MULTI-TASK PREDICTION OF SOLID-PHASE EXTRACTION PROTOCOL COMPONENTS FROM MOLECULAR DESCRIPTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Solid-phase extraction (SPE) is widely used to enrich trace pollutants before instrumental analysis, yet selecting cartridge chemistries, solvents and mixture ratios still relies on literature search and expert adaptation. Learning these choices from published methods is difficult because a pollutant can have several reported alternatives, solvent labels are highly imbalanced, and ratio annotations are incomplete. We propose SPEGen (Solid-Phase Extraction Protocol Generator), a multi-task framework that predicts SPE components from molecular descriptors. We aggregate published methods by pollutant to construct set-valued supervision that retains alternative reported choices. A shared encoder with task-specific heads predicts cartridge sets, workflow-step indicators, step-specific solvent sets and their ratios, with ratio training restricted to observed annotations. Rank-Gauss normalisation handles heavy-tailed descriptors, while a validation-tuned decoder selects label sets and suppresses solvents for absent steps. Bootstrap aggregation combines predictions from resampled training data. Experiments against six tabular baselines on a held-out pollutant split show that SPEGen-Bag achieves the highest cartridge and solvent F1 and the lowest ratio mean absolute error. Cumulative ablations show gains from normalisation and bagging, while decoding analyses reveal how predicted set size affects the trade-off between recovering reported alternatives and exact-match accuracy.

1 INTRODUCTION

Detecting trace pollutants in environmental samples requires enriching the analytes and reducing matrix interference before instrumental analysis. Solid-phase extraction (SPE) is widely used for this purpose because sorbents and solvents can be adapted to different compounds and sample matrices (Hennion, 1999; Poole, 2003). Developing an SPE method nevertheless requires several related decisions: which cartridge chemistry to use, which workflow steps to include, and which solvents and mixture ratios to assign to each step. These choices depend on analyte properties and sorbent–solvent interactions (Fontanals et al., 2007; Andrade-Eiroa et al., 2016). Analysts typically search for published procedures for related compounds and adapt them through expert judgement (Buszewski & Szultka, 2012). This motivates learning to recommend SPE components systematically from the properties of a target pollutant.

Molecular machine learning has made substantial progress in property prediction (Cherkasov et al., 2014; Wu et al., 2018; Yang et al., 2019), including chromatographic retention-time prediction (Domingo-Almenara et al., 2019; Bouwmeester et al., 2019). These molecular-property models support analytical prediction; selecting the components of a sample-preparation method requires a structured output. Related work in synthetic chemistry moves closer to experimental decisions: Gao et al. (2018) predict reaction conditions, and Vaucher et al. (2020) and Vaucher et al. (2021) extract or predict machine-readable procedures. For SPE, the learning target must accommodate both categorical selections and numeric ratios, while allowing multiple reported choices for the same pollutant.

To address this problem, we propose SPEGen (Solid-Phase Extraction Protocol Generator), which predicts SPE components from molecular descriptors. We aggregate cartridge, step and solvent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

labels by pollutant, producing set-valued supervision that retains alternatives across source methods. Reported ratios provide conditional numeric targets, and missing ratios are masked during training. The resulting formulation supports learning from partially annotated methods and multiple reported choices for each pollutant.

The predictor uses a shared encoder with task-specific heads for cartridges, workflow steps, step-wise solvents and ratios. Sharing the molecular representation allows the tasks to use common chemical information while retaining their distinct output types. To address sparse, imbalanced labels, cartridge and solvent prediction use focal and asymmetric losses, respectively. A validation-tuned decoder then selects label sets using label-specific thresholds and gates solvent outputs by predicted step presence. This separates learning candidate scores from deciding which candidates to return.

Input representation is equally important. We use interpretable descriptors of properties such as lipophilicity and polarity, which are relevant to SPE selection, but their strongly skewed distributions remain heavy-tailed after standardisation. SPEGen therefore applies a training-fitted rank-Gauss transform before the shared encoder. Its bagged variant averages predictions from independently resampled training sets before tuning the decoder. These choices address input irregularity and sensitivity to the training sample while preserving the descriptor information available to the models.

