Data−driven pharmacogenomic classification and drug−response target prioritisation

Unsupervised reclassification of drug−ADR genetic associations,
with a latent (drug−triggered) PGx priority score − no expert input

Model parameters

Only VARIANT descriptors and PHENOTYPE descriptors − no gene−specific parameters

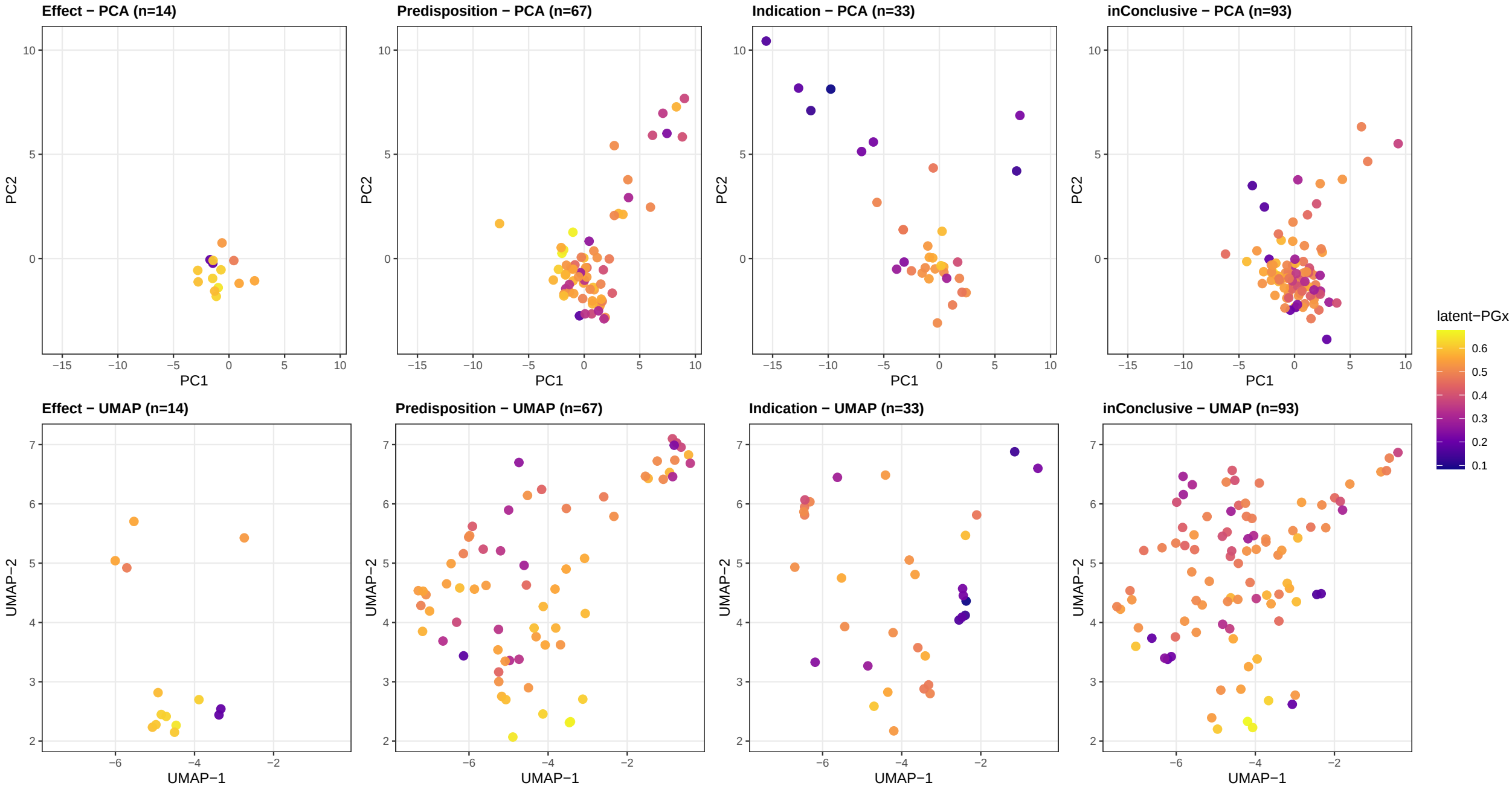
- **GWAS backbone: absolute effect size and −log10 P for four phenotype arms of each variant.**
- **Cross−phenotype contrasts: drug−ADR effect minus disease / dose / prescription effects.**
- **Temporal: counts, median years and onset spread of events BEFORE vs AFTER first prescription.**
- **Held−out test set: genotype counts, case−control frequency difference, effect calibration**

