Supporting Information

Engineering stable and efficient ketol-acid reductoisomerases for industrial biotransformations using ancestral sequence reconstruction

Oscar Paredes Trujillo[a], Gabriel Foley[a], Sebastian Porras[a], Georgina H. Joyce[a], Andrew Douw[a], Sanjana Tule[a], Hannes Ehlert[a], Lachlan Asser[a], Ulban Adhikary[a], Samuel Davis[a], Damian Hine[b], Matthew Marzo[c], Kent Evans[c], Joe Ley[c], Volker Sieber[d,a], Luke Guddat[a], Mikael Bodén[a,*], Gerhard Schenk[a,e,*]

[a] School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia

[b] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD 4072, Australia

[c] Gevo Inc, Englewood, CO 80112, USA

[d] Chair of Chemistry of Biogenic Resources, Technical University of Munich, Campus Straubing, 94315 Straubing, Germany

[e] Australian Institute of Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD 4072, Australia

* Corresponding authors:

Mikael Bodén: m.boden@uq.edu.au

Gerhard Schenk: schenk@uq.edu.au

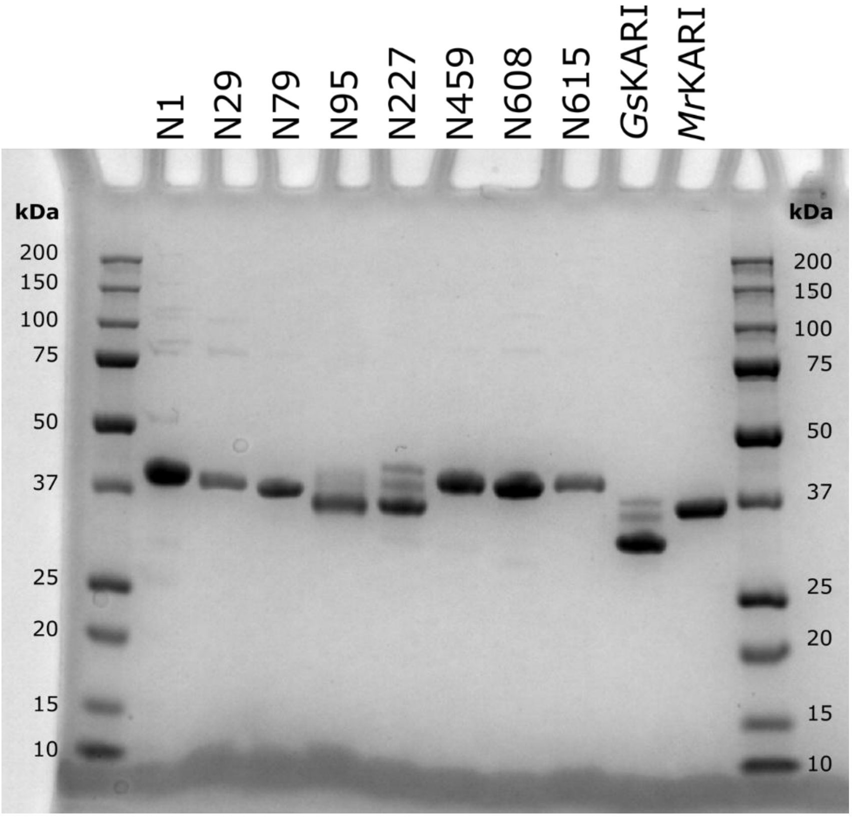


Figure S1: Evaluation of purified ancestral and extant KARIs by SDS-PAGE analysis. Enzymes were purified using a 1-step affinity chromatography procedure.

