportion of the genetic causes of disease suggested by heritability.[3] The initial burst of GWAS-driven discovery has given way to a dearth of truly novel targets. While half of the known druggable targets remain undeveloped, two thirds of global pharmaceutical R&D programmes are on commonly developed targets (with more than 10 drugs in development against each target) and one quarter on highly developed targets (with more than 50 drugs in development against each target).[4] The well-known targets EGFR and PD-1 for cancer, and GLP-1R for obesity, are all in about 200 active programmes.

This isn't because we're doing less development – R&D programmes have doubled over the last 10 years. It's not because we're short of new targets: some 2,500 remain undeveloped. It's because we're taking on fewer novel targets each year – down from 100 novel targets per year 10 years ago, to 30 per year going into development now. But the impact of drugging a new target can be huge – as GLP has shown. Given the increased patient impact and commercial value of first in class therapeutics, how can companies and investors respond to Dr Baker's challenge and increase their exposure to new targets?

Our company, OutSee Ltd, was founded to do this: our novel AI predictive technology, *Nomaly*, uses a different approach to genetics that can discover, validate and triage new targets and mechanisms. Work on the concept and building blocks by founder Julian Gough started over a decade ago at the UK's MRC Laboratory of Molecular Biology, with the development of a library of hidden Markov models (HMMs) leading to Superfamily, a database of structural, functional and evolutionary information for proteins[5], and dcGO, a comprehensive ontology database for protein domains.[6]

This enabled the development and production by Julian and CSO Chang Lu of *Nomaly*, which was patented in 2016 and published in *Nature comms* in 2023.[7] At a high level, *Nomaly* takes genetic information about an individual and suggests the gene, protein and mechanisms linked to any abnormal phenotype (such as a disease). *Nomaly* uses a single human genome input to predict whether the individual will be an outlier on a range of 20,000 ontology terms for phenotypes and diseases – hence the company name 'OutSee'. When these ontology terms are specific to a disease, and that disease appears in the clinical information of the individual, then *Nomaly*'s predictions illustrate the genetic cause and biological mechanisms for that disease. This check against clinical reality gives confidence in the predictive modelling within *Nomaly*. The approach is totally different from correlation, addresses uncommon variants of medium and low penetrance, allows additional biological information to be derived from existing genomic datasets, and enables insights from small patient cohorts.

Nomaly brings together multiple concepts into a single inseparable system, and is driven by the foundational biological datasets and artificial intelligence described in the *Nature Communications* paper and two decades of foundational papers on the individual concepts. *Nomaly* leverages structural and evolutionary intolerance of mutations in protein functional units within which they are identified, referring to a domain-centric platform for protein-phenotype relationships and compares to a background landscape of genomes. Hidden Markov models predict the scale of impact of a genetic change on a protein domain. Statistical information theory predicts which phenotypes and diseases will be impacted by this protein domain change, based on existing knowledge of molecular and cell biology captured in databases and ontologies. Solution of an eigen-problem – similar to the linear algebra used in spectral clustering, plus genetic distance matrices – show if this change would make the individual be an outlier versus the genetic background cohort. A universal distribution function prioritises across all of the potentially outlying diseases for an individual.

Existing non-associative approaches that investigate genotype-to-phenotype relationships mainly use supervised network models, learned from large genomic variant databases. However this requires very large datasets which currently are very sparse, and can be difficult to interpret. With *Nomaly*, instead of disease being used to train a model at an early step, the patient disease/phenotype is used as a final step to evaluate which predictions performed significantly better than expected by chance. The underlying protein knowledge, on which *Nomaly*'s ab initio models are based, can be examined to provide molecular insights into the predicted phenotype, identifying both the target and mechanism.

This work was validated in the peer-reviewed paper on a cohort of over two thousand individuals from direct-to-consumer genotyping such as 23andMe and Ancestry.com, another cohort of one thousand trios (2 parents & 1 child) from the Deciphering Developmental Disorders (DDD) dataset, and at a cellular level using the HipSci

stem cell bank. In the 23andMe cohort, *Nomaly* predictions were found to match self-reported conditions. *Nomaly*'s predictions from DDD were significantly matched to clinical annotations by doctors, and *Nomaly* identified both previously known causes and novel genetics. Exome data from HipSci donors predicted negative regulation of centrosome duplication, which was confirmed experimentally *in vitro* by collaborators at King's College London.

OutSee has since raised £2.4 million investment to run *Nomaly* on over a million diverse genomes from UK Biobank, Genomics England, AllofUs, Our Future Health, FinnGen, and smaller disease-specific cohorts. We have won over £0.6 million in precision medicine grants and almost $0.5 million in cloud computing awards to extend the input data from non-synonymous protein-coding variants to whole genome data, and to adapt the technology for somatic mutations including for cancer. We are now running half a million genomes a quarter and we have identified several credible novel targets, with the most promising internal candidate now progressed to early drug development.

We hope that the use of our technology will be a major contributor to answering Dr Baker's question, and enable biology to advance in the future at the pace that chemistry does now. We know the limitations of current methods, the scale of unexplained genetics and the impact of genetic evidence in development. We see the industry and investors concentrating on well-known targets, despite the value of novel targets. We welcome other advances from great teams such as Google DeepMind, some of which we hope we can adopt to improve our performance – but more importantly that will lead them and others to innovate alongside us in new and creative ways in this wide open space that has been under-exploited for too long. As individuals, we wish to drive the field away from marginal improvements and towards new approaches to discovery that will lead to cures. We are inspired by recent successes in rare diseases and neurology. We look forward to working with others who also believe that the golden age of genomics following the sequencing of the human genome in 2000 might have been delayed, but that it will ultimately realise the anticipated potential for dramatic improvements in human health.

References:

1. A conversation with Nobel Prize winner David Baker on the future of AI in protein design — and beyond. Available at https://endpoints.news/nobel-winner-david-baker-on-ai-in-biology-and-protein-design/ (Requires free account, accessed 5 Aug 2025)

2. Minikel, E.V., Painter, J.L., Dong, C.C. *et al*. Refining the impact of genetic evidence on clinical success. *Nature* 629, 624–629 (2024). https://doi.org/10.1038/s41586-024-07316-0

3. Manolio TA, i. Finding the missing heritability of complex diseases. *Nature*. 2009 Oct 8;461(7265):747-53. doi: 10.1038/nature08494. PMID: 19812666;

4.Is Biopharma Doing Enough to Advance Novel Targets? Available at https://www.lek.com/insights/hea/us/ei/biopharma-doing-enough-advance-novel-targets (Accessed 5 Aug 2025)

5. Gough, J *et al*, Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure, J. Mol. *Biol*. (2001) 313, 903-919. DOI: 10.1006/jmbi.2001.5080

6. Fang, H. and Gough, J. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. D536–D544 Nucleic Acids Research, 2013, Vol. 41, Database issue. DOI:10.1093/nar/gks1080

7 Lu, C., Zaucha, J., Gam, R. et al. Hypothesis-free phenotype prediction within a genetics-first framework. *Nat Commun* 14, 919 (2023). https://doi.org/10.1038/s41467-023-36634-6

This article was written by James Peach, Chief Operating Officer at OutSee Ltd.