  (test−vs−discovery odds−ratio difference and ratio).
- **Variant function (VEP): consequence severity, biotype, population allele frequency.**
- **Pharmacovigilance: FAERS disproportionality. Internal pleiotropy: variant specificity.**
- No gene−specific parameters (no tissue, druggability, gene−disease scores) and no arbitrary

  thresholds − continuous values and canonical ordinals only.

Each expert class, coloured by latent−PGx score

Only the named expert class is shown in each panel (top: PCA, bottom: UMAP); shared latent−PGx colour scale



What the parameters measure

Four GWAS views of one variant resolve drug−specificity

- drug−ADR arm: does the variant associate with the adverse event under the drug?
- disease−alone arm: does it associate with the disease in the general population (no drug)?
- dose arm: does it associate with the dose patients are titrated to?
- prescription arm: does it associate with who is prescribed the drug?
- **Drug−specificity = drug−ADR effect dominating ALL three alternatives −> a drug−conditional signal.**
- **Low internal pleiotropy + concordant direction + replication = a specific, real drug effect.**

Classification − unsupervised methods

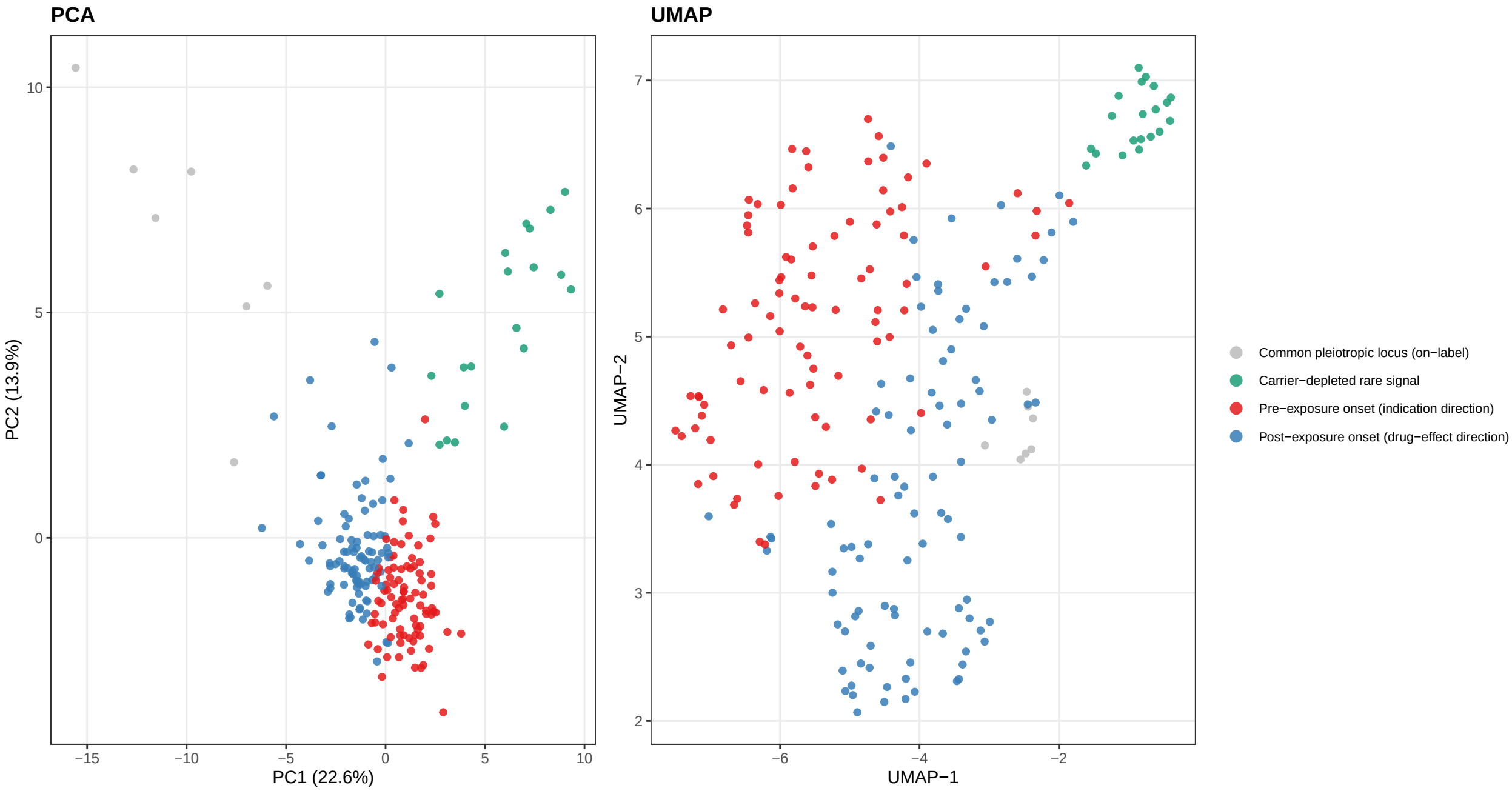
Four complementary algorithms; consensus over 100 seeds; cluster number chosen by the data

