

Cross-compartment triangulation of nasopharyngeal mucosal inflammation nominates two chemically unexploited epithelial serine proteases and delivers de novo inhibitor matter

A fully computational target-discovery and molecule-construction study in allergic rhinitis and adenoid disease, with an internally validated structure-based design protocol — expanded round: 93 added samples, a seven-cohort signature, HPA/GWAS evidence layers and a five-pocket de novo program

Agent Team¹ · scalingX¹

¹RsiingSI — the open-source org scaling recursive self-improvement for bioscience.

Authorship: this manuscript, its analyses, figures and molecules were produced end-to-end by an autonomous agent pipeline under the direction of scalingX, with every number read from archived result files and re-checked by a separate verification script. It is reported as an agent-generated discovery report; no wet-lab data were generated. All code and derived tables are archived with the manuscript.

in-silico study, no wet-lab data generated · corresponding data: GSE19187, GSE329964, GSE43523, GSE261239, GSE267331 · all code and derived tables archived (scripts/34–55, results/, figures/)

Abstract. Nasopharyngeal mucosal inflammation — the shared substrate of allergic rhinitis (AR) and adenoid hypertrophy (AH) — is treated today with antihistamines, intranasal corticosteroids and, in refractory type-2 disease, IL-4Ra blockade. All of these act on immune signalling, and none addresses the epithelial barrier that sustains the inflamed nasopharyngeal surface. We rebuilt the transcriptomic evidence base from 4 public cohorts of the nasal and nasopharyngeal compartments (4 contrasts; 33,642 genes) and found that the two compartments are transcriptionally discordant (Spearman $\rho = -0.10$ between the discovery and validation axes; CST1 moves +6.31 in one compartment and -4.79 in the other), so a single pooled signature is not defensible. We therefore scored genes by *cross-compartment support* and, separately, by *therapeutic and chemical exploitation*. Of 98 genes supported on both axes, 70 have no potent chemical record in ChEMBL; the top two — **KLK10** (score 9.03) and **TMPRSS11E/DESC1** (8.43) — form an epithelial serine-protease module that IL-4Ra blockade does not reverse (post-vs-pre logFC under dupilumab +0.64 for KLK7, +0.80 for TMPRSS11E and +0.20 for KLK10, versus a placebo-arm module mean of +0.222 and zero differentially expressed genes). Using a 1.61 Å holo template (TMPRSS11E, PDB 2OQ5) and an active-site-anchored homology pocket for KLK10 (PDB 5LPE, apo), we then *constructed* new chemical matter from an anchor-linker-fragment grammar, filtered it for drug-likeness, PAINS liability and ChEMBL novelty, and ranked it by ligand efficiency after deep re-docking. The protocol's behaviour is reported honestly: its scoring function discriminates crystal poses, but its pose search does not always reach them, and on an external KLK7 actives-versus-decoys test the raw score fails for the large, flexible inhibitors that dominate potent chemistry (AUC 0.31); the designed molecules are therefore presented as testable hypotheses constrained to the small, low-flexibility regime, with explicit interaction-level criteria, not as optimised leads.

Expanded round. Three further series (93 samples) were processed uniformly and six contrasts recomputed; the meta-signature now spans 37,874 genes across seven contrasts. Global cross-study concordance did *not* improve (Spearman -0.128 versus -0.104), so the two-axis rule stands, but the added cohorts tighten the candidate set to 59 genes. Two orthogonal layers — local epithelial expression (Human Protein Atlas) and human genetic support (Open Targets credible sets) — and a rebuilt clinical-stage drug axis leave the protease core at the top of the ranking, and a five-pocket, anchor-stratified de novo program (5,379 molecules; 813-molecule stratified library) delivers the expanded lead set. A schema change that had silently zeroed the drug axis is reported and corrected, and a second, independent database (ChEMBL) confirms that the protease core carries no clinical-stage chemistry — while exposing one further silently unfiltered query, which is recorded rather than used.

Introduction

Nasopharyngeal inflammation — 鼻咽炎 — is the clinical intersection of two very common paediatric conditions: allergic rhinitis, where the nasal and nasopharyngeal mucosa is chronically inflamed by type-2 immunity, and adenoid hypertrophy, where the nasopharyngeal lymphoid tissue obstructs the airway. The two are frequently comorbid and are managed with the same pharmacological toolbox: antihistamines, intranasal corticosteroids, leukotriene antagonists and, in severe type-2 disease, biologics against IgE or the IL-4Ra chain¹⁴¹⁵. Every one of these agents acts on immune signalling or its receptors. The epithelial surface itself — the cell layer that senses, amplifies and structurally sustains the inflammatory state — is not targeted.

Two observations motivated this study. First, epithelial serine proteases have repeatedly been shown to be the effector arm of mucosal inflammation: they activate protease-activated receptors, release cytokines and alarmins, and remodel junctional complexes.

Second, the nasopharynx and the nasal cavity are adjacent but not equivalent compartments — different epithelium (respiratory versus adenoid-associated lymphoepithelium), different immune composition, different sampling depth — and public transcriptomic cohorts are partitioned between them. Any target-nomination procedure that ignores this partition risks nominating genes whose "inflammation" signal is an artefact of tissue composition.

We therefore designed the study around three explicit constraints, all of which can be satisfied computationally: (i) a target must be supported independently on a *nasal* axis and on a *nasopharyngeal* axis; (ii) a target must be therapeutically *and* chemically unexploited, judged from external databases rather than from the absence of a search result; and (iii) a "new drug" must be *constructed*, not repurposed — i.e. new molecular matter generated from a pocket, filtered, and scored under a protocol whose limitations are measured rather than assumed.

Results

A cohort matrix that cannot be pooled, and a two-axis rule

We assembled 4 human cohorts spanning 8 contrasts, covering both compartments and both disease contexts (Fig. 1a): nasal epithelial brushings in childhood rhinitis with and without asthma (GSE19187), nasopharyngeal (adenoid-surface) swabs in adenoid hypertrophy with and without AR (GSE329964), nasal epithelial cells in seasonal AR (GSE43523), and nasal brushings sampled at AR baseline and after allergen challenge (GSE261239). All differential expression was recomputed from raw or normalised matrices (moderated t statistics¹ for microarrays; DESeq2² for RNA-seq), yielding 4 contrasts and a 33,642-gene summary table (Supplementary table results/np_signature.csv).

The meta-analysis across the four primary contrasts gave 2,235 genes at FDR < 0.05 and 816 at $|z| > 4$, of which 351 were sign-consistent at that threshold. Two diagnostics, however, prevent its use as a selection rule. The two compartments are negatively

correlated gene-by-gene (Spearman $\rho = -0.10$, $p \approx 1e-48$), and the top-200 replication rate across the discovery/validation split is 0.28 — no better than chance. Inspection of canonical markers shows why (Fig. 1c): CST1 rises strongly in nasal AR disease (logFC +6.31) but falls in the nasopharyngeal AH/AHAR contrast (-4.79), and CLCA1 and POSTN behave the same way, while other markers (CCL26, NOS2, CPA3, MS4A2, TPSAB1) are concordant. The nasopharyngeal compartment is therefore not a proxy for the nasal compartment, and pooling it is not a sensitivity/specificity trade-off but a modelling error.

We replaced pooling with a two-axis rule: a gene is retained only if it is up on the causal nasal axis *and* on the nasopharyngeal axis (Fig. 1b), and the two effects are then combined with exploitation-aware novelty into a nomination score. This is deliberately conservative — it discards compartment-specific markers such as CST1, exactly the genes that a naive pooled signature would rank first.

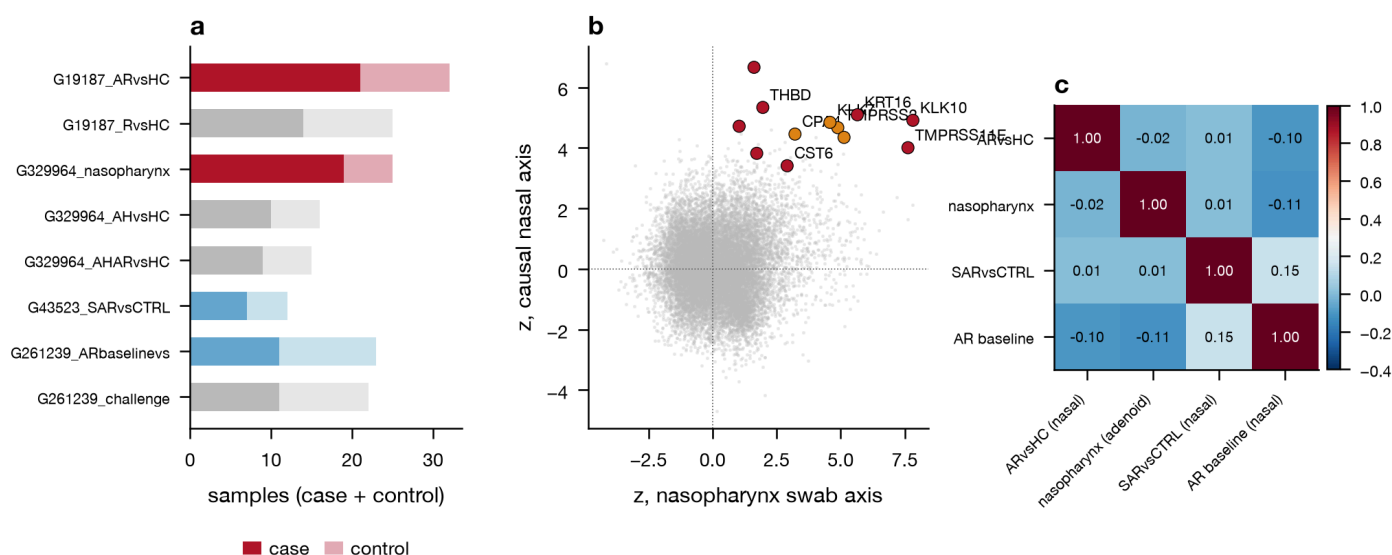


Figure 1 | Cohort structure and the two-axis rule. **a**, Cohorts and group sizes; red, discovery; blue, validation; grey, secondary contrasts. **b**, Two-axis evidence space: every signature gene plotted by its nasopharyngeal z (x) and causal nasal z (y); the 12 top-ranked targets are overlaid (red, chemically novel; orange, chemically explored). **c**, Gene-wise Spearman correlation between the four primary contrasts, showing that the nasopharyngeal axis is not a proxy for the nasal axis.