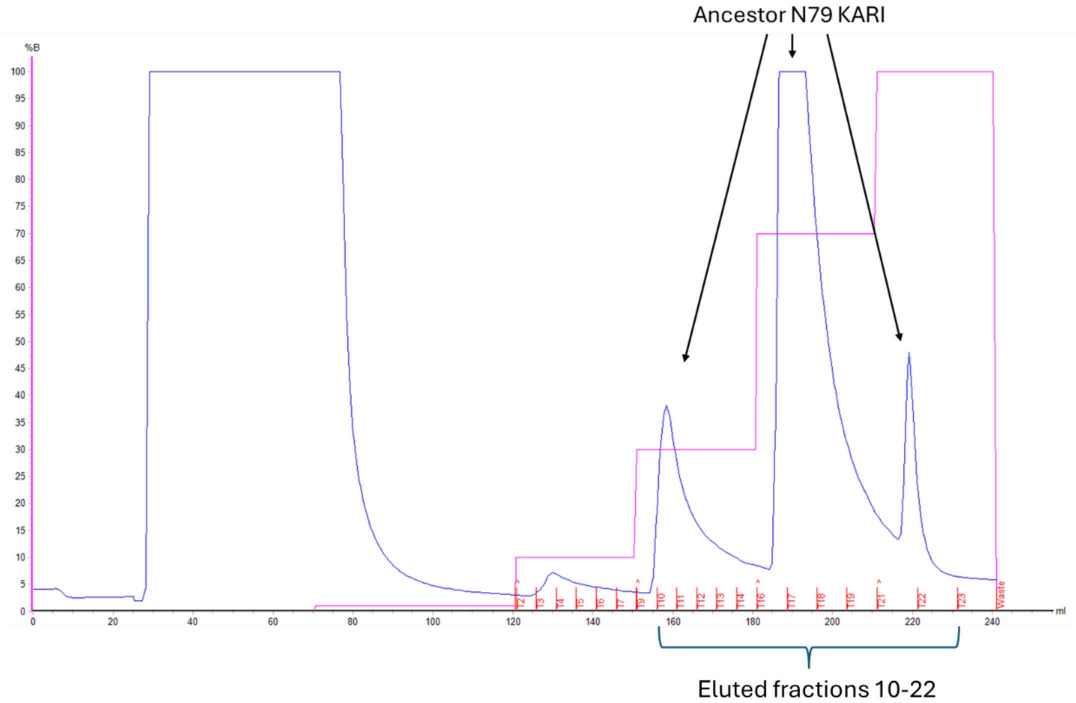


Figure S2: FPLC chromatogram after purification of ancestor N79. Pure KARI was eluted in fractions 10-22. Arrows point to the KARI elution peaks. Approximately 200 mg of enzyme were recovered from a 2 L culture.

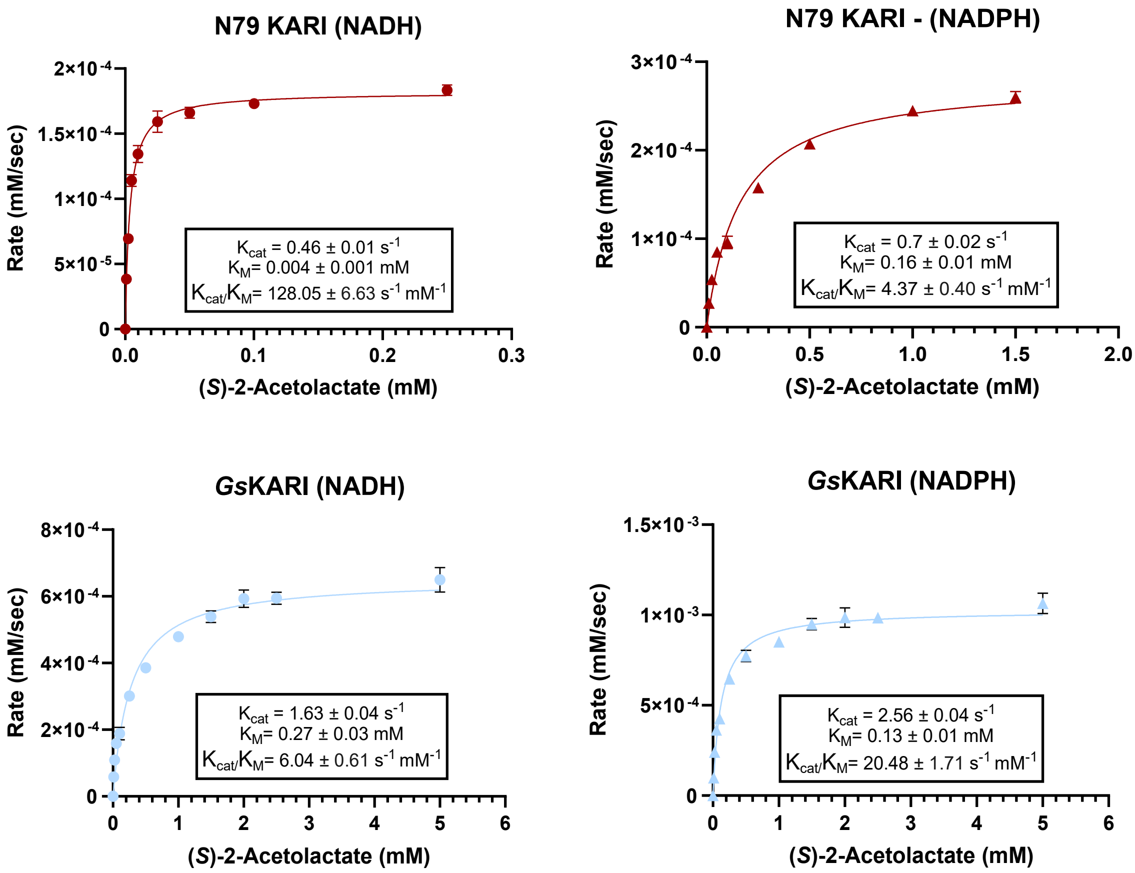


Figure S3: Representative data from catalytic assays with ancestor N79 (top) and *Gs*KARI (bottom), measured in presence of NADH or NADPH. Data were fitted to the Michaelis-Menten equation. Relevant catalytic parameters are shown in the boxes.