- **k−means: centroid / Euclidean partitioning; spherical clusters; k chosen by silhouette.**
- **GMM: Gaussian mixture (tied covariance); soft posterior membership; k chosen by BIC.**
- **LCA: latent class analysis; handles mixed−type features; k chosen by BIC.**
- **HDBSCAN: density−based; discovers k; isolates outliers as a separate group.**
- 100−seed consensus gives a soft membership and a per−association stability score,

  so borderline associations are reported as borderline rather than forced into one class.

Data−driven classification of drug−response associations

Each point = one association; colour = unsupervised class



The discovered classes (by parameter signature)

Groups defined by GWAS parameters alone (self−contained model)

- **On−label indication: high prescription−arm signal; disease precedes the drug (3 sub−groups).**
- **Rare−variant predisposition: large effect size and SE, low allele frequency, drug−amplified.**
- **Drug−amplified mechanism (effect−enriched): drug−ADR dominates other arms, replicated −**

  the group holding the candidate drug−triggered effects.
- **Low/moderate−evidence mass + weak−evidence residual: the large, low−signal background.**
- Self−contained features classify more coarsely than annotation−rich models,

  but the latent−PGx prioritisation (next) is independent of the clustering.

The pharmacogenomically interesting class

Latent predisposition realised only by the drug (gene x drug interaction)

- **Parameter signature: drug−ADR effect AND −log10 P dominate disease, dose and prescription arms,**

  concordant direction, replicated in validation, and low internal pleiotropy.

- Silent in the general population (low disease−alone signal); unmasked by the drug.