Our contributions are threefold. First, we formulate SPE component recommendation as a pollutant-level, multi-task learning problem with set-valued and partially observed targets. Second, we develop SPEGen, combining shared representation learning, task-specific objectives, rank-based normalisation and validation-tuned decoding. Third, we evaluate the framework against tabular baselines and analyse its behaviour through cumulative ablations and set-size diagnostics. The experiments demonstrate improved cartridge and solvent set recovery and lower ratio error with bootstrap aggregation, and clarify the trade-offs between broad recovery, exact match and top-ranked selection.

2 Related Work

2.1 SPE method development and analytical machine learning

SPE reviews describe sorbent selection and the roles of polarity, hydrophobicity and ionisation (Hennion, 1999; Poole, 2003; Fontanals et al., 2007; Andrade-Eiroa et al., 2016; Buszewski & Szultka, 2012). Quantitative structure–property models (Cherkasov et al., 2014) and learned molecular representations (Wu et al., 2018; Yang et al., 2019) provide complementary tools, including retention-time prediction (Domingo-Almenara et al., 2019; Bouwmeester et al., 2019). SPE component recommendation extends this setting to heterogeneous outputs. We use a common descriptor representation to study multi-task prediction, input normalisation and decoding.

2.2 Structured prediction of laboratory procedures

Reaction-condition prediction (Gao et al., 2018), retrosynthetic planning (Segler et al., 2018) and robotic synthesis (Coley et al., 2019) connect predictions to experimental decisions. Vaucher et al. (2020; 2021) represent procedures as machine-readable actions. The Open Reaction Database provides a schema and infrastructure for sharing reaction data (Kearnes et al., 2021). SPE component prediction combines heterogeneous targets aggregated across published methods for each pollutant. This formulation provides a sample-preparation counterpart to work on reaction conditions and laboratory procedures.

2.3 Multi-label and tabular learning

Classifier chains model label dependence through conditioning on preceding labels (Read et al., 2011); multi-task encoders share representations across related outputs (Caruana, 1997). Focal and asymmetric losses address label imbalance (Lin et al., 2017; Ridnik et al., 2021; Zhang & Zhou, 2014), and validation-based threshold selection is important for F1 (Lipton et al., 2014). Strong tree baselines remain essential on tabular inputs (Grinsztajn et al., 2022); neural alternatives include numerical embeddings and attention-based encoders (Gorishniy et al., 2022; 2021). TabPFN uses in-context prediction (Hollmann et al., 2023), and TabR retrieves training examples during prediction (Gorishniy et al., 2024). In SPEGen, descriptor neighbourhoods regularise shared representations during training.

Ensembling offers a complementary approach: independently initialised networks vary their starting parameters (Lakshminarayanan et al., 2017), and bootstrap aggregation also varies the sampled training data (Breiman, 1996).

3 PRELIMINARY

The pollutant-level target collects four types of reported SPE components. For pollutant i, let $\mathbf{x}_i \in \mathbb{R}^{21}$ denote the molecular descriptors and define

$$\mathbf{y}_i = (\mathbf{c}_i, \mathbf{m}_i, \mathbf{S}_i, \mathbf{R}_i), \quad \mathbf{c}_i \in \{0,1\}^{42},\ \mathbf{m}_i \in \{0,1\}^{5},\ \mathbf{S}_i \in \{0,1\}^{5\times 378},\ \mathbf{R}_i \in [0,1]^{5\times 378}. \quad (1)$$

Here $\mathbf{c}_i$ indicates reported cartridge classes; $\mathbf{m}_i$ indexes loading, conditioning, washing, elution and reconstitution; $\mathbf{S}_i$ is a step-by-solvent binary indicator matrix; and $\mathbf{R}_i$ stores ratio summaries. The five entries use a fixed indexing convention for workflow-step categories. The molecular table contributes pollutant identifiers, canonical SMILES, 21 PubChem-derived descriptors (Kim et al., 2023) and 42 cartridge indicators. A second table contributes 36,776 method records, including step-wise solvent strings and 26,204 records with at least one parsed ratio.

