BARKing Up the Right Trees

Bayesian Optimisation for Designing Chemistry Experiments in Mixed and Conditional Domains

Toby Boyne[1], Jose Pablo Folch[2], Robert M. Lee[3], Behrang Shafei[3], Ruth Misener[1]

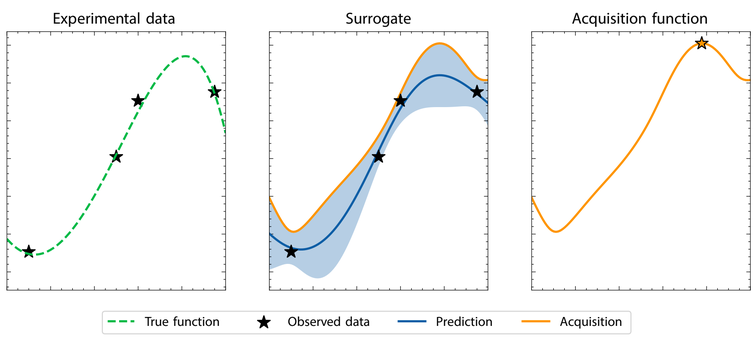
Introduction

Experiments are expensive.
Optimisation is core to many applied sciences, often where each experiment is expensive, and is a "black box":

- Chemistry: pick reaction conditions to maximise yield
- Machine learning: pick architecture to minimise errors

Bayesian optimisation is a framework to select experiments.

To find: $x^* = \arg\max_{x \in \mathcal{X}} f(x)$ → Solve iteratively: $x_{t+1} = \arg\max_{x \in \mathcal{X}} \alpha(x \mid \mathcal{D})$

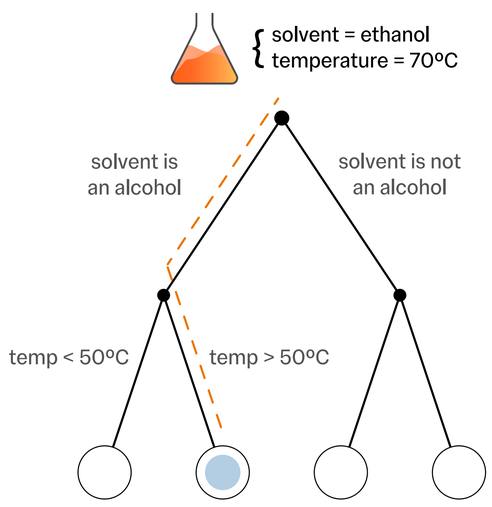


Challenges

Standard BO uses Gaussian process (GP) surrogates: predict the performance of a new experiment according to its similarity to past observations, $k(x, x')$

Standard surrogates face challenges for real-world chemistry:
1. **Conditional features** - inputs that depend on other inputs.
2. **Mixed domains** - continuous and discrete inputs.
3. **Molecular inputs** - comparing different substances.

All three challenges can be tackled by **decision trees**: models that use (learned) rules to assign data to leaves.



Conclusions

Machine learning tools let us **accelerate experimental campaigns** in chemistry.

Trees can improve how we model real-world experiments:
- Capturing interactions between continuous and categorical inputs.
- Understanding the hierarchy of how inputs interact.
- Covering a diverse search space.

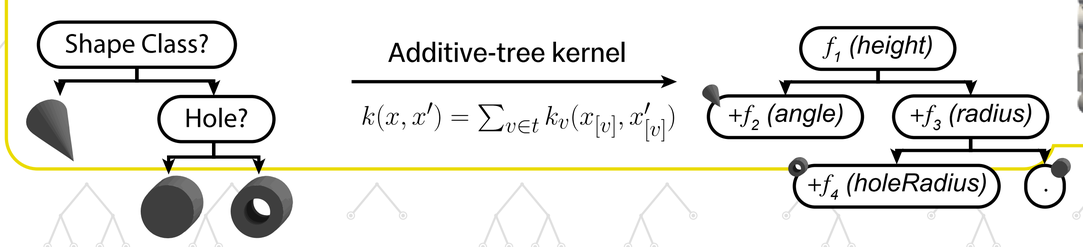
These advancements help bridge the gap between theory and application. Our methodology contributes are being incorporated into BoFire [3], an open-source Python package for BO in chemistry.

Catalyst Design

Task: design catalyst tablets to fill a reactor:
- We can control the shape of the tablet.
- We want to maximise the rate of reaction.