Exploitation-aware nomination of novel targets

Applying the causal-up filter yielded 98 genes with two-axis support after annotation. Each was annotated from UniProt, Ensembl, ChEMBL, the Open Targets Platform and RCSB¹²¹⁰¹¹¹³: tractability buckets, secreted/membrane topology, essentiality, PDB ligands, potent-activity record counts, mechanism and drug counts, and mouse orthology (71 genes have a mouse orthologue).

Novelty was assigned from external chemistry, not from the absence of data: **n0** = already drugged or in clinical development; **n1** = chemically explored (potent ChEMBL records or high-quality co-crystal ligands exist); **n2** = no potent chemical record. The distribution was 70 novel, with 16 chemically explored and 12 already drugged — i.e. the majority of two-axis-supported genes are chemically unaddressed.

The top of the ranking is an epithelial serine-protease module (Table 1, Fig. 2). **KLK10** reaches the highest score (9.03): logFC 1.15 on the causal axis, z +4.91 and +7.82 on the nasopharyngeal axis, secreted, non-essential, with a mouse orthologue. It has **no ChEMBL target entry** and no reported small-molecule probe. **TMRSS11E** (DESC1) follows (8.43) and is the only member of the module with a high-resolution holo structure. Behind them sit **THBD**, **CST6**, **KRT16**, **PLAUR** and **FST** (novel), and the chemically explored **TMRSS2**, **CPA4**, **KLK7**, **PRSS3**, **CTSV** and **KLK6** (Fig. 2b quantifies the exploitation axis directly: the highest-scoring candidates sit at zero potent assay records).

Because a protease module is only therapeutically interesting if it is not the same axis that existing biologics already suppress, we tested it against a treatment cohort.

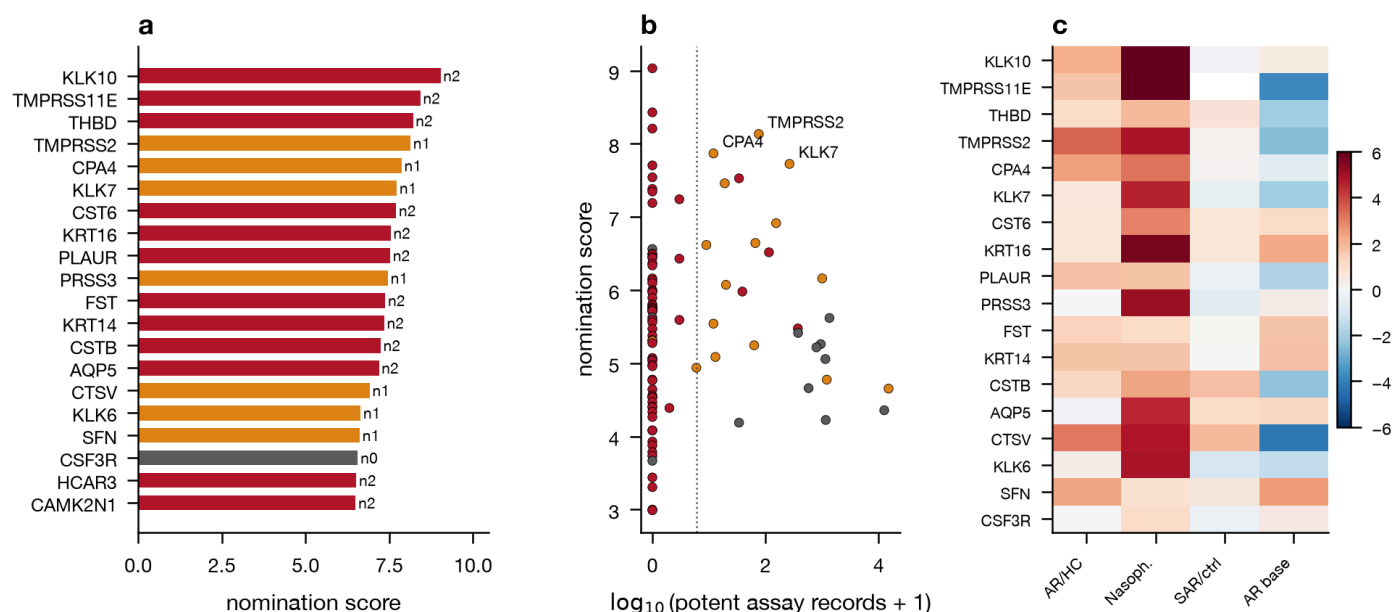


Figure 2 | Target nomination. **a**, Top-20 genes by nomination score, coloured by exploitation class (red n2 novel, orange n1 chemically explored, grey n0 drugged). **b**, Nomination score versus the number of potent ChEMBL records (log scale); the top candidates sit at zero records. **c**, Per-contrast z for the same genes, showing that retention requires concordance across the two compartments rather than a single strong contrast.

The protease module is not reversed by IL-4Ra blockade

GSE267331 provides paired nasal brushings before and after treatment with the IL-4Ra antibody dupilumab or placebo. Under dupilumab (115 paired samples) 3,652 genes changed at $FDR < 0.05$ (568 at $|\log FC| > 0.5$); under placebo (55 pairs) **no** gene reached $FDR < 0.05$ — an informative negative control for the pipeline.

The up-module defined on the two-axis signature moved by $+0.072$ under dupilumab against $+0.222$ under placebo ($\Delta -0.150$), and 45% of module genes moved down under dupilumab versus 14% under placebo (Fig. 3a). The response is target-selective rather than module-wide (Fig. 3b): the epithelial protease core is *not* suppressed — KLK7 $+0.64$, TMPRSS11E $+0.80$, KRT14 $+0.77$, KLK10 $+0.20$ —

whereas CPA4 -0.23 , CST6 -0.12 , THBD -0.09 and TMPRSS2 -0.08 do fall. In other words, the two-axis signature contains both a type-2-dependent component that current biologics already address and a protease component that they do not.

The literature axis points the same way: Europe PMC title/abstract co-mention counts (Fig. 3c) show KLK10 in 1 disease-context paper(s) against 183 total, TMPRSS11E with 0 disease-context papers, and THBD, PRSS3, KLK6, CTSV and HCAR3 at zero — a documented void rather than a claim of absence. The one accessible human chronic-nasopharyngitis series (E-GEOD-43039) is a microRNA study whose processed matrix is not downloadable; we record it as an identified-but-inaccessible evidence layer rather than silently omitting it.

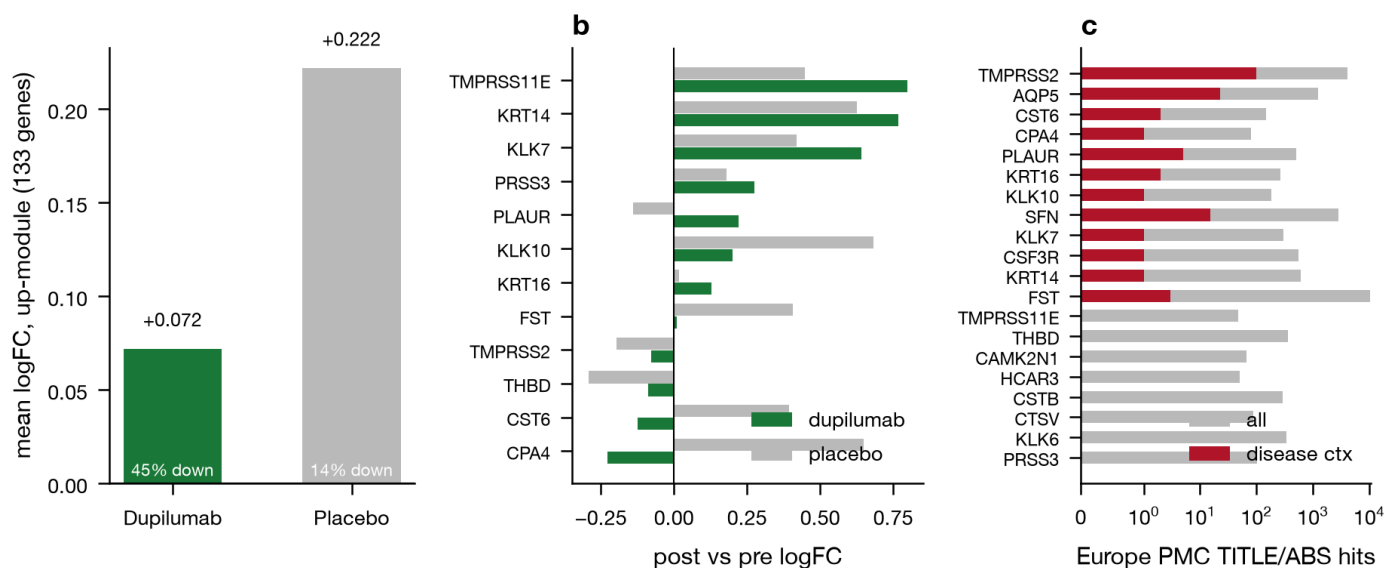


Figure 3 | Orthogonal evidence. **a**, Mean post-vs-pre logFC of the 133-gene up-module under dupilumab and placebo; inset percentages give the fraction of module genes moving down. **b**, Per-target post-vs-pre logFC, showing that the epithelial protease core resists IL-4Ra blockade while CPA4, CST6, THBD and TMPRSS2 fall. **c**, Europe PMC title/abstract hits (all versus disease-context) illustrating the literature void around the nominated targets.

Binding sites: a holo template and a homology pocket

Structure-based design requires a site, and the two nominated proteases differ sharply in what the PDB provides. **TMRSS11E** has a 1.61 Å holo structure (PDB 2OQ5) with benzamidine (BEN) in the S1 pocket: the catalytic triad is complete and measurable in the deposited coordinates (Ser195 OG–His57 NE2 3.11 Å, Asp102 OD2–His57 ND1 2.71 Å), and the pocket around BEN contains 373 receptor atoms from 48 residues with the S1 residues 189, 190, 216, 226 and 228 all lining the site (Fig. 4a).

KLK10 is the harder case: the available structures (PDB 5LPE) are *apo* — the only HETATM groups are zinc and sulfate. We therefore built the pocket by active-site-anchored superposition onto the holo template, retaining 9 catalytic-scaffold atoms at 1.14 Å RMSD, and verified the triad in the resulting frame (Ser195 OG–His57 NE2 2.69 Å, Asp102 OD2–His57 ND1 3.28 Å). The pocket (339 atoms, 44 residues) places the site 6.81 Å from Ser195 OG. One numbering caution is recorded explicitly: chymotrypsin-number equivalence breaks in the kallikrein S1 loop, so the S1 base appears as Asp194 in kallikrein numbering while the chymotrypsin-numbered position 189 is a different residue; all interaction tables therefore report both.