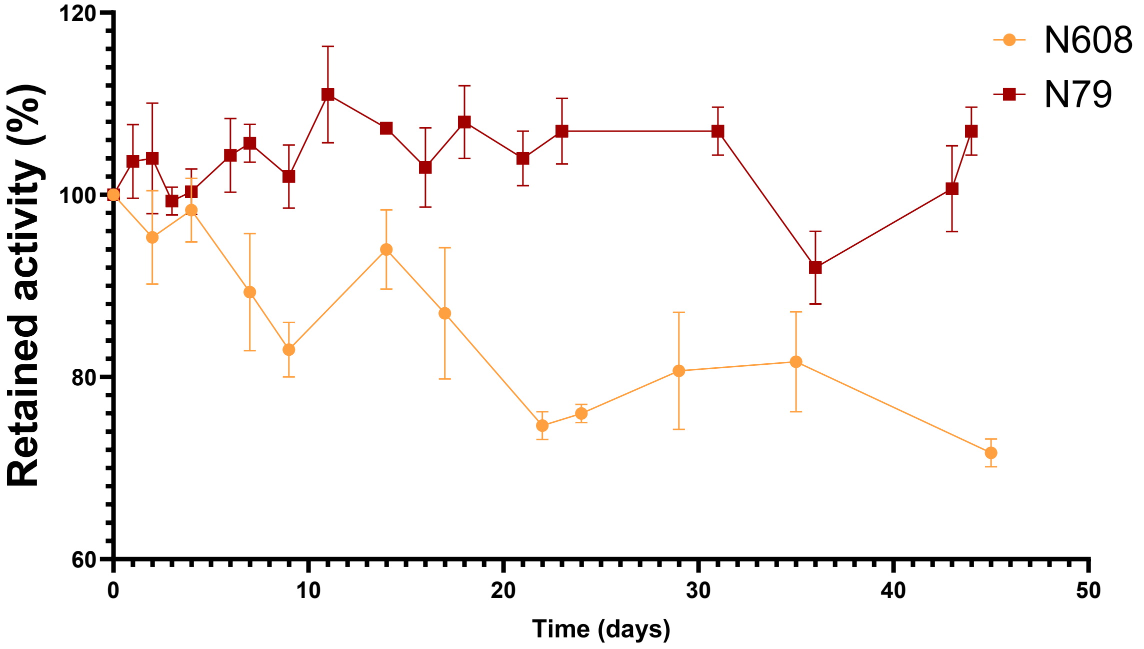


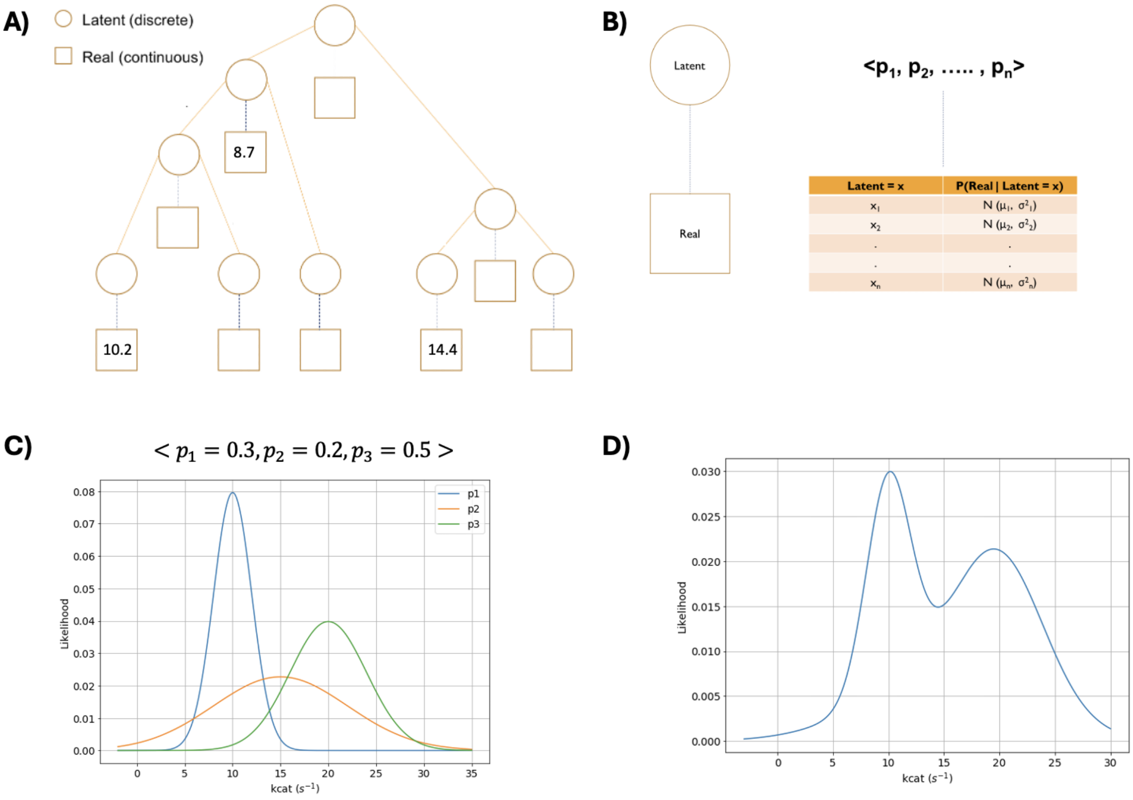
Figure S4: Thermal stability studies over 45 days. Ancestors N79 and N608 were incubated at 37 °C to evaluate if these are commercially viable candidates and overcome the minimum total turnover number (TTN $>10^6$) required for industrial deployment. TTN refers to the total number of conversions from substrate to product that an enzyme can do in its lifetime. Retained activity of the enzymes is shown as percentages of corresponding activity at 25 °C at each timepoint. Assays were run at 25 °C, and the mean of three replicates are displayed with standard error bars.

Modelling propagation of whole-of-protein properties with TreeGazer

To expand the view from real, experimental values (*i.e.* catalytic parameters) of a subset of members of the Class I KARI enzyme family we developed a tool TreeGazer that adopts classical evolutionary models to operate on a nominated set of discrete, latent states suited to the experimental property. TreeGazer is integrated into the GRASP Suite of tools [7,68] and described below. To infer real-valued properties for all branch points in a user-provided tree, TreeGazer first constructs a Bayesian network using a mixture of discrete and continuous nodes. Each branch point in the phylogenetic tree corresponds to a node in the Bayesian network that is conditioned on its parent, matching the structure of the phylogenetic tree (**Fig. S5A**). The specific nodes map to ancestral (internal) and extant nodes on a phylogenetic tree and specify discrete latent states. Each latent state specifies a Gaussian density, with the set of all latent states attempting to capture how an arbitrary trait varies across the tree (**Fig. S5B**). The probability of transitioning between these latent states is adapted from the Jukes-Cantor's model for nucleotide substitution where rates of transitioning to any other latent state are uniform. This framework has been generalised to include any number of latent states.

