PROTO-NOOS:

Orchestrating Open-Access Bioinformatics for Seamless Drug Discovery

Konrad Gorzelańczyk[1]○, Szymon Krawczyk[1], Maksymilian Korbik[1], Florian Hołubowski[1],

[1]Faculty of Computing and Telecommunications, Poznan University of Technology, Piotrowo 3a, 60-965 Poznań, Poland

○ Correspondence: sknwpl@proton.mail

01 MOTIVATION

From single-score docking to multi-signal evidence

Computational antibiotic discovery often ranks compounds using isolated signals such as docking score, predicted affinity, or drug-likeness rules. However, **favourable affinity does not guarantee success**: candidates can fail due to poor **Gram-negative entry**, **efflux**, unstable binding, irrelevant metabolism, or lack of a feasible synthetic route.

PROTO-NOOS integrates open tools, including REINVENT4, RDKit, an ODE-based Cell Target Engagement model, Boltz-2, GROMACS, BioTransformer3, COBRApy/iML1515, and AiZynthFinder, into a **reproducible workflow** for the E. coli DHFR context, PDB 6XG5, target gene folA. The pipeline **does not claim antibacterial activity**, but generates **structured evidence** for triaging large generated populations into a smaller set of candidates for deeper analysis.

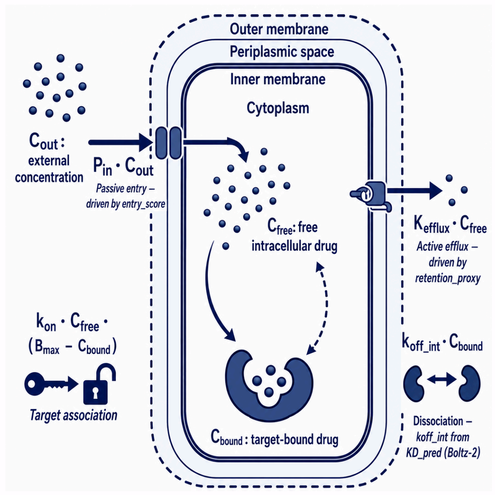
Can independent computational signals — chemistry, structure, dynamics, metabolism, synthesis — be combined into one defensible ranking?

24,089	**3,234**	**1,538,032**	**557 / 424**
DE NOVO COMPOUNDS	STOKES PHENOTYPIC SET	REINVENT SEARCHED	CHEMBL / BINDINGDB

03 STAGE 03 — CELL TARGET ENGAGEMENT **ODE module separates structural affinity from cellular accessibility**

A compound with favourable predicted *KD* may still fail intracellularly if it does not cross the **outer membrane** or is actively **effluxed**. **PROTO-NOOS inserts a mechanistic ODE** between cheap chemistry and expensive structure prediction. The module integrates **free intracellular drug** (*Cfree*) and **target-bound drug** (*Cbound*) over time, driven by passive entry, efflux, target association, and dissociation. It runs twice: **before Boltz-2** (kinetic priors only) and **after Boltz-2** (KD_pred replaces the prior).

VISUALIZATION 2 — FIG 2



Governing equations

$$\frac{dC_{free}}{dt} = P_{in}\,C_{out} - k_{efflux}\,C_{free} - k_{on}\,C_{free}\,(B_{max} - C_{bound}) + k_{off,int}\,C_{bound}$$

$$\frac{dC_{bound}}{dt} = k_{on}\,C_{free}\,(B_{max} - C_{bound}) - k_{off,int}\,C_{bound}$$

Pre-Boltz-2 pass

KD_pred unavailable - kinetic priors used. Ranking by entry + efflux only.

Post-Boltz-2 rescore

KD_pred integrated. Full dynamics: affinity + accumulation + rebinding.

Fig. 2. CellTE models entry, efflux, intracellular free drug and target-bound drug. It is the mechanistic bridge between Boltz2 KD and cellular target occupancy.