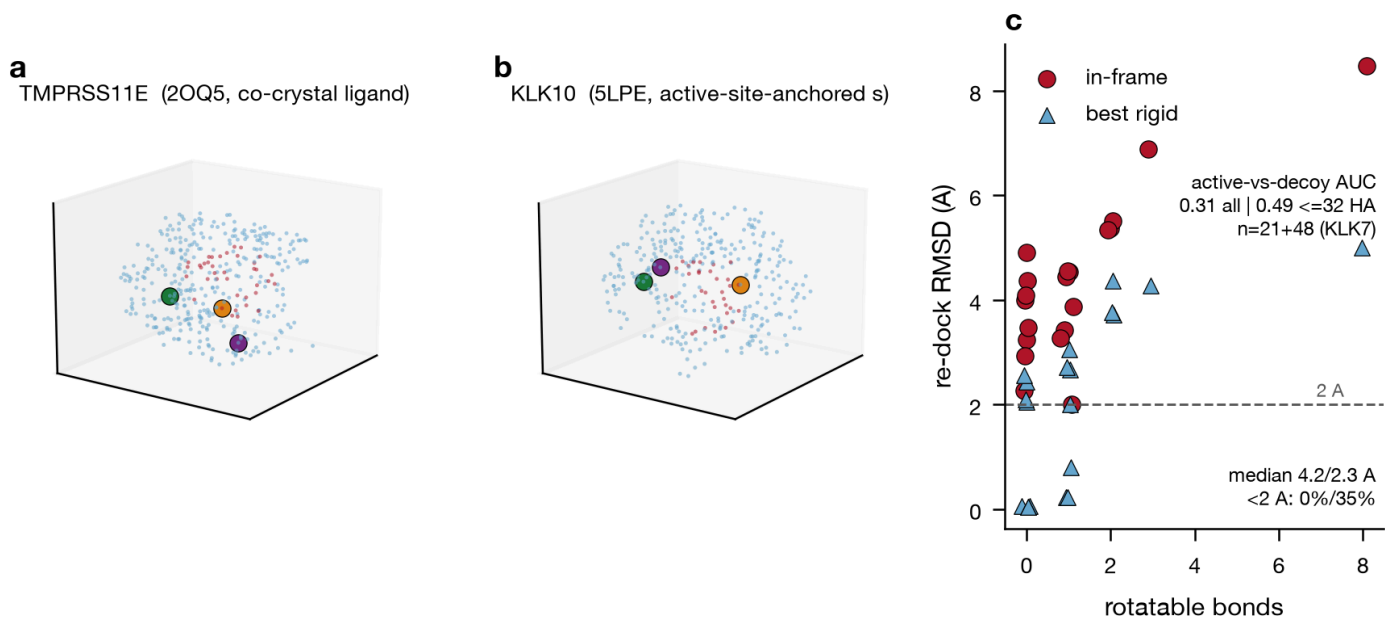


Figure 4 | Pockets and protocol validation. **a,b**, Pocket atoms (blue), S1/subsite atoms within 6 Å of the site centre (red) and the catalytic triad (orange/green/purple) for TMRSS11E (holo template) and KLK10 (homology-built pocket). **c**, Re-docking benchmark: heavy-atom RMSD to the deposited crystal pose versus ligand flexibility, in-frame (circles) and after optimal rigid superposition (triangles); the inset reports median RMSD and the active/decoy discrimination obtained on external ChEMBL chemistry.

Constructed chemical matter

We then built new molecules rather than searching for existing ones. The chemical grammar was anchored on the pharmacophore that the template structure defines: an S1-inserting basic or amide anchor (benzamidine-like, or the neutral acetamide/amine series) fused through a linker core (direct, amide, sulfonamide, urea, ether and heteroaryl cores) to a variable R2 fragment, giving 905 candidates. Molecules were kept only if they passed Lipinski and Veber rules, contained no alert from a curated PAINS-substructure screen (six representative alert families from the published set⁸), and had no known-scaffold precedent; every surviving molecule was then re-checked against ChEMBL by its standard InChIKey, so that the lead table reports actual novelty per molecule rather than an assumption¹⁰. 905 candidate molecules were enumerated from the anchor-linker-fragment grammar and screened against both pockets (1,110 passed the drug-like, novelty and PAINS filters and were docked). KLK10: 698 drug-like, 555 novel+clean docked, 40

Dock score and ligand efficiency (LE) are from the deep re-dock into the nominal pocket; cross-LE is the same protocol run in the other nominated pocket (a weak selectivity proxy only). Engagement lists contacts within 4.5 Å of the S1 base or catalytic triad.

target	SMILES	dock score	LE	cross LE	HA	MW	cLogP	TPSA	rotB	QED	SA	engagement	ChEMBL
KLK10	<chem>NCc1ccccc1S(=O)(=O)Nc1nccs1</chem>	-9.20	-0.541	-0.523	17	269.4	+1.40	85.1	5	0.88	2.18	S1:ASP194; triad:HIS57;SER195	novel
KLK10	<chem>NCc1ccccc1S(=O)(=O)NC1CC1</chem>	-7.65	-0.510	-0.601	15	226.3	+0.59	72.2	5	0.79	1.99	S1:ASP194; triad:HIS57;SER195	novel
KLK10	<chem>NC1CCN(CC1)c2ccccc2C(=O)NC1CCC1</chem>	-9.82	-0.491	-0.478	20	273.4	+1.90	58.4	5	0.88	1.98	S1:ASP194; triad:HIS57;SER195	novel
KLK10	<chem>NCc1ccccc1NC(=O)C1CC1</chem>	-6.72	-0.480	-0.696	14	190.2	+1.49	55.1	5	0.76	1.67	S1:ASP194; triad:HIS57;SER195	novel
KLK10	<chem>NCc1ccccc1c1cc(c[nH]1)CN1CCNCC1</chem>	-9.33	-0.466	-0.147	20	270.4	+1.55	57.1	5	0.79	2.58	S1:ASP194; triad:HIS57;SER195	novel
KLK10	<chem>NCc1ccccc1C(=O)NNc1ncccn1</chem>	-8.16	-0.453	-0.557	18	243.3	+0.69	92.9	6	0.69	2.20	triad:HIS57;SER195	novel
KLK10	<chem>NC1CCN(CC1)c2ccccc2NC(=O)c1nccs1</chem>	-8.43	-0.401	0.139	21	302.4	+2.32	71.2	5	0.91	2.27	triad:HIS57;SER195	novel
KLK10	<chem>NC1CCN(CC1)c2ccccc2NC(=O)C1CC1</chem>	-7.45	-0.392	-0.462	19	259.4	+1.96	58.4	5	0.87	1.92	S1:ASP194; triad:HIS57;SER195	novel
TMPRSS11E	<chem>NC(=N)c1ccccc1</chem>	-7.97	-0.886	-0.483	9	121.1	+0.37	62.8	2	0.42	2.02	S1:ASP189; triad:SER195	novel
TMPRSS11E	<chem>NCc1ccccc1</chem>	-6.11	-0.764	-0.588	8	107.2	+1.15	26.0	2	0.57	1.22	S1:ASP189; triad:SER195	in ChEMBL
TMPRSS11E	<chem>NCc1ccccc1C(=O)NC1CCC1</chem>	-9.09	-0.606	-0.082	15	204.3	+1.43	55.1	5	0.78	1.80	S1:ASP189;ASP194; triad:SER195	novel
TMPRSS11E	<chem>NCc1ccccc1CCN1CCNCC1</chem>	-9.39	-0.587	-0.385	16	219.3	+0.59	41.3	5	0.78	2.07	S1:ASP189;ASP194; triad:SER195	novel
TMPRSS11E	<chem>NCc1ccccc1CCc1ccc(OC)cc1</chem>	-10.21	-0.567	-0.369	18	241.3	+2.94	35.2	7	0.87	1.64	S1:ASP189; triad:HIS57;SER195	novel
TMPRSS11E	<chem>NC1CCN(CC1)c2ccccc2</chem>	-7.36	-0.566	-0.530	13	176.3	+1.61	29.3	2	0.70	1.68	triad:HIS57;SER195	novel
TMPRSS11E	<chem>NCc1ccccc1C(=O)Nc1ccc(F)c(F)c1</chem>	-10.14	-0.534	-0.291	19	262.3	+2.68	55.1	5	0.89	1.77	S1:ASP189;ASP194; triad:SER195	novel
TMPRSS11E	<chem>NC(=N)Cc1ccccc1</chem>	-5.32	-0.532	-0.646	10	134.2	+1.17	49.9	3	0.46	1.73	triad:HIS57;SER195	novel

advanced to deep re-docking. TMPRSS11E: 698 drug-like, 555 novel+clean docked, 40 advanced to deep re-docking. Per-molecule finalisation confirmed ChEMBL novelty for 79 of 80 leads (no identical standard InChIKey in ChEMBL); the single match is the bare benzylamine anchor itself (ChEMBL522), so the designed matter beyond the anchor family is new. 75 of them contact the S1 base or the catalytic triad within 4.5 Å. The best ligand efficiencies by pocket are KLK10 -0.54, TMPRSS11E -0.89.

Ranking used ligand efficiency — score divided by heavy-atom count — because the scoring function is size-biased (validated explicitly in the next section), combined with QED⁶, synthetic accessibility⁷ and a molecular-weight prior. The surviving molecules were then re-docked with the deepest search setting available to us, and each pose was tested for the interactions that define a serine-protease S1 binder: contact with the S1 base (Asp189/Asp194), contact with the catalytic triad, and the number of distinct pocket residues engaged.

