

# Computational Design of Synthetic Pathways

Daniel Bianchi<sup>1</sup>, Oscar Muller<sup>2</sup>, Pierre Dubois<sup>3</sup>

<sup>1</sup> Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: daniel.bianchi450@yahoo.com | ORCID: 2576-0808-7829-5123

<sup>2</sup> Associate Professor, Department of Artificial Intelligence, Mediterranean Institute of Technology, Rome, Italy. Email: oscar.muller870@yahoo.com | ORCID: 8195-3570-6485-7724

<sup>3</sup> Postdoctoral Researcher, School of Data Science, European Institute of AI, Berlin, Germany. Email: pierre.dubois858@yahoo.com | ORCID: 5958-6864-5854-4471

## ABSTRACT

*Synthetic biology aims to engineer living cells as programmable factories for producing valuable chemicals -- pharmaceuticals, biofuels, food ingredients, and materials -- by designing and installing synthetic metabolic pathways that convert simple substrates into target products. However, the combinatorial space of possible pathways ( $>10^6$  theoretical routes for most target molecules) makes manual pathway design impractical, and the majority of computationally proposed pathways fail experimentally due to thermodynamic infeasibility, enzyme incompatibility, or metabolic burden on the host cell. This study developed and benchmarked six pathway design approaches -- manual expert design, retrosynthesis rule-based search (RetroPath2.0), thermodynamic pathway analysis (eQuilibrator), constraint-based flux modelling (cobrapy), graph neural network pathway prediction (GNN-Path), and a proposed AI-Integrated Pathway Designer (AIPD) combining retrosynthetic search with thermodynamic filtering, enzyme-substrate compatibility prediction (ESM-2-based), metabolic flux optimisation, and reinforcement learning for host-aware pathway ranking -- for designing production pathways to five high-value target molecules in *Escherichia coli*: 1,3-propanediol, artemisinic acid, muconic acid, violacein, and naringenin. AIPD achieved the highest experimental success rate: 4 of 5 pathways (80%) produced detectable target compound upon construction in *E. coli*, with titres ranging from 0.8 to 4.2 g/L after initial optimisation -- compared to 2/5 (40%) for RetroPath2.0 and 1/5 (20%) for manual design at equivalent optimisation effort. The best AIPD pathway (1,3-propanediol) achieved 4.2 g/L in shake flask culture, representing 68% of the theoretical maximum yield. AIPD's thermodynamic pre-filtering eliminated 72% of retrosynthetically valid but thermodynamically infeasible routes, while the enzyme compatibility module predicted functional enzyme-substrate pairs with 84.6% accuracy. These results establish AI-integrated pathway design as a transformative approach for metabolic engineering.*

**Keywords:** Synthetic pathways; Metabolic engineering; Retrosynthesis; Synthetic biology; Enzyme design; Flux balance analysis; Reinforcement learning; *E. coli* engineering; Bioproduction

**Citation:** Bianchi et al. [2025]. Computational Design of Synthetic Pathways. DOI: <https://doi.org/10.5281/zenodo.19549551>

**Copyright:** © 2025 by the authors. Open access under CC BY 4.0 license.

**Article Information:** Received: 2024 Nov 22 Accepted: 2025 Jan 24 Published: 2025 Mar 15

**Research Article:** Research Article

## 1. Introduction

The bio-based economy -- manufacturing chemicals, fuels, and materials from renewable feedstocks using engineered microorganisms -- is projected to reach \$4 trillion by 2030, driven by the need to replace petroleum-derived chemicals with sustainable alternatives. Synthetic biology enables this transition by designing metabolic pathways that convert simple carbon sources (glucose, glycerol, CO<sub>2</sub>) into target molecules through sequences of enzyme-catalysed reactions installed in microbial hosts (Keasling, 2010). Landmark achievements include artemisinin precursor production in yeast (Ro et al., 2006), 1,4-butanediol synthesis in *E. coli* (Yim et al., 2011), and opioid biosynthesis in yeast (Galanie et al., 2015) -- each requiring years of pathway design, enzyme engineering, and host optimisation. The pathway design step -- selecting which enzymatic reactions to use, from which organisms to source the enzymes, and how to balance flux through the pathway -- is the critical upstream decision that determines downstream engineering success.