01 TABLE STAGE-LEVEL CHARACTERISATION

Table 1: Per-stage characteristics on the **de novo branch** (**n = 15,282 unique compounds, 15,900 rows**). Mean runtime per input molecule and share of total pipeline runtime are computed from *runtime_by_stage_branch.csv*. Availability is the fraction of compounds for which the stage produced a **non-null primary metric** (*stage_metric_availability_by_branch.csv*). **Stage 05 alone consumes 47%** of the pipeline budget and **Stage 04 a further 32%**: any meaningful cost reduction must come from **earlier-stage gating**.

01 **TABLE**

Stage	Partition	Primary tool	Output metric	Mean runtime / cmpd[s]	Share of total time [%]	Availability(de novo)	Note
01 De novo generation	GPU(ngpu)	REINVENT4	stage1_score (QED)	0.55	0.7	100.0%	Receptor-focused prior, deduplicated SMILES.
02a Chemistry & filters	CPU(obl)	RDkit	entry_score	1.84	2.2	100.0%	Lipinski, Veber, PAINS, Gram-negative heuristic.
02b Retention proxy	CPU(obl)	Surrogate model	retention	0.06	0.1	100.0%	Canonical artefact resolution at Stage 02b.
03 CellTE pre-Boltz-2	CPU(obl)	ODE module	score_cellTE (pre)	0.06	0.1	100.0%	Kinetic priors for kon and koff.
03 CellTE rescore	CPU(obl)	ODE module	score_cellTE (post)	0.05	0.1	95.5%	KD,pred from Boltz-2 replaces the prior.
04 Structure & affinity	**GPU(hgx A100)**	**Boltz-2**	**KD_pred, ipTM**	**26.82**	**32.3**	**95.5%**	**GPU bottleneck #2. Resume + dedup enforced.**
05 MD stability	**GPU(hgx A100)**	**GROMACS**	**stability, RMSD**	**39.05**	**47.8**	**91.0%**	**Dominant cost. Fast-MD mode flagged as low confidence.**
06a Xenobiotic fate & FBA	CPU(obl)	BioTransformer3 + COBRApy / iML1515	growth_ratio	13.64	16.3	25.8%	MNX mapping is sparse for OOD de novo compounds.
06b Target impact (folA / DHFR)	CPU(obl)	COBRApy synleth / varB	inhibited_growth_ratio	1.09	1.3	91.0%	synleth and varB give identical IGR (ρ = 1.00).
07 Multi-criteria ranking	CPU(obl)	GUB scorer + desirability	GUB_score	< 0.5	< 0.5	100.0%	Composite of structural, MD, systems, retrosynthesis blocks.
08 Retrosynthesis	CPU(obl)	AiZynthFinder (USPTO)	is_solved, route_score	var.	var.	top-N only	Depth ≤ 5, 300 s per compound, Enamine stock.

04 INTER-STAGE COUPLING **Where the pipeline is redundant, where it is informative**

A **cross-stage correlation audit** revealed three deterministic links and one near-duplicate. *KD_pred* from Stage 04 and inhibited growth ratio from Stage 06b are perfectly coupled, **Spearman ρ = 1.00**, so **Stage 06b adds no independent ranking signal** once *KD_pred* is included. Stage 06b *synleth* and *varB* modes are also redundant, **ρ = 1.00**, while Stage 05 stability and RMSD show the expected inverse coupling, **ρ≈−0.94**. In contrast, Stage 03 pre-Boltz-2 CellTE and rescored CellTE correlate only moderately, **ρ=0.76**, indicating that the **Boltz-2** *KD* **pass adds new information** to cellular ranking.

VISUALIZATION 4 — FIG 4

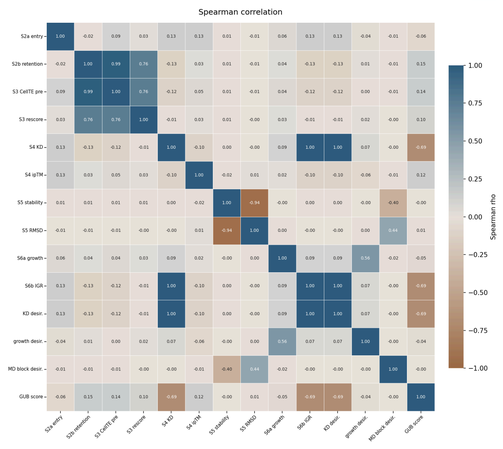


Fig. 4. Inter-stage Spearman correlation matrix across 14 pipeline metrics. Three deterministic KD-driven links (Stage 04 KD,pred ↔ Stage 06b IGR ↔ desirability d_KD_pred, all **ρ=1.00**) and one near-redundant CellTE link (**pre vs rescore ρ=0.76**) are highlighted.