A continuous node takes on real values and is conditioned on a latent node. Predictions of real values can be generated *via* joint reconstruction or a marginal reconstruction. For a joint reconstruction, the maximum likelihood latent state is assigned for each latent node, and the prediction of the real node comes from sampling the corresponding Gaussian distribution for that latent state (**Fig. S5C**). For a marginal reconstruction, variable elimination is used to determine a marginal probability distribution of the latent states. A Gaussian mixture can then be used to predict values by using the probabilities from the marginal distribution to weight each latent Gaussian distribution (**Fig. S5D**).

The assumptions of modelling protein traits in this way can be directly tied to a biochemical context. For nodes that share the same latent state, a Gaussian distribution captures the normally distributed trait variation between homologous proteins as a result of small sequence changes. Different latent states serve to capture more significant changes that can arise from major mutations; note however, states are entirely independent of sequence data.



Latent = x	P(Real | Latent = x)
x_1	N (μ_1, $\sigma^2{}_1$)
x_2	N (μ_2, $\sigma^2{}_2$)
.	.
.	.
x_n	N (μ_n, $\sigma^2{}_n$)

Figure S5: A: A depiction of the Phylo-Bayesian network. The latent nodes (discrete) mimic the structure of the phylogenetic tree and real (continuous) nodes are populated with known features such as kinetic properties. B: The user can specify any number of latent states. Each

latent state maps to a Gaussian distribution with parameters learnt from the data. C: Visualisation of different Gaussian distributions for a model with three latent states. The marginal distribution shows the probability of each of those states for a latent node on the tree. D: The probabilities from the marginal distribution can be used to weight each latent state to create a Gaussian mixture.

See supporting information **Table S1** to find ancestor sequences.

Data that supports TreeGazer, ASR and phylogenetic analysis are available from the corresponding author upon request.