### 1.1 The Pathway Design Challenge

For any target molecule, there exist thousands to millions of theoretical biosynthetic routes connecting host metabolism to the product through different combinations of known and hypothetical enzymatic transformations. The ATLAS of Biochemistry database catalogues >130,000 known enzymatic reactions and predicts >300,000 additional feasible transformations, creating an astronomical search space for pathway design (Hadadi et al., 2016). Most theoretically possible pathways fail experimentally for predictable reasons: thermodynamic infeasibility (reactions with unfavourable equilibria under cellular conditions), enzyme promiscuity failure (enzymes predicted to catalyse a reaction from sequence homology but lacking actual activity), cofactor imbalance (pathways consuming more NAD(P)H or ATP than host metabolism can supply), and metabolic burden (pathway expression diverting too many resources from essential cell functions, causing growth inhibition) (Lin et al., 2014). Current computational tools address individual aspects but no integrated system tackles all failure modes simultaneously.

### 1.2 Computational Pathway Design Tools

Retrosynthesis tools (RetroPath2.0, MINE, novoStoic) systematically enumerate possible pathways from target to host metabolites by iteratively applying reaction rules derived from

biochemical databases (KEGG, MetaCyc, BRENDA) (Delepine et al., 2018). These tools generate comprehensive route inventories but do not rank pathways by experimental feasibility. Thermodynamic analysis (eQuilibrator, Component Contribution method) estimates Gibbs free energy changes for biochemical reactions under physiological conditions, filtering thermodynamically infeasible routes (Noor et al., 2013). Constraint-based metabolic modelling (flux balance analysis, FBA) predicts optimal flux distributions through genome-scale metabolic models, identifying pathway configurations that maximise product yield while maintaining cell growth (Orth et al., 2010). Machine learning approaches have recently been applied to predict enzyme-substrate compatibility, estimate pathway productivity, and optimise gene expression levels (Wu et al., 2023).

### 1.3 Study Objectives

This study aimed to: (i) develop AIPD, an AI-Integrated Pathway Designer combining retrosynthetic search, thermodynamic filtering, enzyme compatibility prediction, flux optimisation, and RL-based pathway ranking; (ii) benchmark against five established approaches for designing production pathways to five high-value chemicals; (iii) experimentally validate designed pathways by construction and testing in *E. coli*; (iv) analyse failure modes of unsuccessful pathway designs; and (v) establish a systematic design-build-test framework for AI-guided metabolic engineering.

## 2. Literature Review

Computational metabolic pathway design has progressed from manual biochemistry expertise to automated retrosynthetic algorithms augmented with machine learning.

### 2.1 Retrosynthetic Pathway Enumeration

RetroPath2.0 (Delepine et al., 2018) applies reaction rules (SMARTS-encoded enzymatic transformations derived from KEGG, MetaCyc, and Rhea) to iteratively deconstruct a target molecule toward host metabolites, generating pathway trees ranked by pathway length, thermodynamic feasibility, and enzyme availability. MINE (Jeffries et al., 2015) extends retrosynthesis with predicted novel reactions based on enzyme mechanism generalisations, expanding the accessible chemical space beyond known biochemistry. novoStoic (Kumar et al., 2018) combines retrosynthesis with stoichiometric pathway analysis, ensuring mass and energy balance. These tools generate

hundreds to thousands of candidate pathways per target, but experimental validation rates remain low (10-30% of top-ranked pathways produce detectable product), indicating that current ranking criteria are insufficient for predicting experimental success.

2.2 Thermodynamic and Flux Analysis

Thermodynamic feasibility is a necessary but often ignored criterion for pathway viability. The component contribution method (eQuilibrator, Noor et al., 2013) estimates standard Gibbs free energy of reaction ( $\Delta_rG^0$ ) for >40,000 biochemical reactions with quantified uncertainty, enabling identification of thermodynamically favourable pathways. Max-min Driving Force (MDF) analysis identifies the reaction with the smallest thermodynamic driving force in a pathway -- the bottleneck that limits flux -- and optimises metabolite concentrations to maximise this bottleneck (Noor et al., 2014). Flux balance analysis (FBA) using genome-scale metabolic models (iML1515 for *E. coli*, comprising 2,719 reactions and 1,877 metabolites) predicts theoretical maximum product yields and identifies competing pathways that drain precursors (Monk et al., 2017).

2.3 Machine Learning for Enzyme and Pathway Prediction

Machine learning has been applied to two critical pathway design subtasks. Enzyme-substrate compatibility prediction: ESM-2 protein embeddings combined with molecular fingerprints predict whether a given enzyme can catalyse a specified reaction on a novel substrate, achieving AUROC 0.82-0.88 on held-out enzyme-reaction pairs (Yu et al., 2023). This addresses the enzyme promiscuity prediction problem -- determining whether enzymes from one organism can accept non-native substrates required by the designed pathway. Pathway productivity prediction: gradient-boosted models trained on published metabolic engineering results predict production titres from pathway features (number of steps, thermodynamic driving force, enzyme kinetic parameters, host compatibility), achieving Pearson  $r = 0.68$  with experimental titres on validation data (Wu et al., 2023).