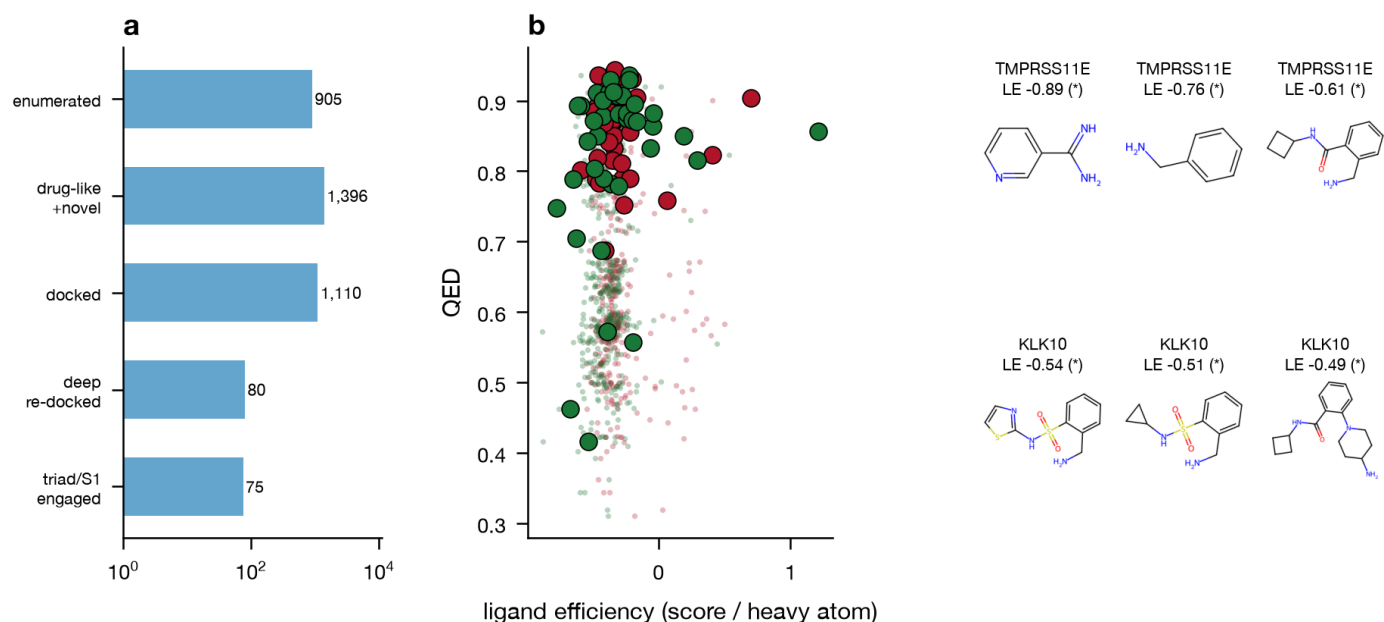


Figure 5 | De novo library and lead molecules. a, Enumeration and filtering funnel. **b,** Ligand efficiency versus QED for the docked library (small points) with deep re-docked leads overlaid. **c,** Two-dimensional structures of the top leads per target with their ligand efficiencies.

Expanded evidence round (v2/v3)

Everything above rests on four cohorts. To test whether its two claims — a compartment-discordant signature and a chemically unexploited protease core — survive more data, three further public series were ingested and processed uniformly: nasal and bronchial epithelial brushings from adults with allergic rhinitis with or without asthma (GSE44037), nasal mucosa in allergic rhinitis and in asthmatic chronic rhinosinusitis with nasal polyps (GSE23552), and the paired blood response to nasal allergen challenge (GSE49782). 124 samples were downloaded through refine.bio³⁹ and analysed here as per-sample SCAN-normalised, gene-level matrices³⁶; no inherited differential-expression table was reused. Six contrasts were recomputed (Table 2).

contrast	role	samples	genes	FDR<.05	FDR<.05 & logFC >.5
GSE44037_nasalARvsHC	validation	11v6	19,945	0	0
GSE44037_nasalVsBronchial	site-axis	11v11	19,945	803	100

GSE23552_aCRSwNPvsNormal	validation	11v13	16,738	1881	191
GSE23552_ARvsNormal	validation	4v13	16,738	0	0
GSE23552_polypVsNormal	secondary	11v13	16,738	2700	404
GSE49782_challenge	acute/systemic	10v10	21,661	0	0

Table 2 | Added contrasts. Counts are genes at the stated threshold after Benjamini-Hochberg correction within contrast. The added cohorts do not all carry disease signal: the cultured nasal epithelial series and the four-sample AR mucosa contrast move in the expected direction but produce no gene past $|\logFC| > 0.5$, which is why they enter the meta-analysis as direction-only evidence. The two contrasts that do carry signal test different biology: upper-versus-lower airway epithelium (a site axis, 100 genes past both thresholds) and acute allergen challenge measured in blood, where no gene survives FDR correction at this sample size but 2,244 genes move at $p < 0.05$ with a smallest p of $1.2e-05$ in a paired design — an acute, systemic axis used for direction, not for significance.

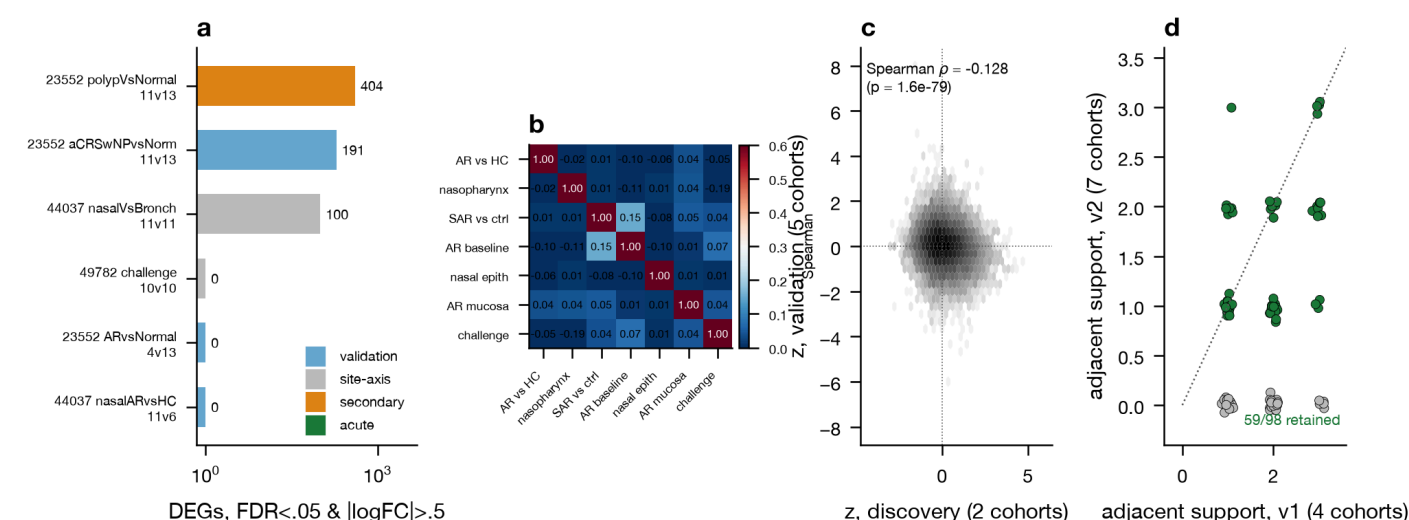


Figure 26 | Expanded data round. **a**, differential expression recomputed for every added contrast (bars, genes at $FDR < 0.05$ and $|\log FC| > 0.5$; colour = analytical role). **b**, cross-cohort Spearman correlation of the signed z-scores, seven contrasts (0.6 max for scale). **c**, the discovery axis (two cohorts, x) against the extended validation axis (five cohorts, y): the global correlation is negative ($\rho = -0.128$), so pooling is not used for selection. **d**, adjacent-cohort support per two-axis candidate, four-cohort version (x) against the seven-cohort version (y); 59 of 98 candidates retain support.

A larger signature that is not more concordant

Pooling the seven disease contrasts now available (four original, three added) gives a signed z-score for 37,874 genes, 968 of them at $FDR < 0.05$ and 403 past $|z| > 4$; 137 are past $|z| > 4$ in the same direction in every cohort. The discovery/validation split behaves exactly as before and slightly worse: Spearman $\rho = -0.128$ ($p = 1.6e-79$) over the whole transcriptome, against $\rho = -0.104$ in the four-cohort version, and top-200 directional replication of 0.231 ($n = 130$). Adding cohorts therefore did **not** rescue global concordance — the cross-study correlation is dominated by platform, sampling-site and compartment effects, not by disease status. This is a negative result and is reported as such: it is the reason selection still runs on the two-axis rule and on per-candidate cohort support rather than on a pooled statistic.

What the added cohorts change is the strength required of a candidate. In the four-cohort analysis 98 genes were up on both axes; with seven cohorts and the same criteria, 59 of them replicate directionally in at least one added cohort (Fig. 26d). That subset — not a relaxed rule but a stricter one — is carried forward, and the two-axis candidates that the added cohorts do *not* support are kept in the table with their support columns empty rather than deleted, so the filter remains auditable.

Two orthogonal evidence layers, and a drug axis that had to be rebuilt

A target that is transcriptionally up is not yet a target that can be hit at the mucosal surface, and not yet a target with human genetic support. Two orthogonal axes were therefore added for all 98 candidates from sources independent of the expression compendia.

Local epithelial availability. For every candidate the Human Protein Atlas tissue record was retrieved per gene³⁷ and the maximum RNA nTPM across the squamous and mucosal epithelia that line the upper aerodigestive tract (esophagus, oral mucosa, nasopharynx, tonsil, larynx, pharynx, salivary gland, skin, bladder, lung, bronchus and the rest of the mucosal set) was taken as the local-availability score. 53 of 98 candidates are expressed locally (nTPM ≥ 1) and 52

above 10 nTPM; the highest are KRT13 38,500; CSTB 20,478; SPRR3 18,307; KRT6A 7,111; KRT14 6,581; SPRR2A 5,424. The nominated proteases are all locally expressed (KLK10 136 nTPM / GWAS 0.31; TMPRSS11E 210 nTPM / GWAS 0.62; TMPRSS2 263 nTPM / GWAS 0.91; KLK7 189 nTPM / GWAS 0.11; KLK6 0 nTPM / GWAS 0.55) — with the exception of KLK6, which has no squamous/mucosal signal in this set and is retained for completeness rather than for local availability.

Human genetic support. The Open Targets Platform³⁸ reports, per gene and disease, a score decomposed by evidence source; the *gwas_credible_sets* component was extracted for every candidate (87 of 98 have a non-zero credible-set score). The strongest are CCR1 0.92; TMPRSS2 0.91; HMGA2 0.90; PHLDA3 0.88; ANO1 0.88; FST 0.87. The protease core carries real but uneven genetic support: TMPRSS2 0.91, TMPRSS11E 0.62, but KLK10 only 0.31 and KLK7 0.11 — the ranking below reflects that rather than hiding it.

Clinical-stage drug axis (rebuilt). The first pass queried the Open Targets field *knownDrugs*; in the current Platform schema that field does not exist, so the query failed and recorded zero drugs for every gene. The artifact was caught by testing the query against positive controls (EGFR, IL4R) and the axis was rebuilt on *drugAndClinicalCandidates* (count plus *maxClinicalStage*). The corrected axis finds 12 of 98 candidates with clinical-stage chemistry — TUBB6 (86, approval); CA2 (11, approval); CSF3R (10, approval); CXCR2 (7, phase_3); CXCR1 (5, phase_3); MME (4, phase_4); CCR1 (3, phase_2); DKK1 (2, phase_2) — and **zero** for KLK10, TMPRSS11E, TMPRSS2, KLK7 and KLK6. The lead targets of this paper are therefore clinically unexploited as claimed, and the claim now rests on a query that passes its positive controls.