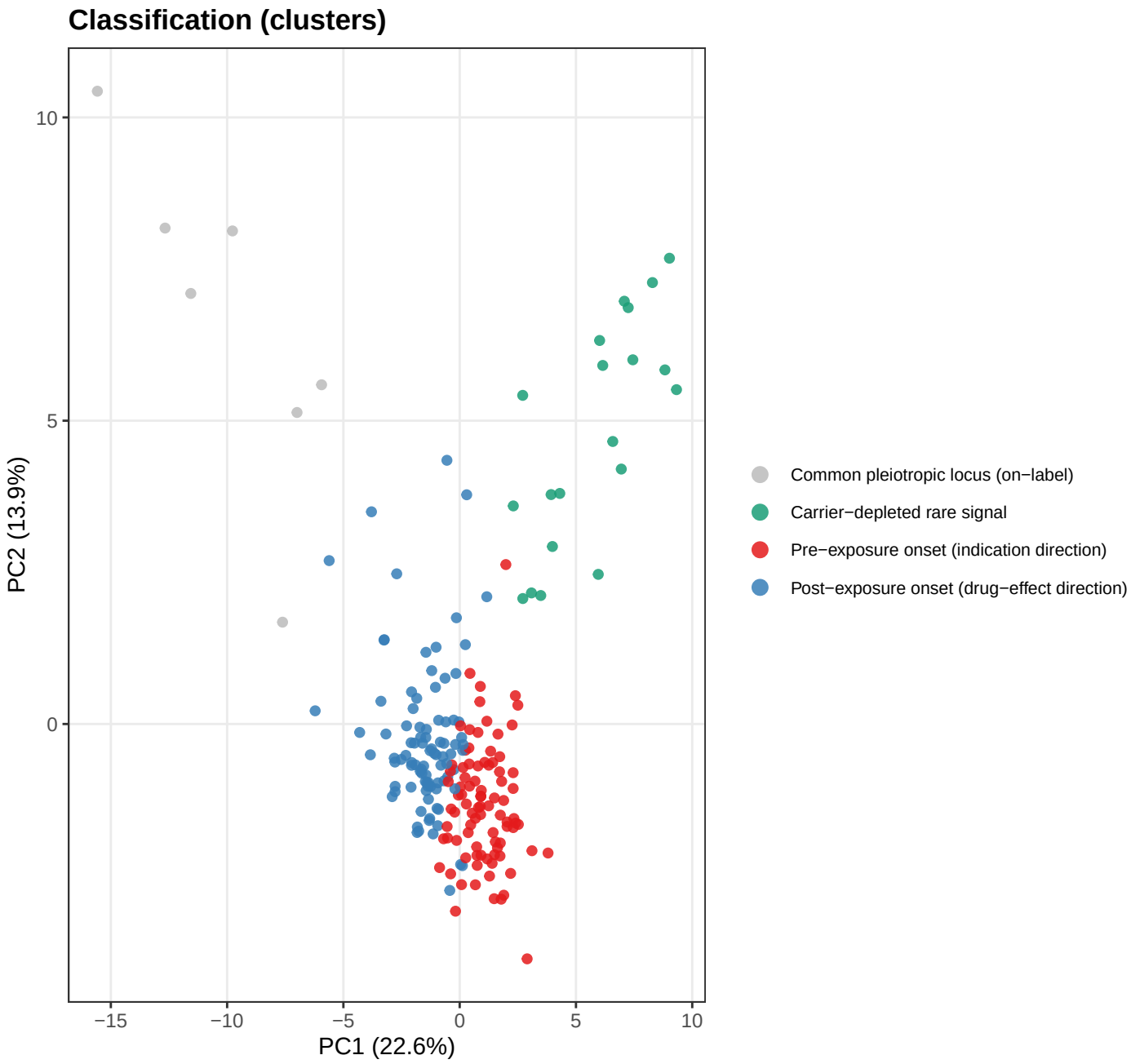
- Distinct from disease−driven variants (which associate with the disease without the drug).