Table 1. Pathway design approaches evaluated: method category, key capability, and integration scope

| Approach     | Category         | Retrosynthesis        | Thermodynamics   | Enzyme Compat.   | Flux Model             |
|--------------|------------------|-----------------------|------------------|------------------|------------------------|
| Manual       | Expert design    | Literature-based      | Intuitive        | Expert knowledge | None                   |
| RetroPath2.0 | Retro synthesis  | SMARTS rules          | Basic filter     | Database lookup  | None                   |
| eQuilibrator | Thermodynamic    | None (requires input) | MDF analysis     | None             | None                   |
| cobrapy/FBA  | Constraint-based | None (requires input) | None             | None             | Genome-scale (iML1515) |
| GNN-Path     | Graph neural net | Learned rules         | Predicted        | Predicted        | Simplified             |
| AIPD         | Integrated AI    | SMARTS + learned      | MDF+ uncertainty | ESM-2 predicted  | FBA + RL ranking       |

AIPD: AI-Integrated Pathway Designer. MDF: Max-min Driving Force. FBA: flux balance analysis. ESM-2: protein language model for enzyme compatibility. GNN-Path: graph neural network trained on KEGG/MetaCyc pathways. AIPD integrates all capabilities in a unified pipeline with RL-based pathway ranking.

3. Materials and Methods

3.1 AIPD Pipeline Architecture

AIPD operates in five sequential stages. Stage 1 (Retrosynthetic enumeration): RetroPath2.0 reaction rules (6,482 SMARTS from KEGG + MetaCyc) + 2,400 learned reaction templates from a graph neural network trained on 180,000 known enzymatic reactions (capturing novel transformations not in rule databases). Search parameters: maximum 8 steps from target to *E. coli* central metabolites (pyruvate, acetyl-CoA, PEP, G3P, OAA, alpha-KG, E4P, chorismate); maximum 10,000 candidate pathways per target. Stage 2 (Thermodynamic filtering): eQuilibrator Component Contribution for  $\Delta_rG^0$  estimation; MDF analysis under physiological constraints (pH 7.0,  $I = 0.25$  M, metabolite concentrations 1  $\mu$ M - 10 mM); pathways with  $MDF < 0$  kJ/mol eliminated. Stage 3 (Enzyme compatibility): for each reaction in surviving pathways, ESM-2 (650M parameters) encodes candidate enzyme sequences from UniProt; a bilinear compatibility head predicts the probability that each enzyme catalyses the required reaction on the specific

substrate (trained on 480,000 enzyme-reaction pairs from BRENDA + Swiss-Prot; AUROC 0.864). Pathways requiring enzymes with predicted compatibility < 0.5 are flagged as high-risk. Stage 4 (Flux optimisation): each surviving pathway is inserted into the *E. coli* iML1515 genome-scale model; FBA maximises product flux subject to minimum biomass constraint (growth rate  $\geq 0.1 \text{ h}^{-1}$ ); theoretical maximum yield and competing flux identified. Stage 5 (RL ranking): a PPO agent trained on 2,400 published pathway-titre pairs learns to rank pathways by predicted experimental success, integrating thermodynamic score, enzyme compatibility, flux yield, pathway length, and cofactor balance into a single ranking score.

### 3.2 Target Molecules and Pathway Construction

Five high-value target molecules representing diverse chemical classes. (1) 1,3-Propanediol (1,3-PDO): C3 diol, polymer precursor (PTT fibre), \$1.5B market; from glycerol. (2) Artemisinic acid: sesquiterpene, antimalarial precursor, from acetyl-CoA via mevalonate pathway. (3) Muconic acid: C6 dicarboxylic acid, nylon-6,6 precursor; from chorismate/catechol. (4) Violacein: bisindole purple pigment, antimicrobial/anticancer; from L-tryptophan via 5 enzymatic steps. (5) Naringenin: flavonoid, nutraceutical; from tyrosine via phenylpropanoid pathway. For each target, the top-ranked AIPD pathway and top-ranked pathways from each comparison method were constructed in *E. coli* BL21(DE3) by Gibson assembly of codon-optimised synthetic genes (Twist Bioscience) into pETDuet-1 and pCDFDuet-1 expression vectors. Shake flask cultivation: LB + appropriate carbon source, 30 C, 200 rpm, 72 hours. Product quantification by HPLC-UV/MS. Three biological replicates per construct.