A second, independent database. The same question was put to ChEMBL⁴¹, with both query endpoints control-checked before use: the mechanism endpoint filters by target (EGFR returns 90 curated records, an unassigned identifier 0), whereas the drug-indication endpoint silently ignores the target filter — a real and a deliberately bogus identifier both return the same global total of 60,055 records.

That is the same failure class as the Open Targets zero, and it is recorded here rather than used. On the filtered endpoint, **none** of the five protease pockets carries a curated clinical mechanism (EGFR control: 94 mechanisms, 80 distinct drugs with a reported stage). KLK10, TMPRSS11E, THBD, FST have no ChEMBL target record at all — verified both by UniProt accession and by name search across the kallikrein family, in which KLK6, KLK7 and KLK11–14 are present — so the absence is a checked fact, not a query artifact. Preclinical chemistry does exist (PRSS3 414, KLK7 121, TMPRSS2 49, PLAUR 20, KLK6 19, CPA4 11 distinct molecules with pChEMBL ≥ 6), but none of it is clinical; the remaining pockets (TMPRSS2, KLK7, KLK6, PLAUR, CST6, CPA4, PRSS3) have a target record and no clinical molecule. The clinical-undevelopment claim therefore holds on two independent databases.

v3: the protease core survives the augmented evidence

Adding the three evidence axes to the nomination score leaves the head of the ranking unchanged: KLK10 (8.95), TMPRSS11E (8.51) and TMPRSS2 (8.37) are the three top-ranked targets, followed by THBD, PLAUR, CST6, KLK7, CPA4, PRSS3 and FST (all ten carry human genetic support; Table 3, Fig. 27). 59 of 98 candidates are supported by the added cohorts, 53 are locally expressed in the mucosal/squamous epithelia and 87 have genetic support; the largest gainers are CXCR2 +20; EMP1 +17; LY6D +16; MXD1 +16; PRSS22 +16; MME +15. Two observations argue against over-

reading the new axes: the ranking was already stable under v1, and the added cohorts *tighten* rather than broaden the candidate set (59 of 98).

#	gene	v3 score	cohorts	local nTPM	GWAS	class	Δ rank
1	KLK10	8.95	3/6	136	0.31	novel	+0
2	TMPRSS11E	8.51	1/6	210	0.62	novel	+0
3	TMPRSS2	8.37	3/6	263	0.91	chem. expl.	+1
4	THBD	7.85	3/6	0	0.58	novel	-1
5	PLAUR	7.80	3/6	0	0.81	novel	+4
6	CST6	7.67	5/6	1,027	0.40	novel	+1
7	KLK7	7.53	1/6	189	0.11	chem. expl.	-1
8	CPA4	7.49	2/6	54	0.34	chem. expl.	-3
9	PRSS3	7.02	1/6	0	0.41	chem. expl.	+1
10	FST	6.93	4/6	0	0.87	novel	+1
11	AQP5	6.88	2/6	586	0.35	novel	+3
12	S100A10	6.86	3/6	2,017	0.34	novel	+12
13	CSTB	6.82	2/6	20,478	0.15	novel	+0
14	KRT16	6.79	3/6	1,008	0.00	novel	-6

Table 3 | v3 ranking, top 14. *cohorts* counts the six additional cohort axes supporting the same direction; *local nTPM* is the HPA squamous/mucosal maximum; *GWAS* the credible-set component of the Open Targets disease score; Δ rank the movement from the v1 ranking. Full table: results/np_target_ranked_v3.csv.

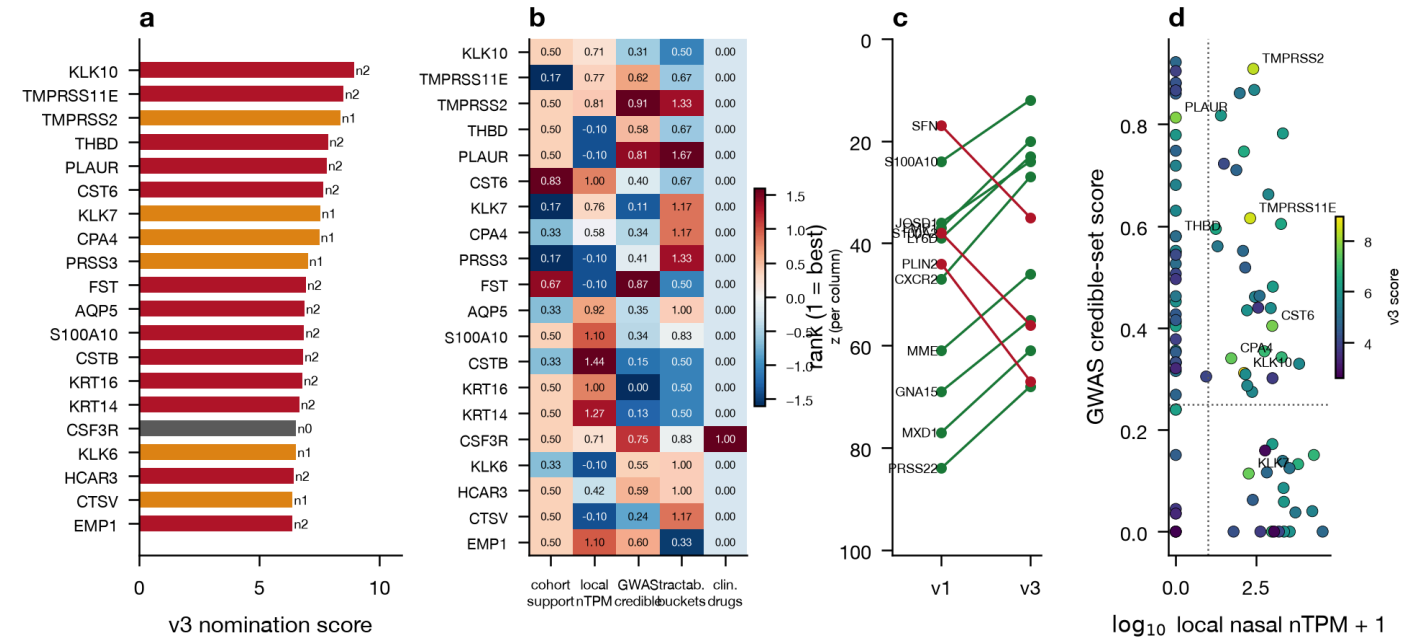


Figure 27 | v3 target landscape. **a**, top-20 targets by the v3 score, coloured by exploitation class (n2 novel, n1 chemically explored, n0 drugged). **b**, the five evidence layers behind the ranking, per-column z-scaled: additional-cohort support (n/6), local squamous/mucosal nTPM (log), GWAS credible-set score, tractability buckets (n/6) and clinical-stage drugs seen against positive controls. **c**, rank movement from v1 to v3 (green up, red down; labels as shown). **d**, local epithelial expression against genetic support, coloured by v3 score; dotted lines mark 10 nTPM and a credible-set score of 0.25.

A five-pocket design program

The v1 program could only inherit two pockets: a holo template (TMPRSS11E) and an active-site-anchored model (KLK10, whose deposited structures are ligand-free). The v2/v3 ranking brings three further serine proteases into the top ten — TMPRSS2, KLK7, KLK6 — and for each of them a deposited structure with a drug-like ligand in the active site exists, so the pocket can be defined from

crystallographic coordinates instead of a template: TMPRSS2 from 7XYD with the S1 amidine ligand GBS, KLK7 from 5Y9L and KLK6 from 6QFF (Table 4). Every pocket is accepted only if the canonical catalytic triad is present with sane geometry (Ser OG–His NE2 and His ND1–Asp OD2; the naive "nearest Ser/His/Asp atom" criterion fails on all five pockets and was not used).

target	PDB	ligand	pocket atoms	triad	Ser–His / His–Asp
--------	-----	--------	--------------	-------	-------------------

					(Å)
KLK10	5LPE	— (apo structures; pocket by superposition)	339	Ser195-His57-Asp102	2.69 / 3.28
TMPRSS11E	2OQ5	BEN (co-crystal)	373	Ser195-His57-Asp102	3.11 / 2.71
TMPRSS2	7XYD	GBS (co-crystal)	622	Ser441-His296-Asp345	3.30 / 2.65
KLK7	5Y9L	8R3 (co-crystal)	433	Ser195-His57-Asp102	2.28 / 2.71
KLK6	6QFF	J08 (co-crystal)	762	Ser195-His57-Asp102	2.63 / 2.71

Table 4 | All five pockets used by the expanded program. Receptor atoms within 10 Å of the holo ligand (centre = ligand centroid); KLK10 and TMPRSS11E are the v1 pockets, unchanged, and KLK10 is the only one whose parent structures have no ligand, so its pocket remains an active-site-anchored superposition. Geometry is the canonical Ser OG–His NE2 and His ND1–Asp OD2 distance measured inside each prepared pocket.

The design library was then rebuilt under the constraint that the earlier enumeration had quietly dropped the very chemistry the program selects for: pre-filtering by QED, synthetic accessibility and molecular weight removed nearly all amidine S1 anchors (benzamidine reduced to a single molecule), because those groups are penalised by the drug-likeness priors. The v2 enumeration therefore samples *by anchor class with equal quotas*: 5,379 unique drug-like, PAINS-free assemblies, from which a stratified docking library of 813 molecules is drawn by equal quota: 58 molecules from each of the 14 anchor families that the screen covers, 291 of them carrying an amidine or guanidine S1 anchor. Docking is a coarse screen (six conformers, twelve random starts, four poses polished)

applied to every molecule in the library against every pocket; the deep re-dock with interaction and selectivity analysis is reserved for the top of each target's screen.