VISUALIZATION 5 — FIG 5

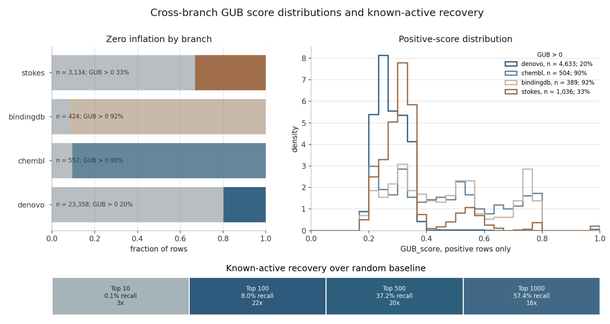


Fig. 5. De novo rows are strongly zero-inflated, whereas **ChEMBL** and **BindingDB are enriched for positive GUB scores**. Top-1000 recovery captures 57.4% of known actives, 16× over random baseline.

02 PIPELINE ARCHITECTURE

The workflow is split into a GPU partition for generation, structure, affinity, and MD stability, and a CPU partition for filtering, retention, Cell Target Engagement, xenobiotic fate, target impact, ranking, and retrosynthesis.

Stages communicate through **CSV / Pydantic typed contracts**, enabling each module to be re-run, audited, or replaced without breaking downstream steps.

VISUALIZATION 1 — FIG 1

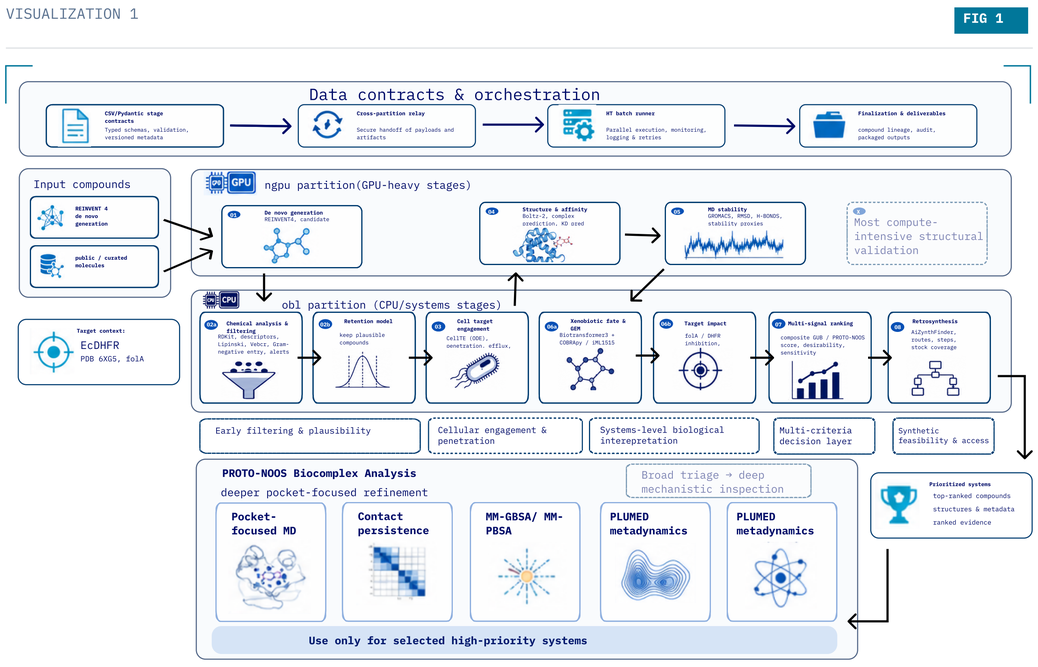


Fig. 1. PROTO-NOOS connects generation, filtering, CellTE, Boltz2 affinity, GROMACS stability, BioTransformer/COBRA target logic, ranking and retrosynthesis into one stage-contract workflow.