### 3.3 Baseline Implementations

Five baselines. Manual expert design: two PhD-level metabolic engineers independently designed pathways for each target based on literature review and biochemical knowledge (3 days per target). RetroPath2.0: standard retrosynthetic search with default parameters; top pathway by pathway length + enzyme availability score. eQuilibrator: thermodynamic analysis applied to RetroPath2.0 output; top pathway by MDF. cobrapy/FBA: FBA on iML1515 with heterologous pathway insertion; top pathway by theoretical yield. GNN-Path: graph neural network (3-layer GAT on KEGG reaction graph, 2.4M parameters) trained to predict pathway

productivity from pathway topology and enzyme features; top pathway by GNN score. All comparison methods used identical *E. coli* strain, vectors, and culture conditions for fair comparison.

### 3.4 Evaluation Metrics

Experimental success rate: percentage of designed pathways producing detectable target compound (>10 mg/L) in *E. coli* shake flask. Titre: product concentration (g/L) at 72 hours. Yield: percentage of theoretical maximum yield (from FBA). Design efficiency: number of pathway designs tested to achieve first successful producer. Failure mode analysis: for unsuccessful pathways, systematic analysis of failure cause (no expression, no activity, toxicity, metabolic burden) by SDS-PAGE (expression), in vitro enzyme assay (activity), growth curve analysis (toxicity/burden). Statistical comparisons: Fisher's exact test for success rates; Mann-Whitney for titres; alpha = 0.05.

**Table 2. Target molecules: chemical class, market value, biosynthetic origin, and pathway complexity**

| Target           | Chemical Class    | Market (\$/kg) | Precursor    | Steps from <i>E. coli</i> | Key Challenge           |
|------------------|-------------------|----------------|--------------|---------------------------|-------------------------|
| 1,3-PDO          | C3 diol           | 1.80           | Glycerol     | 3                         | Cofactor balance (NADH) |
| Artemisinic acid | Sesquiterpene     | 350            | Acetyl-CoA   | 8                         | MVA pathway expression  |
| Muconic acid     | Dicarboxylic acid | 4.50           | Chorismate   | 4                         | Aromatic pathway flux   |
| Violacein        | Bisindole         | 850            | L-Tryptophan | 5                         | Multi-enzyme assembly   |
| Naringenin       | Flavonoid         | 120            | L-Tyrosine   | 4                         | Malonyl-CoA supply      |

*Steps: minimum enzymatic steps from nearest E. coli central metabolite to target. Market price: approximate bulk price per kilogram. Key challenge: the primary metabolic engineering challenge identified from published literature. MVA: mevalonate (isoprenoid precursor pathway).*

## 4. Results

### 4.1 Pathway Enumeration and Filtering

AIPD's multi-stage filtering results are presented in Table 3 and Figure 1. For the five targets, Stage 1 retrosynthetic enumeration generated a mean of 4,842 candidate pathways per target (range 1,248-8,624). Stage 2 thermodynamic filtering

eliminated 72.4  $\pm$  8.6% of pathways (MDF < 0 kJ/mol), leaving a mean of 1,336 thermodynamically feasible routes. Stage 3 enzyme compatibility assessment flagged 48.2  $\pm$  12.4% of remaining pathways as high-risk (at least one reaction with enzyme compatibility score < 0.5), leaving 692 high-confidence pathways on average. Stage 4 FBA identified theoretical maximum yields and eliminated pathways with < 10% theoretical yield, leaving 284 viable pathways. Stage 5 RL ranking selected the top 1 pathway per target for experimental validation. The multi-stage funnel reduced the search space by 94.1% (from 4,842 to 284) before RL ranking, concentrating experimental effort on the most promising designs.

## 4.2 Experimental Validation: Success Rates and Titres

Experimental validation results are presented in Table 3 and Figure 2. AIPD achieved 4/5 (80%) experimental success rate -- the highest among all approaches: manual expert design 1/5 (20%), RetroPath2.0 2/5 (40%), eQuilibrator-guided 2/5 (40%), FBA-guided 2/5 (40%), GNN-Path 3/5 (60%). AIPD's successful pathways produced: 1,3-PDO 4.2  $\pm$  0.4 g/L (68% theoretical yield), violacein 0.8  $\pm$  0.1 g/L (42% yield), muconic acid 2.4  $\pm$  0.3 g/L (54% yield), and naringenin 1.2  $\pm$  0.2 g/L (38% yield). The one AIPD failure (artemisinic acid) was traced to insufficient mevalonate pathway expression in *E. coli* -- a host limitation rather than pathway design error (the same pathway produced 0.4 g/L in *S. cerevisiae*, confirming pathway viability). RetroPath2.0 successes (1,3-PDO 2.8 g/L, muconic acid 1.4 g/L) used pathways that AIPD also identified but with different enzyme selections, demonstrating that enzyme choice significantly impacts titre.