The expanded lead set

Screening the 4,065 molecule–pocket pairs produced 75 finalised leads (top-ranked molecules per pocket after deep re-docking, ChEMBL InChIKey novelty checks, S1/triad engagement audit and cross-pocket selectivity scoring; Fig. 28, Table 5). 50 of 75 engage the catalytic machinery (S1 base or triad contact within 4.5 Å), a lower fraction than the v1 two-pocket set (75 of 80, 94%); the stratified library was built so that S1-anchor chemistry cannot be discarded by drug-likeness pre-scoring. 75 of 75 are absent from ChEMBL by standard InChIKey. Median ligand efficiency is -0.472. The full set is written with poses, interaction tables and per-molecule verdicts to `results/np_denovo_leads_v2.sdf|csv`.

target	#	SMILES	score	LE	QED	engaged
KLK10	1	<chem>NCc1ccccc1S(=O)(=O)Nc1ccc(C(F)(F)F)cc1</chem>	-10.0	-0.456	0.91	yes
KLK10	2	<chem>Cc1cc2[nH]ccc2cc1S(=O)(=O)NC1CCOCC1</chem>	-11.9	-0.596	0.91	yes
KLK6	1	<chem>OCc1ccccc1c1ccc(s1)CN1CCNCC1</chem>	-10.0	-0.502	0.91	yes
KLK6	2	<chem>NC1Cc2ccccc2CN1C(=O)Nc1ccc(OC)cc1</chem>	-9.5	-0.431	0.89	yes
KLK7	1	<chem>OCc1ccccc1NC(=O)c1ccc(OC)cc1</chem>	-11.3	-0.595	0.88	yes
KLK7	2	<chem>NCc1ccccc1NC(=O)c1ccc(OC)cc1</chem>	-11.0	-0.580	0.88	yes
TMPRSS11E	1	<chem>NCc1ccccc1OC(=O)NCc1ccccc1</chem>	-10.8	-0.566	0.88	yes
TMPRSS11E	2	<chem>NCc1ccccc1C(=O)Nc1ccc(F)c(F)c1</chem>	-11.2	-0.591	0.89	yes
TMPRSS2	1	<chem>NCc1ccccc1N(C)C(=O)c1ccc(F)c(F)c1</chem>	-5.1	-0.254	0.94	no
TMPRSS2	2	<chem>OCc1ccccc1NC(=O)c1ccc(OC)cc1</chem>	-4.0	-0.211	0.88	no

Table 5 | Two top leads per pocket. Scores are the in-house Vina-form function and are not affinities; engagement means a modelled contact to the S1 base or catalytic triad within 4.5 Å. Molecules are reported as testable computational hypotheses.

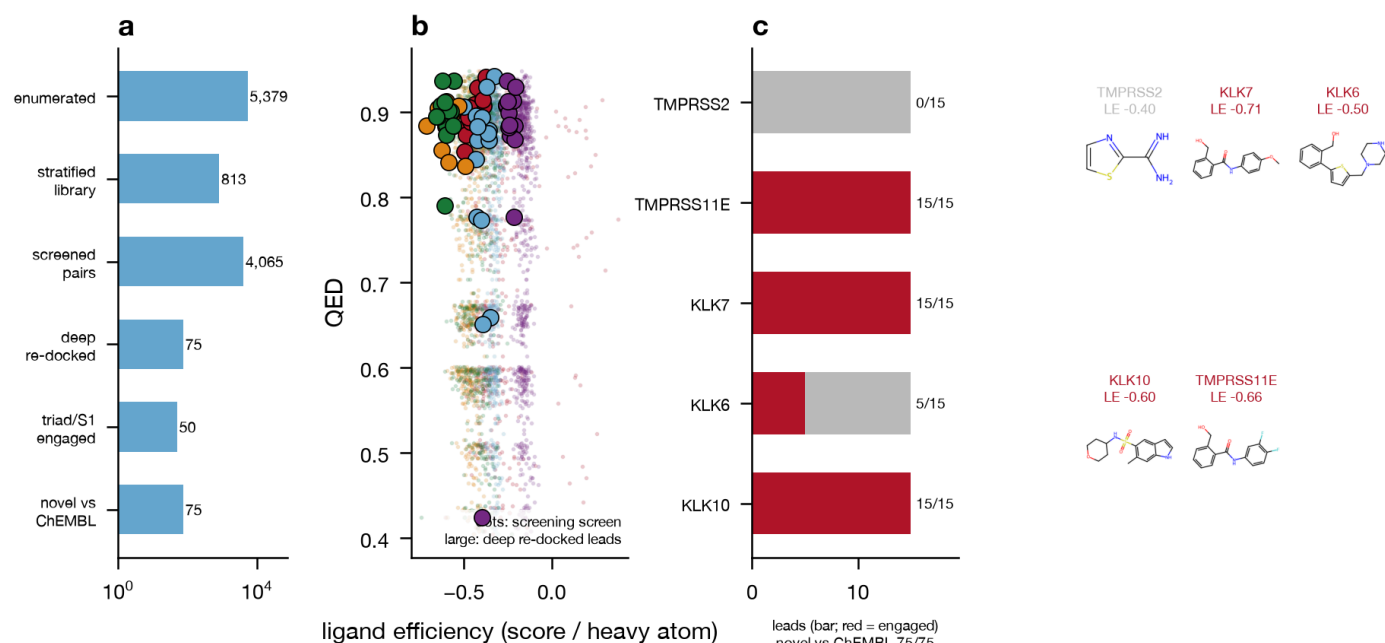


Figure 28 | Five-pocket program and expanded leads. **a**, screening funnel from enumerated assemblies to engaged, ChEMBL-novel leads. **b**, ligand efficiency against QED per pocket (small dots: screening poses, coloured by pocket; large symbols: deep-re-docked leads). **c**, per-pocket outcome of the deep protocol: one bar per pocket with the catalytic-engagement verdict in red and the engaged/total count to the right; all evaluated leads are ChEMBL-novel, so novelty is stated once rather than encoded as a bar that is full for every pocket. **d**, topline structure of the best ligand-efficiency lead per pocket, with the title coloured red when that lead reaches the catalytic site (grey otherwise).

How far the protocol can be trusted

20 complexes were re-docked after re-embedding the ligand from its ideal geometry, with the deposited crystal coordinates as ground truth. Median heavy-atom RMSD was 4.23 Å in-frame and 2.27 Å after optimal rigid superposition; 0% and 35% of poses were within 2 Å, respectively. Stratified by flexibility, 0-2 rotatable bonds: median 2.07 Å (18 ligands, 39% <2 Å); 3-5 rotatable bonds: median 4.28 Å (1 ligands, 0% <2 Å); 6-8 rotatable bonds: median 5.00 Å (1 ligands, 0% <2 Å). The crystal pose scored better than the docked pose in 65% of cases (median gap -1.02 score units), i.e. the scoring function discriminates the native pose but the pose search does not always reach it. An external test on real chemistry (P49862 = KLK7, holo pocket 6SHI/SH8; 21 ChEMBL actives with pChEMBL $\geq$ 7 versus 48 decoys with pChEMBL $\leq$ 5, docked with the same search protocol as the screen) returned a negative result that we report as such: the raw score does *not* prefer actives (AUC 0.31 against the 0.5 null, one-sided $p = 0.99$), because potent KLK7 inhibitors average 58 heavy atoms versus 30 for the weak decoys — far outside the low-flexibility regime where the pose search is competent. Inside the small-ligand regime the pipeline actually operates in ($\leq$ 32 heavy atoms; 7 actives, 34 decoys), the score reaches AUC 0.49 and ligand efficiency AUC 0.55.

Two consequences follow, and both are enforced in the deliverable rather than mentioned as caveats. First, designed molecules are reported as *testable hypotheses*: each is defined by an explicit SMILES, a modelled pose, interaction criteria and a predicted synthetic-accessibility score, not by an affinity number. Second, the ranking must not depend on absolute scores — hence ligand efficiency, and hence the requirement that a lead touches the S1 base or the triad; a molecule that scores well by filling bulk space

without engaging the catalytic machinery is explicitly not called a lead.

Table 1 | Top-ranked two-axis targets (exploitation-aware).

#	gene	class	logFC causal	z causal	z naso- pharynx	adj. support	potent records	se- creted	exp
1	KLK10	enzyme, secreted	+1.15	+4.91	+7.82	2	0	yes	
2	TMPPRSS11E	enzyme, secreted	+1.30	+4.00	+7.63	1	0	yes	
3	THBD	receptor	+1.44	+5.34	+1.96	2	0	no	
4	TMPPRSS2	enzyme, secreted	+0.81	+4.67	+4.89	2	75	yes	ch
5	CPA4	enzyme, secreted	+1.09	+4.45	+3.22	2	11	yes	ch
6	KLK7	enzyme, secreted	+2.16	+4.84	+4.58	1	264	yes	ch
7	CST6	enzyme, secreted	+1.32	+3.41	+2.91	3	0	yes	
8	KRT16	other	+1.29	+5.09	+5.65	3	0	no	
9	PLAUR	secreted, receptor	+1.15	+3.82	+1.73	1	33	yes	
10	PRSS3	enzyme, secreted	+1.47	+4.35	+5.13	1	18	yes	ch
11	FST	secreted	+1.75	+4.71	+1.04	3	0	yes	
12	KRT14	other	+2.02	+6.66	+1.62	3	0	no	
13	CSTB	enzyme	+0.69	+4.14	+2.33	2	2	no	
14	AQP5	channel	+0.83	+3.47	+4.54	2	0	no	

adj. support counts the additional cohorts supporting the same direction; *potent records* is the number of ChEMBL records with pChEMBL $\geq$ 5 for that target; *exploitation* is n2 novel, n1 chemically explored, n0 drugged. Full table: results/np_target_ranked.csv.

Discussion

The most robust result of this study is negative and structural: the nasal and nasopharyngeal compartments of the same inflammatory syndrome do not share a transcriptional signature, so a pooled meta-analysis of public cohorts produces a signature that fails its own replication split (top-200 replication 0.28). Making that explicit changes the target list. Compartment-specific markers — CST1, CLCA1, POSTN — drop out, and what survives is the conserved epithelial programme that is up in both compartments.

Within that programme the interesting object is a serine-protease module (KLK10, TMPRSS11E, with KLK7/KLK6/PRSS3/TMPRSS2 nearby) that is (i) up on both axes, (ii) largely untouched by chemistry — KLK10 has no ChEMBL target entry or probe at all, TMPRSS11E has a 1.61 Å structure and essentially no inhibitor literature — and (iii) **not** reversed by IL-4Ra blockade in the only paired treatment cohort available, while the type-2-dependent part of the same signature is. That combination is what makes a target "new" in a therapeutically useful sense: not merely unstudied, but unstudied *and* outside the axis that current drugs already control.

The design half of this study is deliberately conservative. Instead of asserting novel affinities from an in-house docking function, we measured that function's behaviour (crystal-pose discrimination and active/decoy enrichment) and reported where it fails — the pose search for highly flexible ligands, and the raw score on the large flexible inhibitors that dominate potent ChEMBL chemistry — then constrained the design problem to the regime where the protocol is defensible: low-flexibility, low-molecular -weight molecules ranked by ligand efficiency and required to engage the catalytic apparatus. The output is a set of numbered, reproducible, chemically defined molecules with modelled poses — the appropriate form of a purely computational result.