We aggregate methods linked to the same pollutant to retain the reported alternatives in component-wise supervision. This representation pools the marginal choices across source methods, while method-specific joint assignments remain outside the target. Let $\mathcal{M}_i$ be the linked methods, $\mathcal{S}(\mu, t)$ the solvent tokens reported by method μ at step t, and $\mathcal{M}_{i,t,k}$ the methods that report a ratio for step t and solvent k. For annotated pollutants, the aggregate labels are

$$\begin{aligned} m_{i,t} &= \max_{\mu\in\mathcal{M}_i} \mathbb{1}[|\mathcal{S}(\mu,t)| > 0], \\ s_{i,t,k} &= \max_{\mu\in\mathcal{M}_i} \mathbb{1}[k \in \mathcal{S}(\mu,t)], \\ r_{i,t,k} &= \frac{1}{|\mathcal{M}_{i,t,k}|} \sum_{\mu\in\mathcal{M}_{i,t,k}} r_{\mu,t,k}. \end{aligned} \quad (2)$$

The ratio mean is defined for $|\mathcal{M}_{i,t,k}| > 0$, and unobserved cells are masked during training. The unions collect marginal alternatives; averaging conflicting ratios can yield a value between the source observations. Appendix F discusses the implications for selecting compatible components. Figure 1(a) illustrates elution alternatives for one pollutant; Figure 1(b) shows the resulting supervision tensors. The complete five-step example and solvent-label distribution appear in Appendix A. Aggregation produces 11,441 pollutants; 10,335 have protocol annotations and 1,106 have cartridge labels only. A protocol-availability indicator a_i and an observed-ratio mask $o_{i,t,k}$ restrict direct task supervision to available annotations. The split and loss reductions are specified in Section 5.1 and Appendix C.

4 METHODOLOGY

Figure 2 presents SPEGen. Molecular descriptors pass through a training-fitted transform, a shared encoder and task-specific towers. Aggregated labels and ratio masks provide task-specific training supervision. Classification scores and conditional ratio estimates are decoded into candidate components.

4.1 RANK-GAUSS INPUT

For descriptor j, let $\ell_j(x)$ and $u_j(x)$ count training values strictly below and at most x, respectively. The transform is

$$\begin{aligned} p_j(x) &= \mathrm{clip}\left(\frac{\ell_j(x) + u_j(x)}{2(n_{\mathrm{tr}}+1)}, 10^{-6}, 1-10^{-6}\right), \\ v_{ij} &= \Phi^{-1}(p_j(x_{ij})), \qquad \widetilde{x}_{ij} = \frac{v_{ij} - \mu_j^{\mathrm{tr}}}{\max(\sigma_j^{\mathrm{tr}}, 10^{-6})}. \end{aligned} \quad (3)$$

Here Φ is the standard normal CDF; the final mean and population standard deviation are fitted on transformed training values. Left/right insertion counts assign ranks to tied values and to values unseen during training. Values beyond the training extrema use the same count rule and clipping. The map preserves the 21 input dimensions; exact feature order and boundary behaviour are specified in Appendix C.