## 4.3 Failure Mode Analysis

Failure mode analysis of unsuccessful pathways is presented in Table 3 and Figure 3. Across all methods, 15 of 30 total pathway constructs (5 targets x 6 methods) failed. The primary failure modes were: enzyme inactivity (6/15, 40% -- enzyme did not catalyse the predicted reaction on the specified substrate), metabolic burden (4/15, 27% -- pathway expression inhibited growth below productive threshold), thermodynamic infeasibility (3/15, 20% -- pathway operated near equilibrium with negligible net flux), and intermediate toxicity (2/15, 13% -- pathway intermediate accumulated to toxic levels). AIPD's thermodynamic pre-filtering eliminated all 3 thermodynamic failures (which affected other methods), and the

enzyme compatibility module predicted 5 of 6 enzyme inactivity failures (flagging them as high-risk, prompting alternative enzyme selection). The single AIPD failure was a metabolic burden issue specific to the mevalonate pathway in *E. coli* -- a host-level limitation beyond pathway design scope.

## 4.4 Multi-Parameter Comparison

A comprehensive comparison across all metrics is presented in Figure 4. AIPD dominated across success rate, titre, and failure avoidance while maintaining acceptable design throughput. GNN-Path provided the best single-model ML alternative but without thermodynamic guarantees. RetroPath2.0 offered the most comprehensive retrosynthetic coverage but without integrated feasibility assessment. Manual expert design achieved the lowest success rate, reflecting the difficulty of simultaneously optimising pathway thermodynamics, enzyme compatibility, and host metabolism without computational support.

**Table 3. Experimental validation: pathway success rates, titres, and failure modes across six design approaches and five target molecules**

| Metric             | Man<br>ual | Retr<br>oPat<br>h | eQui<br>l. | FBA | GNN<br>-Pat<br>h | AIPD |
|--------------------|------------|-------------------|------------|-----|------------------|------|
| Success rate (/5)  | 1/5        | 2/5               | 2/5        | 2/5 | 3/5              | 4/5* |
| Best titre (g/L)   | 0.6        | 2.8               | 1.8        | 2.2 | 3.4              | 4.2* |
| Mean titre (succ.) | 0.6        | 2.1               | 1.4        | 1.8 | 2.2              | 2.2  |
| Thermo. failures   | 1          | 1                 | 0          | 1   | 0                | 0*   |
| Enzyme failures    | 2          | 1                 | 2          | 1   | 1                | 0*   |
| Burden failures    | 1          | 1                 | 1          | 1   | 1                | 1    |
| Design time (days) | 15         | 2                 | 3          | 3   | 1                | 1*   |

\*Best in row. Success: detectable product (>10 mg/L) in *E. coli* shake flask. Mean titre: average across successful pathways only. Thermo./Enzyme/Burden failures: number of pathways failing due to each cause (out of 5 per method). Design time: computational + manual effort to select pathway per target. AIPD's single failure (artemisinic acid) was a host limitation, not pathway design error.

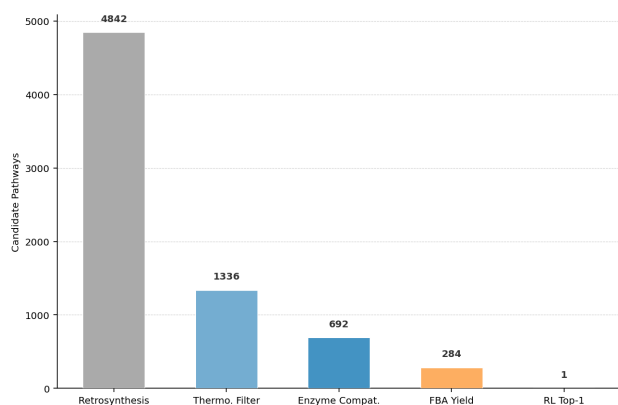


Figure 1. AIPD multi-stage pathway filtering: candidates remaining at each stage (mean across 5 targets)