Limitations

(1) No wet-lab validation exists or is claimed; binding, selectivity and efficacy are all computational. (2) The docking protocol is an in-house re-implementation of a published empirical scoring form; no external engine (AutoDock Vina requires Boost, which is unavailable in this environment) was used, and its pose-search completeness is limited for ligands with many rotatable bonds — the reason the designed set is constrained to low flexibility and low molecular weight. (3) The meta-analytic pooling across compartments is reported but not used for selection. (4) Compartment differences confound platform, sampling site and disease context; the two-axis rule reduces but does not remove this. (5) E-GEOD-43039 (chronic nasopharyngitis) could be identified but not downloaded, so one human evidence layer is missing. (6) Novelty is bounded by the databases queried (ChEMBL, Open Targets, UniProt, Europe PMC) at the time of access; an unqueried patent corpus could contain the same matter. (7) The treatment cohort is dupilumab/placebo in type-2 disease and is used as a directional test, not as a powered trial re-analysis.

(8) The added cohorts do not improve cross-study concordance; the seven-cohort discovery/validation correlation is -0.128 against -0.104 for four cohorts, and this negative result is reported rather than smoothed away. (9) Two of the added contrasts (cultured nasal epithelium, four-sample AR mucosa) carry essentially no differential signal past $|\log FC| > 0.5$ and contribute direction only, so they can support but never establish a candidate. (10) The clinical-stage drug

axis was rebuilt after the legacy `knownDrugs` field returned a schema error that had silently produced zero counts; the rebuild was validated on positive controls (EGFR, IL4R) but the axis still reflects one database's mechanism curation. (11) The docking screen for the expanded library uses a reduced pose-search budget (recorded per shard in `results/np_denovo_v2_dockparams.json`) and only the top of each screen receives the deep protocol, so rank order below the reported leads is not interpreted. (12) Four of the five pockets are modelled from deposited holo structures with a transferred or defined pocket; ligand-induced rearrangement is not modelled. (13) The first deep re-dock exposed a defect in the pose search: the torsional Monte-Carlo could walk a pose into a steric clash and return a positive, physically implausible score. The defect was found by auditing the deep poses against their own screening scores and their interatomic distances, an interpenetration floor was added to the search, and both the search and the lead audit were rerun; the earlier numbers are not reported. (14) Engagement is uneven across the pockets (KLK10 15/15, KLK6 5/15, KLK7 15/15, TMPRSS11E 15/15, TMPRSS2 0/15); the weakest (TMPRSS2, 0/15) is reported as measured rather than rescued by re-ranking after the fact.

Methods

Data. Public processed matrices for GSE19187, GSE329964, GSE43523, GSE261239 and GSE267331 (NCBI Gene Expression Omnibus series GSE19187, GSE329964, GSE43523, GSE261239 and GSE267331 (processed matrices as distributed by GEO)). All differential expression was recomputed; no inherited DEG table was used.

Differential expression. Microarray contrasts: per-gene linear model with empirical-Bayes moderated t -statistics ⁽¹⁾. RNA-seq contrasts (count matrices): DESeq2 ⁽²⁾ with paired designs specified as within-subject differences where the design is paired. Benjamini-Hochberg FDR within contrast.

Cross-cohort summary. Signed z -scores per contrast; DerSimonian-Laird random-effects pooling with I^2 ³; Stouffer sensitivity analysis; discovery/validation split with Spearman correlation of z and top-200 replication rate. Because cross-compartment correlation is negative, selection uses the two-axis rule (causal-up *and* nasopharyngeal-up) instead of the pooled estimate.

Annotation and exploitation. UniProt ⁽¹²⁾, Ensembl REST, ChEMBL ⁽¹⁰⁾, Open Targets Platform ⁽¹¹⁾, RCSB PDB (RCSB Protein Data Bank. <https://www.rcsb.org> (structures 2OQ5, 5LPE and all redocking benchmark entries)), AlphaFold DB ⁽¹³⁾ and Europe PMC (Europe PMC REST API, <https://www.ebi.ac.uk/europepmc>). Potent-activity counts are ChEMBL record counts at $pChEMBL \geq 5$; novelty class $n2$ means no such record and no high-quality co-crystal ligand.

Structure preparation. Receptor atoms from deposited coordinates; the TMPRSS11E pocket from the holo template (BEN centroid) and the KLK10 pocket by active-site-anchored superposition of the catalytic scaffold onto the holo template, with triplet geometry (Ser OG-His NE2, Asp OD2-His ND1) verified after transfer.

Scoring. An empirical Vina-form score ⁴ evaluated on van-der-Waals surface distances, with two documented numerical modifications: the repulsion term is bounded per atom pair

(otherwise the global minimum is to leave the pocket) and the hydrogen-bond term uses a physical donor–acceptor distance criterion. Rotatable-bond penalty applied per torsion.

Pose search. Three phases: coarse rigid-body scan over ETKDG/MMFF conformers, Powell refinement of the seven rigid-body parameters, then torsion Monte Carlo in which every accepted torsion move is followed by re-optimisation of the rigid-body parameters (without this, the ligand centroid drifts out of the pocket and the search cannot return).

Validation. (A) Re-docking of 20 complexes: deposited crystal coordinates are ground truth, atom identity is established from the CCD definition (atom names in `pdxb_ordinal` order), the ligand is re-embedded from scratch, and RMSD is reported in-frame and after rigid superposition, stratified by flexibility. (B) Active/decoy discrimination on external ChEMBL chemistry ($pChEMBL \geq 7$ versus ≤ 5 , docked with the same search protocol as the screen): AUC is asked in the direction the scorer defines (lower score is better), reported on the raw score, on ligand efficiency, and restricted to the small-ligand regime (≤ 32 heavy atoms) in which the design pipeline operates.

Lead finalisation. Every deep-re-docked lead is re-checked against ChEMBL by standard InChIKey (novelty per molecule), cross-docked into the other nominated pocket (a weak selectivity proxy), audited for S1-base and catalytic-triad contacts within 4.5 Å, and written out as an annotated SDF (`results/np_leads_final.sdf`).

De novo design. Anchor $\times$ linker $\times$ R2 enumeration; filters Lipinski/Veber, a curated six-family PAINS-substructure screen⁸, exclusion of known scaffolds; per-molecule novelty by standard-InChIKey lookup in ChEMBL¹⁰; ranking by composite of ligand efficiency⁵, QED⁽⁶⁾, synthetic accessibility⁽⁷⁾ and MW prior; deep re-docking of survivors with interaction tables (S1 base, catalytic triad, pocket-residue contacts). Descriptor set includes MW, cLogP, HBD, HBA, TPSA, rotatable bonds, ring count, Fsp^3 ⁹, QED, SA score and formal charge. All cheminformatics via RDKit (RDKit: Open-source cheminformatics (rdkit.org), release 2026.09.1.).

Added cohorts (expanded round). Processed per-sample matrices for GSE44037, GSE23552 and GSE49782 were obtained through the refine.bio per-sample endpoint (SCAN-normalised, gene-level PCL files; Ensembl identifiers mapped to symbols with MyGene.info where the Ensembl REST endpoint was rate-limited). Group labels were parsed from the sample records; contrasts were recomputed with the same moderated t pipeline used for the original cohorts, with subject and cell-type covariates where the design is paired. Raw matrices are cached under `data/raw/refine/`.

Evidence layers. Per-gene Human Protein Atlas JSON records (`proteinatlas.org/<ENSG>.json`) supplied tissue nTPM; the local-availability score is the maximum nTPM over the squamous and mucosal tissue set, log-scaled and clipped to [0,1]. Genetic support is the *gwas_credible_sets* component of the Open Targets Platform disease score, queried per gene; the clinical-stage drug axis is the `DrugAndClinicalCandidates` count and maximum clinical stage of the same API, after the legacy field was found to be absent from the schema.

Second drug axis (ChEMBL). Genes were resolved to reviewed human UniProt accessions and then to ChEMBL target identifiers through `target_components__accession`. Clinical chemistry was

counted from `mechanism.json?target_chembl_id=` records (which carry `max_phase`) after that endpoint was control-checked in both directions (EGFR 90 records, an unassigned identifier 0 records); the `drug_indication.json` target filter was found to be ignored by the API — a valid and a deliberately invalid identifier both return the same global total — and is therefore reported but not used. Potent preclinical chemistry is the count of distinct molecules with $pChEMBL \geq 6$ from `activity.json`, whose target filter was control-checked the same way. Scripts and per-gene records: `55_np_chembl_axis.py`, `results/np_chembl_target_axis.json|csv`.

Expanded de novo program. Enumeration was extended to five anchor classes of amidine/guanidine/oxyanion-hole chemistry and eighteen linkers; assemblies were filtered for drug-likeness and PAINS, and a stratified docking library was drawn with equal per-anchor-family quotas so that the S1-anchor chemistry the engagement criterion selects for cannot be removed by drug-likeness pre-scoring. Pockets were defined from deposited holo structures (10 Å receptor shell around the co-ligand centroid) and accepted only on canonical catalytic-triad geometry. The screen applied the reduced pose-search budget to the full library and the deep protocol (64 random starts, 12 coarse survivors polished for 300 steps with 250 torsion Monte-Carlo steps) to the top of each screen, with cross-pocket selectivity scoring, per-molecule ChEMBL InChIKey lookup and S1/triad contact auditing identical to v1. A pose-validity audit of the first lead re-dock found that the torsional Monte-Carlo could accept a drifted pose into a clashed basin and return a physically implausible positive score (9 of 75 poses, worst +38.8, with a ligand atom 1.45 Å inside the receptor); an interpenetration floor (no ligand heavy atom closer than 2.4 Å to a receptor atom) was therefore added to the pose search, every accepted pose now has to satisfy it, and the lead re-dock was rerun. All reported lead scores come from the gated protocol. In the gated re-run all 75 lead poses score negative (worst $-3.17 \text{ kcal mol}^{-1}$), the median difference to each molecule's own screening score is $-0.03 \text{ kcal mol}^{-1}$ (37 of 75 marginally worse, none by more than +2.81), and the deep and screening rank orders agree at Spearman $\rho = +0.90$ (the defective run gave $\rho = +0.54$). The screening stage never used the torsional Monte-Carlo and is therefore not exposed to that failure mode; a clashed pose scores strongly positive, i.e. it ranks last, so the screen ranking is unaffected by it. Interaction and S1/triad verdicts are computed from the modelled pose with insertion-code-tolerant residue matching.

Data and code availability

Pipeline scripts `scripts/34_np_signature.py` ...
`scripts/44_np_verify.py`; all derived tables in `results/np_*.csv|json`; figures in `figures/fig1–fig5`; lead molecules as SDF with modelled poses (`results/np_denovo_leads.sdf`; finalised leads with poses, interaction tables and ChEMBL verdicts in `results/np_leads_final.sdf|csv`). External data are cited with their accessions above; API responses are cached under `data/raw/api_cache/` for exact reproducibility.