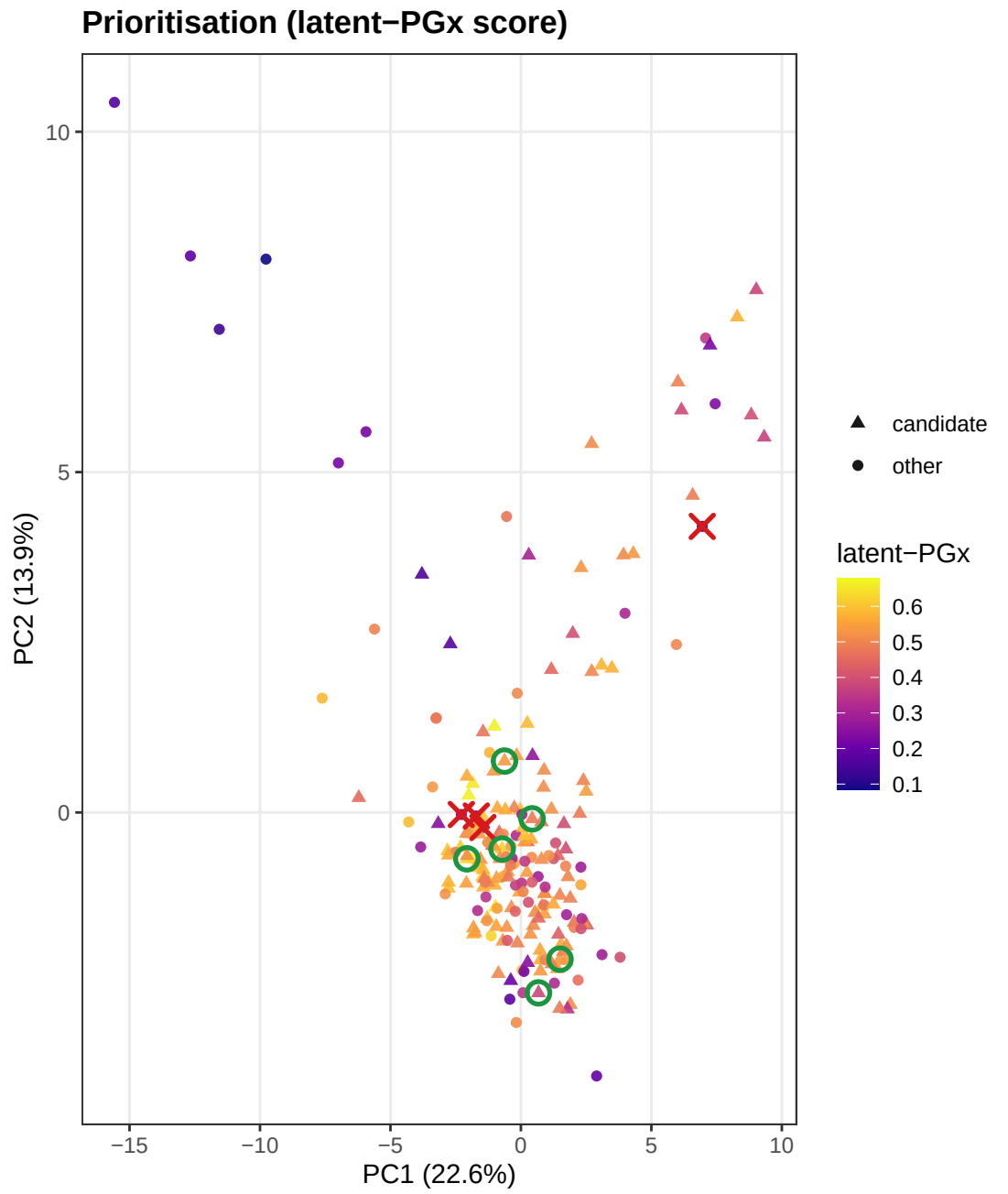
- Captured by a continuous latent−PGx score, layered on top − not a clustering input.