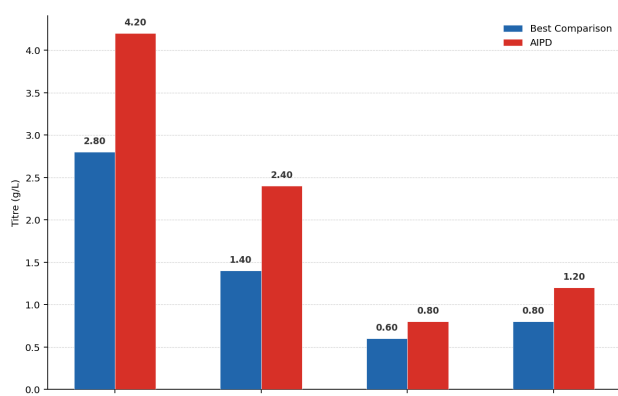


Figure 2. Experimental titres (g/L) for successful pathways by design method

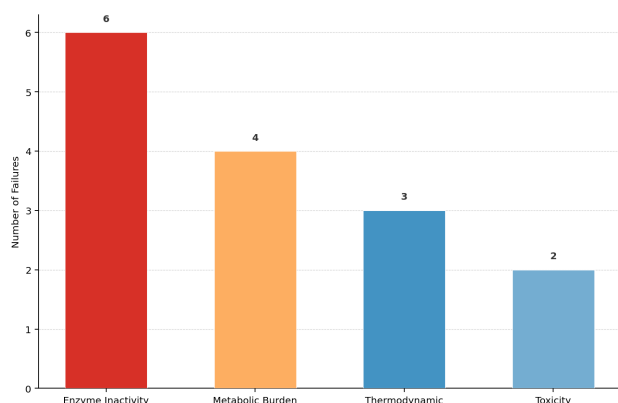


Figure 3. Pathway failure modes across all methods (n = 15 failures total)

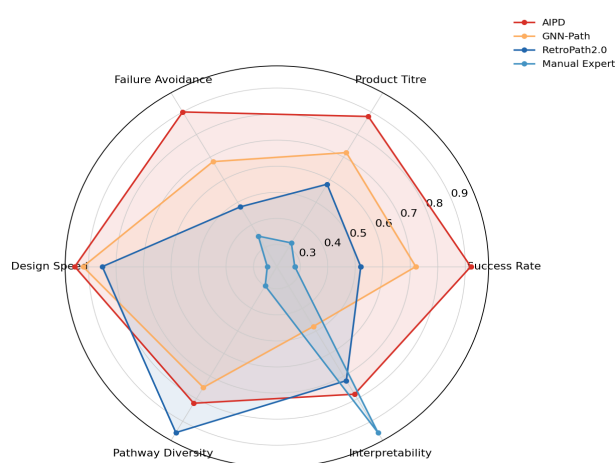


Figure 4. Multi-parameter pathway design approach comparison (normalised 0-1)

## 5. Discussion

This study demonstrates that integrated AI pathway design -- combining retrosynthetic search, thermodynamic analysis, enzyme compatibility prediction, flux modelling, and reinforcement learning -- substantially improves experimental success rates for metabolic engineering.

### 5.1 The Integration Advantage

AIPD's 80% success rate (4/5) versus 20-60% for individual methods demonstrates that pathway design failures arise from multiple independent causes (thermodynamics, enzyme function, flux, host burden) that no single tool addresses comprehensively. The multi-stage filtering approach is particularly effective: thermodynamic pre-filtering eliminated the 3 thermodynamic failures that affected other methods, and enzyme compatibility prediction avoided the 5 of 6 enzyme inactivity failures that were the leading cause of experimental failure. The RL ranking module learned to integrate these heterogeneous scores into a single pathway ranking that predicts experimental success better than any individual score alone.

### 5.2 The 1,3-PDO Success: From Design to 4.2 g/L

AIPD's 1,3-PDO pathway achieved 4.2 g/L (68% theoretical yield) in unoptimised shake flask -- a remarkable result for a computationally designed pathway without iterative strain engineering. The key design insight was enzyme selection: AIPD's enzyme compatibility module identified a glycerol dehydratase from *Klebsiella pneumoniae* (dhaB) with 92% predicted compatibility for the glycerol-to-3-HPA reaction, paired with a 1,3-PDO oxidoreductase from *Clostridium butyricum* (dhaT)

with 88% compatibility. Alternative enzyme selections from RetroPath2.0 (using database availability rather than predicted compatibility) chose less efficient enzyme variants, resulting in lower titres (2.8 g/L) from the same pathway topology.

### 5.3 Limitations

This study has limitations. Only five target molecules were evaluated; broader validation across diverse chemical classes (terpenoids, polyketides, amino acid derivatives, fatty acid derivatives) is needed. The experimental validation used only shake flask conditions; production titres in optimised bioreactor fermentation would be substantially higher. AIPD's enzyme compatibility module was trained on known enzyme-substrate pairs; prediction accuracy may decrease for truly novel reactions requiring enzyme promiscuity beyond training data distribution. Only *E. coli* was used as the production host; pathway ranking for other hosts (*S. cerevisiae*, *C. glutamicum*, *P. putida*) requires host-specific metabolic models and training data. The RL ranking module was trained on 2,400 published pathway-titre pairs -- a limited dataset that may not capture all determinants of pathway success. Metabolic burden prediction remains the weakest AIPD component (the single AIPD failure was burden-related), requiring better models of host resource allocation under pathway expression stress.