03 STAGE 03 — CELL TARGET ENGAGEMENT **Loud failure is preferable to a corrupted score**

Multi-stage pipelines fail silently when stale artefacts, partial outputs, or unit mismatches propagate into later scoring. PROTO-NOOS enforces validated boundary checks at every stage: **molecular yield threshold**, **canonical artefact resolution**, **KD_pred** and **CIF path existence**, **MD confidence flagging**, and **FBA mapping coverage**. Critical breaches halt the pipeline with a non-zero exit code rather than flowing into the GUB score.

VISUALIZATION 3 — FIG 3

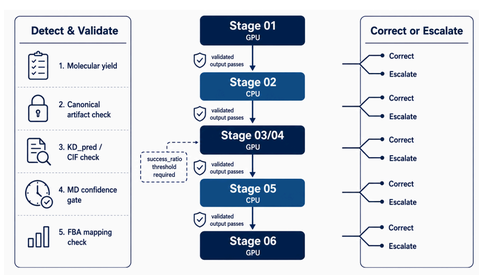


Fig. 3. Each stage uses **boundary checks** with explicit **correct / escalate** paths. Transient failures, such as GPU errors or filesystem timeouts, trigger **retry with exponential backoff**. Stale artefacts are **pruned and deduplicated** by *compound_id*. Silent semantic errors, including **partial CIF files**, **missing** *KD_pred*, or **yields below threshold**, cause a **hard stop** to prevent downstream error propagation.

06 CROSS-BRANCH SCREENING BEHAVIOUR **Branches differ sharply in zero-inflation and positive-score density**

On four independent input streams — de novo (REINVENT4), **Stokes phenotypic actives**, **BindingDB**, and **ChEMBL** — the GUB scorer concentrates non-zero scores in known-active branches (**ChEMBL 90%**, **BindingDB 92%**) and aggressively zeroes out the broad de novo population (**20% non-zero**) and the Stokes set (**33% non-zero**). On **a top-1000 cut**, recall of known actives is **57.4%**, a **16× enrichment over a random baseline**, which validates the composite score as a **triage instrument** rather than a hard activity classifier.

07 BOLTZ RESCORE **Boltz-2 adds affinity context that entry-efflux priors cannot predict**

The **pre-Boltz-2 CellTE** score ranks compounds using entry and efflux priors. The **post-Boltz-2 rescore** replaces the kinetic prior with *KD_pred* and re-integrates the ODE. The rank correlation is moderate, **ρ=0.76** with **median |Δrank| = 12**, showing that **the affinity pass adds new information**. Low-*KD_pred* compounds shift upward, while high-*KD* compounds drop despite strong entry scores. Thus, CellTE prevents compounds with **good penetration but weak affinity** from remaining highly ranked.