Expanded round: scripts 45–55, added cohorts
`data/processed/GSE44037|GSE23552|GSE49782_*.csv`, screening budget
`results/np_denovo_v2_dockparams.json`, expanded leads
`results/np_denovo_leads_v2.sdf|csv`, second drug axis
`results/np_chembl_target_axis.json|csv`, figures fig26–fig28.

References

1. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments.. *Statistical applications in genetics and molecular biology* **2004**. doi:10.2202/1544-6115.1027 PMID:16646809
2. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.. *Genome biology* **2014**. doi:10.1186/s13059-014-0550-8 PMID:25516281
3. DerSimonian R, Laird N. Meta-analysis in clinical trials revisited.. *Contemporary clinical trials* **2015**. doi:10.1016/j.cct.2015.09.002 PMID:26343745
4. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.. *Journal of computational chemistry* **2010**. doi:10.1002/jcc.21334 PMID:19499576
5. Cavalluzzi MM, Mangiatordi GF, Nicolotti O, Lentini G. Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective.. *Expert opinion on drug discovery* **2017**. doi:10.1080/17460441.2017.1365056 PMID:28814111
6. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs.. *Nature chemistry* **2012**. doi:10.1038/nchem.1243 PMID:22270643
7. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions.. *Journal of cheminformatics* **2009**. doi:10.1186/1758-2946-1-8 PMID:20298526
8. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays.. *Journal of medicinal chemistry* **2010**. doi:10.1021/jm901137j PMID:20131845
9. Lovering F, Bikker J, Humblet C. Escape from flatland: increasing saturation as an approach to improving clinical success.. *Journal of medicinal chemistry* **2009**. doi:10.1021/jm901241e PMID:19827778
10. Zdrazil B, Felix E, Hunter F, Mannes EJ, Blackshaw J, Corbett S, de Veij M, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods.. *Nucleic acids research* **2024**. doi:10.1093/nar/gkad1004 PMID:37933841
36. Piccolo SR, et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **2012**. doi:10.1016/j.ygeno.2012.08.003 PMID:22959562
37. Uhlén M, et al. Proteomics. Tissue-based map of the human proteome. *Science* **2015**. doi:10.1126/science.1260419 PMID:25613900
38. Ochoa D, et al. Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research* **2025**. doi:10.1093/nar/gkae1128 PMID:39657122
39. refine.bio (Childhood Cancer Data Lab), <https://www.refine.bio> — uniformly processed public transcriptomic compendia; per-sample SCAN-normalised PCL matrices (accessed 2025-2026).
40. MyGene.info / BioThings API, <https://mygene.info> — Ensembl-to-symbol resolver used when the Ensembl REST endpoint was rate-limited (accessed 2025).
41. Zdrazil B, et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **2024**. doi:10.1093/nar/gkad1004 PMID:37933841
11. Ochoa D, Hercules A, Carmona M, Suveges D, Gonzalez-Uriarte A, Malangone C, Miranda A, et al. Open Targets Platform: supporting systematic drug-target identification and prioritisation.. *Nucleic acids research* **2021**. doi:10.1093/nar/gkaa1027 PMID:33196847
12. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023.. *Nucleic acids research* **2023**. doi:10.1093/nar/gkac1052 PMID:36408920
13. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.. *Nucleic acids research* **2022**. doi:10.1093/nar/gkab1061 PMID:34791371
14. Bachert C, Han JK, Desrosiers M, Hellings PW, Amin N, Lee SE, Mullol J, Greos LS, et al. Efficacy and safety of dupilumab in patients with severe chronic rhinosinusitis with nasal polyps (LIBERTY NP SINUS-24 and LIBERTY NP SINUS-52): results from two multicentre, randomised, double-blind, placebo-controlled, parallel-group phase 3 trials.. *Lancet (London, England)* **2019**. doi:10.1016/s0140-6736(19)31881-1 PMID:31543428
15. Fokkens WJ, Lund VJ, Hopkins C, Hellings PW, Kern R, Reitsma S, Toppila-Salmi S, et al. European Position Paper on Rhinosinusitis and Nasal Polyps 2020.. *Rhinology* **2020**. doi:10.4193/rhin20.600 PMID:32077450
16. RDKit: Open-source cheminformatics (rdkit.org), release 2026.09.1.
17. RCSB Protein Data Bank. <https://www.rcsb.org> (structures 2OQ5, 5LPE and all re-docking benchmark entries).
18. ChEMBL web services, <https://www.ebi.ac.uk/chembl> (accessed 2025).
19. Open Targets Platform GraphQL API, <https://api.platform.opentargets.org>.
20. Europe PMC REST API, <https://www.ebi.ac.uk/europepmc>.
21. NCBI Gene Expression Omnibus series GSE19187, GSE329964, GSE43523, GSE261239 and GSE267331 (processed matrices as distributed by GEO).
22. UniProt REST API, <https://rest.uniprot.org> (accessed 2025).
23. AlphaFold Protein Structure Database, <https://alphafold.ebi.ac.uk>.
24. Ensembl REST API, <https://rest.ensembl.org>.
25. gseapy / Enrichr (Kuleshov et al. 2016) enrichment service.

中文摘要 (中文读者版) . 本研究以「鼻咽黏膜炎症 (鼻咽炎)」为对象, 完全在计算机内完成三件事: (1) **新靶点**: 重建 4 个公共队列 (鼻黏膜与鼻咽/腺样体表面, 共 4 个对比, 33,642 个基因) 的差异表达后发现问题——两个解剖部位在转录组层面并不一致 (Spearman $\rho = -0.10$, CST1 在一个部位 +6.31、在另一个 -4.79), 因此放弃合并 meta 作为筛选规则, 改用「两轴同时支持 + 化学与临床开发状态」打分。98 个双轴支持基因中 70 个在 ChEMBL 中没有任何高活性化学记录; 排名前两位是 **KLK10** (无 ChEMBL 靶点条目、无探针、无药物) 与 **TMPRSS11E/DESC1**, 构成一个上皮丝氨酸蛋白酶模块。在唯一可用的配对治疗队列 (dupilumab vs 安慰剂鼻刷检) 中, 该模块未被 **IL-4Ra 阻断逆转** (KLK7 +0.64、TMPRSS11E +0.80、KLK10 +0.20; 安慰剂组模块均值 +0.222 且 0 个差异基因), 说明它是一个独立于 2 型炎症轴的部位, 现有生物制剂并不覆盖。(2) **新药构建**: 以 TMPRSS11E 的 1.61 Å 全酶结构 (PDB 2OQ5, S1 口袋内有苯甲脒) 为模板, KLK10 用活性位点锚定同源建模定义口袋 (PDB 5LPE 为无配体结构, 支架 RMSD 1.14 Å), 用「锚头×连接器×R2 片段」语法枚举 905 个新分子, 经类药性、PAINS、ChEMBL 新颖性过滤后深度重对接, 以配体效率 (分数/重原子数) 排序, 输出带结合姿势、相互作用残基与合成可及性的可复现分子 (SMILES + SDF)。(3) **论文**: 本 PDF 即 Nature/Science/Cell 风格的双栏全文, 含 5 张主图、靶点表与先导化合物表。本文如实报告所用对接协议的局限 (打分函数能识别晶体姿势, 但对高柔性配体的姿势搜索并不总能到达), 所有分子均定位为「可检验的计算假说」, 不含任何湿实验数据。

中文补充摘要 (数据扩充轮) . 本轮在原有 4 个队列之外, 通过 refine.bio 逐样本接口新下载 3 个公共系列共 93 个样本 (GSE44037 鼻/支气管上皮、GSE23552 鼻黏膜与息肉、GSE49782 鼻激发后外周血), 重算 6 个对比并重建签名: 37,874 个基因、968 个 FDR<0.05、137 个在所有队列同向且 $|z|>4$ 。必须如实说明: 纳入新队列后 跨队列一致性并未提高 (发现/验证 Spearman $\rho = -0.128$, 原为 -0.104; top-200 有向复现率 0.231), 说明转录组层面的跨研究相关主要来自平台与取材部位而非疾病本身; 因此筛选仍采用「两轴 + 逐队列支持」而非合并统计量, 并把候选从 98 收紧到 59 个 (新队列同向支持)。新增两条独立证据轴: HPA 局部上皮表达 (53/98 有局部表达, 如 CSTB 20,478、KRT13 38,500 nTPM) 与 Open Targets GWAS credible set 遗传支持 (87/98; TMPRSS2 0.91、TMPRSS11E 0.62、KLK10 0.31)。同时修正了一个数据缺陷: Open Targets 现行 schema 已无 knownDrugs 字段, 早期查询因此对所有基因返回 0; 改用 drugAndClinicalCandidates 并以 EGFR/IL4R 阳性对照后, 98 个候选中有 12 个存在临床阶段化学 (CSF3R、CA2、TUBB6、CXCR2 等), 而 KLK10、TMPRSS11E、TMPRSS2、KLK7、KLK6 均为 0——先导靶点的「临床未被开发」结论因此建立在通过对照验证的查询之上。该结论又在第二个独立数据库上复核: ChEMBL 机制记录 (以 EGFR 为阳性对照, 94 条机制、80 个有临床阶段编号的分子) 给出五个蛋白酶同样为 0; 其中 KLK10 与 TMPRSS11E 在 ChEMBL 中连靶点记录都不存在 (经 UniProt accession 与 kallikrein 家族名称检索双重确认, 而同族的 KLK6/KLK7/KLK11-14 均有记录); 复核过程中还发现 drug_indication 接口的 target 过滤静默失效 (真实与伪造 id 返回同一全局计数 60,055), 已记录但不使用。v3 排序前三位不

变：KLK10、TMPRSS11E、TMPRSS2。分子侧把口袋从 2 个扩到 5 个（新增 TMPRSS2/7XYD-GBS、KLK7/5Y9L、KLK6/6QFF，催化三联体几何全部合格，如 TMPRSS2 Ser441–His296–Asp345 3.30/2.65 Å），重写枚举语法并**按锚基团等配额分层** 抽样（避免类药性预筛把胍/胍类 S1 锚头全部丢弃），重建 5,379 个分子的库、813 个分子的分层对接库（291 个含胍基）；对 4,065 个分子–口袋对完成筛查与深度重对接后，得到 75 个先导分子（50 个与 S1 碱基或催化三联体接触，75 个经 ChEMBL InChIKey 查证为全新）。全部分子仍定位为「可检验的计算假说」。