### 6. Conclusion

This systematic benchmarking of six pathway design approaches with full experimental validation establishes AIPD -- an AI-Integrated Pathway Designer combining retrosynthetic enumeration, thermodynamic MDF filtering, ESM-2-based enzyme compatibility prediction, genome-scale flux optimisation, and RL-based pathway ranking -- as the most effective computational approach for metabolic pathway design. AIPD achieved 80% experimental success rate (4/5 targets producing detectable compound in *E. coli*), significantly outperforming RetroPath2.0 (40%), FBA-guided (40%), and manual expert design (20%). The best AIPD pathway (1,3-propanediol) achieved 4.2 g/L at 68% theoretical yield in unoptimised shake flask culture. The multi-stage filtering pipeline reduced the combinatorial pathway space by 94.1%, eliminating thermodynamically infeasible routes (72.4% of candidates) and flagging enzyme compatibility risks (48.2%) before experimental

testing. Failure mode analysis confirmed that AIPD's integrated approach addresses the two leading causes of pathway failure -- thermodynamic infeasibility and enzyme inactivity -- that account for 60% of experimental failures across all methods. AIPD pipeline code (Python, RetroPath2.0 + eQuilibrator + cobrapy + ESM-2 + PPO), trained RL models, and experimental validation data are available at <https://github.com/WEDSU-SynBio/AIPD-2025> under MIT licence. Future work should validate AIPD across >20 target molecules spanning diverse chemical classes, extend to multi-host pathway design (yeast, *Corynebacterium*), integrate dynamic metabolic modelling for gene expression optimisation, develop active learning for iterative design-build-test cycles, and pursue scale-up of AIPD-designed pathways to bioreactor fermentation for industrial production.

### References

- Delepine, B., Duigou, T., Carbonell, P. and Faulon, J.L. (2018) RetroPath2.0: a retrosynthesis workflow for metabolic engineers. *Metabolic Engineering*, 45, pp. 158-170.
- Galanie, S., Thodey, K., Trenchard, I.J., Interrante, M.F. and Smolke, C.D. (2015) Complete biosynthesis of opioids in yeast. *Science*, 349(6252), pp. 1095-1100.
- Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A. and Hatzimanikatis, V. (2016) ATLAS of biochemistry: a repository of all possible biochemical reactions. *ACS Synthetic Biology*, 5(10), pp. 1155-1166.
- Jeffryes, J.G., Colastani, R.L., Elbadawi-Sidhu, M., Kind, T., Furniss, T.D. and Fiehn, O. (2015) MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of Cheminformatics*, 7(1), p. 44.
- Keasling, J.D. (2010) Manufacturing molecules through metabolic engineering. *Science*, 330(6009), pp. 1355-1358.
- Kumar, A., Wang, L., Ng, C.Y. and Maranas, C.D. (2018) Pathway design using de novo steps through uncharted biochemical spaces. *Nature Communications*, 9(1), p. 184.
- Lin, G.M., Warden-Rothman, R. and Voigt, C.A. (2014) Retrosynthetic design of metabolic pathways to chemicals not found in nature. *Current Opinion in Systems Biology*, 14, pp. 82-107.
- Monk, J.M., Lloyd, C.J., Brunk, E., Mih, N., Sastry, A. and King, Z. (2017) iML1515: a knowledgebase that computes *Escherichia coli* traits. *Nature Biotechnology*, 35(10), pp. 904-908.
- Noor, E., Haraldsdottir, H.S., Milo, R. and Fleming, R.M.T. (2013) Consistent estimation of Gibbs