VISUALIZATION 6 — FIG 6

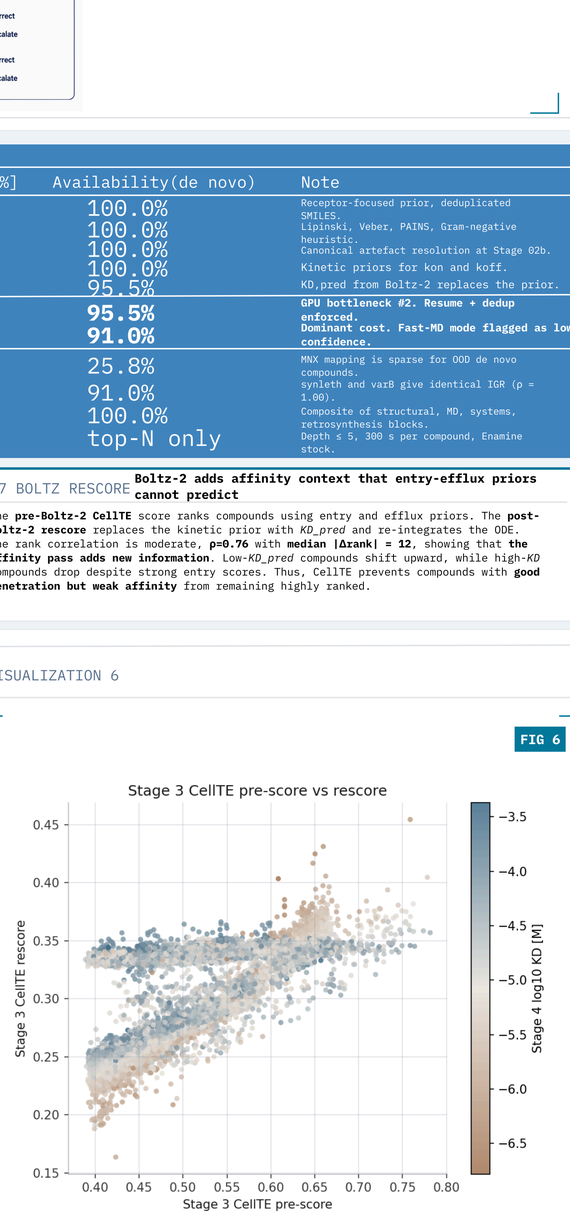


Fig. 6. Pre vs post CellTE ranking has Spearman **ρ=0.76**. **KD_pred induces** a **bimodal split** and **changes 99%** of rank **positions**.

09 CONCLUSIONS **PROTO-NOOS is a reproducible triage system, not a claim of confirmed antibacterial activity, yet.**

01 The pipeline successfully integrates de novo generation, chemistry, retention, CellTE, Boltz-2, GROMACS, BioTransformer3, COBRApy, and AiZynthFinder under one typed contract. The contribution is **not a confirmed antibiotic**, but a **reproducible, auditable triage infrastructure** for the **E. coli DHFR** context.

02 **Boltz-2 contributes non-trivial information** to the cellular ranking. The pre vs post Boltz-2 CellTE **Spearman is 0.76** with **median |Δrank| = 12** and **max = 50** on the top compounds. This is the **right amount of disruption**: large enough that running Boltz-2 changes the answer, small enough that **pre-Boltz-2 gating** on entry and efflux priors is still informative.

03 **Stage 06a** (xenobiotic FBA) reaches only **27% availability** on de novo compounds. The remaining **73%** carry no parent or predicted product in MetaNetX. This is a **database coverage limit**, not a property of the compounds: the very novelty that motivates de novo generation places the molecules outside the training distribution of metabolic annotation. Until this is addressed by **scaffold-based generalisation** or a richer metabolite mapping, **Stage 06a contributes a sparse signal** on novel chemistry.

05 The **inter-stage Spearman audit** reveals which signals carry independent information and which do not. *KD_pred* (S04) and IGR (S06b) are **ρ=1.00**: once *KD_pred* is in scope, **IGR adds nothing**. The *synleth* and *varB* modes of Stage 06b also collapse (**ρ=1.00**), and Stage 05 stability and RMSD are **ρ≈−0.94**. **Stage 06b in its current configuration confirms target essentiality rather than discriminating compounds** and is the **next module to refactor**.

04 **Cost is concentrated in a single stage. Stage 05** (GROMACS MD) alone consumes **47% of the pipeline budget** while operating in low-confidence fast-MD mode and contributing **weight 0.0** to the final GUB score. **Stage 04** (Boltz-2) adds another **32%**. The cheap mechanistic ODE in **Stage 03 is the right place to gate**, which is consistent with the early-exit results reported in the companion ML poster.

06 The **execution integrity layer is a precondition for any of the above to be trustworthy**. Yield thresholds, canonical artefact resolution, KD,pred and CIF coverage checks, MD confidence flagging, and **hard halts on coverage breach** prevent stale or partial artefacts from entering the GUB score. **All ranking conclusions remain computational until supported by synthesis and antibacterial assays!**

ρ(pre, post Boltz-2) = 0.76	**47%**	**ρ(KD, IGR) = 1.00**	**27%**
RESCORE ADDS INFORMATION	RUNTIME IN STAGE 05 (MD)	REDUNDANCY S04 ↔ S06b	FBA MAPPING COVERAGE (DE NOVO)

PREPRINT x REPOSITORY

PREPRINT

https://komputerowe-projektowanie-lekow.github.io/webpage/

CODE

https://komputerowe-projektowanie-lekow.github.io/webpage/proto-noos_system.html