Challenge: catalyst classes have different parameters. We want to share information across classes, to be more efficient with our experiments.

Solution: use a tree structure to encode which inputs are shared between different classes. A pair of catalysts can jointly model their shared parameters.

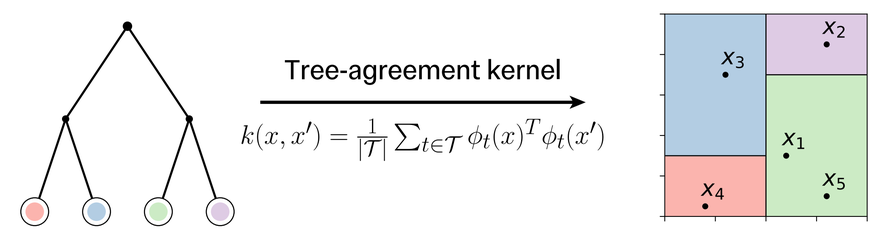


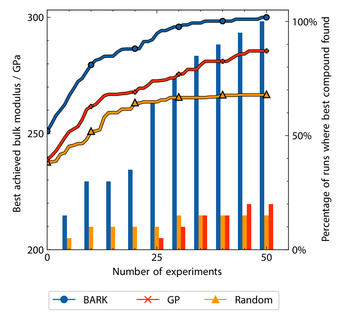
BARK

Task: optimise reaction conditions for mixed spaces:
- Continuous inputs (temperature, concentration)
- Discrete inputs (choice of solvent, catalyst)

Challenge: standard surrogates in BO are purely continuous.

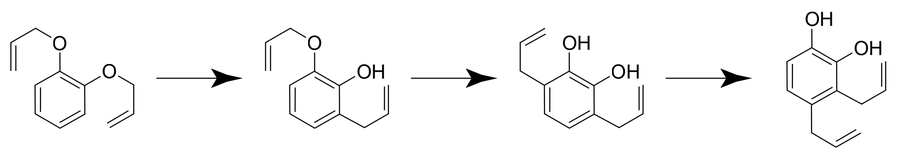
Solution: use tree-agreement kernels [1]. This combines the expressivity of tree models in mixed domains with uncertainty quantification of GPs.





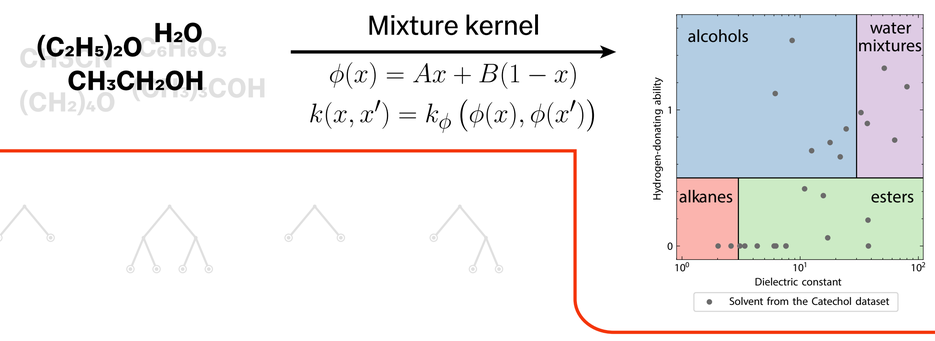
Solvent Selection

Task: replace toxic or unsustainable solvents with green alternatives. We benchmark existing methods on a flow chemistry dataset [2], generated by collaborators at SOLVE Chemistry, a London-based startup.



Challenge: search space of all solvents - and solvent mixtures - is too large.

Solution: use solvent properties to group solvents, and predict performance.



Affiliations

[1] Imperial College London; [2] SOLVE Chemistry; [3] BASF SE

Toby Boyne is funded by StatML CDT, grant no. EP/Y034813/1; IConIC, grant no. EP/X025292/1; BASF SE.

Contact: t.boyne23@imperial.ac.uk

References

[1] T. Boyne et al. BARK: A Fully Bayesian Tree Kernel for Black-box Optimization. ICML 2025.
[2] T. Boyne et al. The Catechol Benchmark: Time-series SolventSelection Data for Few-shot Machine Learning. NeurIPS 2025.
[3] J. P. Dürholt et al. BoFire: Bayesian Optimization Framework Intended for Real Experiments. JMLR 2025.