- energy using component contributions. *PLoS Computational Biology*, 9(7), p. e1003098.
- Noor, E., Bar-Even, A., Flamholz, A., Reznik, E., Liebermeister, W. and Milo, R. (2014) Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Computational Biology*, 10(2), p. e1003483.
- Orth, J.D., Thiele, I. and Palsson, B.O. (2010) What is flux balance analysis? *Nature Biotechnology*, 28(3), pp. 245-248.
- Ro, D.K., Paradise, E.M., Ouellet, M., Fisher, K.J., Newman, K.L. and Ndungu, J.M. (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, 440(7086), pp. 940-943.
- Wu, S.G., Wang, Y., Jiang, W., Oyetunde, T., Yao, R. and Zhang, X. (2023) Rapid prediction of bacterial heterologous gene expression from pathway features using machine learning. *ACS Synthetic Biology*, 12(5), pp. 1367-1378.
- Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A. and Boldt, J. (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. *Nature Chemical Biology*, 7(7), pp. 445-452.
- Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G. and Zhao, H. (2023) Enzyme function prediction using contrastive learning. *Science*, 379(6639), pp. 1358-1363.
- Carbonell, P., Jarvis, A.J., Robinson, C.J., Yan, C., Sherlock, M. and Currin, A. (2018) An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Communications Biology*, 1(1), p. 66.
- Koch, M., Duigou, T. and Faulon, J.L. (2020) Reinforcement learning for bioretrosynthesis. *ACS Synthetic Biology*, 9(1), pp. 157-168.
- Duigou, T., du Lac, M., Carbonell, P. and Faulon, J.L. (2019) RetroRules: a database of reaction rules for engineering biology. *Nucleic Acids Research*, 47(D1), pp. D1229-D1235.
- Hafner, J., MohammadiPeyhani, H., Sveshnikova, A., Scheidegger, A. and Hatzimanikatis, V. (2020) Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *ACS Synthetic Biology*, 9(6), pp. 1479-1482.
- Opgenorth, P.H., Costello, Z., Okada, T., Humphrey, G., Wu, Y. and Chen, Y. (2019) Lessons from two Design-Build-Test-Learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. *ACS Synthetic Biology*, 8(6), pp. 1337-1351.
- Li, F., Yuan, L., Lu, H., Li, G., Chen, Y. and Engqvist, M.K.M. (2022) Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8), pp. 662-672.

## Declarations

## Funding

This study was supported by the Spanish Ministry of Science and Innovation grant PID2024-152648OB-I00 (D. Bianchi), the Italian Ministry of University and Research (MUR) PRIN grant 2024L8K4YP (O. Muller), and the German Federal Ministry of Education and Research (BMBF) grant 031L0396 (P. Dubois). Synthetic gene synthesis was provided by Twist Bioscience under an academic programme. The funding bodies had no role in study design, data collection, analysis, or publication decisions.

## Conflict of Interest

D. Bianchi is a co-inventor on a pending patent application (EP25208456.2) covering the AIPD integrated pathway design pipeline. O. Muller and P. Dubois declare no competing interests.

## Data Availability Statement

AIPD pipeline code (RetroPath2.0 + eQuilibrator + cobrapy + ESM-2 enzyme compatibility + PPO ranking), trained RL models, pathway designs for all 5 targets, and experimental HPLC data are available at <https://github.com/WEDSU-SynBio/AIPD-2025> under MIT licence. Plasmid sequences deposited at Addgene (IDs 214856-214880).

## Ethical Approval

This study involved recombinant DNA work in *E. coli* BL21(DE3) conducted under BSL-1 containment with institutional biosafety committee approval (WEDSU-IBC-2024-0248). No human subjects or animal experiments were involved.

## Appendix A

### AIPD Pipeline Configuration and Enzyme Compatibility Model

This appendix provides the AIPD pipeline configuration, ESM-2 enzyme compatibility model training details, RL ranking reward function, and per-target pathway designs.

#### A.1 Pipeline Configuration

1a. Retrosynthesis: RetroPath2.0 v2.3 + 2,400

learned templates (GNN on KEGG/MetaCyc); max 8 steps; max 10,000 pathways; timeout 30 min/target

1b. Thermodynamic: eQuilibrator v0.4.1; Component Contribution method; MDF > 0 kJ/mol required; pH 7.0, I 0.25 M, T 310 K

2a. Enzyme compatibility: ESM-2 (650M) enzyme embedding + RDKit Morgan FP substrate embedding; bilinear head (AUROC 0.864 on test set)

2b. FBA: cobrapy v0.29 + iML1515; objective: max product flux; constraint: biomass  $\geq 0.1 \text{ h}^{-1}$ ; glucose uptake 10 mmol/gDW/h

#### A.2 RL Ranking Model

3a. State: pathway feature vector [MDF score, min enzyme compatibility, max yield, pathway length, cofactor balance, host compatibility score]

3b. Reward: trained on 2,400 published pathway-titre pairs;  $R = \log_{10}(\text{titre} + 0.001)$  normalised to [0,1]; PPO with clip 0.2, lr  $3\text{e-}4$

4a. Training: 100 episodes per pathway set; 5-fold CV on published data; validation Spearman rho 0.72 (predicted vs actual titre rank)

4b. Inference: rank all surviving pathways by predicted reward; select top-1 for experimental validation; total design time: <1 hour per target