- **The most interesting and the rarest class.**

Same PCA space: class membership vs latent−PGx priority

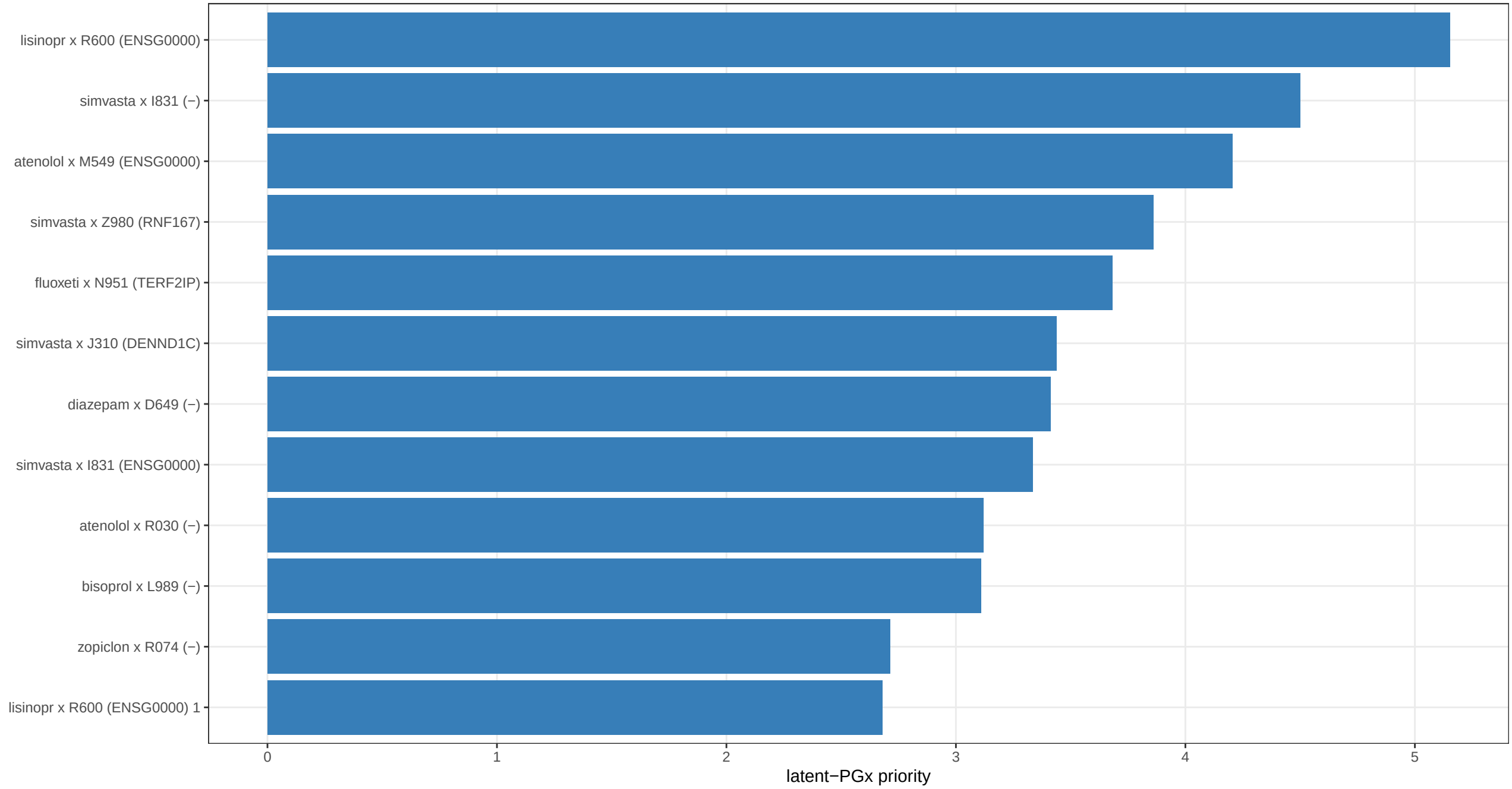
Left: unsupervised class. Right: latent−PGx score (triangles = gated candidates)





Top latent−PGx priority candidates

Ranked drug−gene−ADR triples for fine−mapping and replication



Example 1 − a latent−PGx target

lisinopril x angioedema (R60.0)

- ACE−inhibitor angioedema is a known drug−triggered reaction (bradykinin pathway).
- **Latent−PGx score high (+2.7): drug−ADR effect dominates disease, dose and prescription arms.**
- Internal pleiotropy = 3 (variant−specific); concordant direction; replicated.
- Lead variant is a tag SNP −> flagged FINE−MAPPING NEEDED.
- Archetype: a real gene x drug pair where the causal variant must be resolved.

Example 2 − correctly excluded

Factor V Leiden (F5) x thrombophlebitis (I80.2)

- F5 is a strong genetic risk for thrombosis − but it causes disease WITHOUT any drug.
- **Latent−PGx score deeply negative (−7.0): flagged MANIFEST RISK.**
- High disease−alone signal; drug−specificity near zero; internal pleiotropy elevated.
- Disease−driven / genetic comorbidity − not pharmacogenomics.
- Shows the score separates drug−triggered effects from plain genetics.

Example 3 − deprioritised

ramipril x acute myocardial infarction (I21.9)

- Ramipril is prescribed after MI − confounding by indication, not an adverse reaction.
- **Latent−PGx score negative: the drug−ADR signal does not exceed the prescription arm.**
- High prescription−arm association (the variant tracks who receives the drug).
- Correctly placed in the indication class and excluded from latent candidates.
- Demonstrates indication confounding is separated by the prescription−arm parameter.