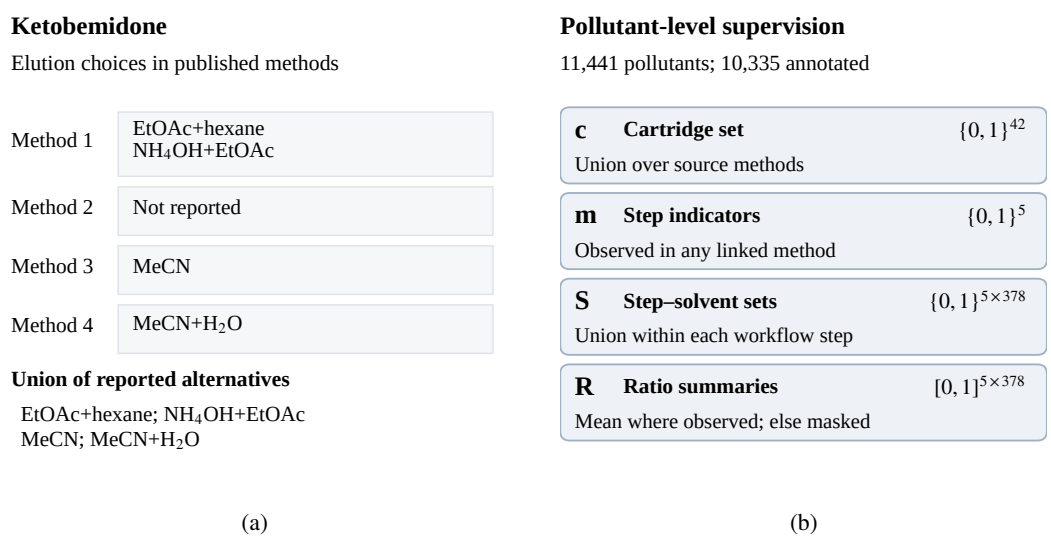


Figure 1: From published methods to set-valued supervision. (a) Elution alternatives in four literature methods for ketobemidone. Blank entries indicate missing reports; the full step-wise example is in Figure A2. (b) Aggregation yields cartridge, step, solvent and partially observed ratio targets.

4.2 Shared Encoder and Task Heads

The shared MLP has two 512-unit layers with ReLU and dropout 0.3 (Srivastava et al., 2014). Four task-specific MLP towers T_a, each with two 512-unit layers, map the shared representation to features for the corresponding output heads. Cartridge and step heads are linear; a step-guided logit term couples step and solvent predictions:

$$\begin{aligned} \mathbf{z}_i &= f_\theta(\widetilde{\mathbf{x}}_i), \qquad \mathbf{h}_i^a = T_a(\mathbf{z}_i), \quad a \in \{c, m, s, r\}, \\ \mathbf{u}_i^c &= W_c \mathbf{h}_i^c + \mathbf{b}_c, \qquad \mathbf{u}_i^m = W_m \mathbf{h}_i^m + \mathbf{b}_m, \\ \mathbf{U}_i^s &= \mathrm{reshape}_{5\times 378}(W_s \mathbf{h}_i^s + \mathbf{b}_s) + \alpha \mathbf{u}_i^m \mathbf{1}_{378}^\top, \\ \mathbf{q}_{i,t,k} &= \mathrm{softmax}\left(g([\mathbf{h}_i^r; \mathbf{e}_t; \mathbf{f}_k])\right), \qquad \widehat{r}_{i,t,k} = \frac{\arg\max_{b\in\{0,\dots,19\}} q_{i,t,k,b} + 0.5}{20}. \end{aligned} \tag{4}$$

The learned scalar α starts at 1. Step and solvent embeddings have widths 32 and 64; the ratio MLP is $608 \to 256 \to 20$. Ratio training uses adjacent-bin-smoothed cross-entropy, and single-model inference returns the centre of the highest-probability bucket. Each estimate describes one step–solvent pair; method-level grouping and mixture compatibility are discussed in Appendix F.

4.3 Objective and Decoding

The active objective combines masked task losses, two ranking terms and non-negative neighbourhood regularisation:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_c + \mathcal{L}_m + \lambda_s \mathcal{L}_s + \mathcal{L}_r + \lambda_{rc} \mathcal{L}_{rc} + 0.05 \mathcal{L}_{rs} + 0.01 \mathcal{L}_{\mathrm{sm}} \\ + \mathbb{1}_{\mathrm{EM}}\, 0.3 (\mathcal{L}_{\mathrm{count,c}} + \mathcal{L}_{\mathrm{count,s}}). \end{aligned} \tag{5}$$

Cartridge classification uses positive-weighted focal loss (Lin et al., 2017); steps use weighted binary cross-entropy and solvents use asymmetric loss (Ridnik et al., 2021). Protocol-availability and observed-ratio masks select the samples and pairs contributing to each task loss. Ranking compares positives with the full label vocabulary; neighbourhood weights are clipped cosine similarities. All reductions, masks and class weights are given in Appendix C.

For threshold decoding, validation selects per-label thresholds and task-wide shifts; writing their clipped sum as τ gives

$$\begin{aligned} \widehat{c}_j &= \mathbb{1}[\sigma(u_j^c) \geq \tau_j^c], \qquad \widehat{m}_t = \mathbb{1}[\sigma(u_t^m) \geq \tau_t^m], \\ \widehat{s}_{t,k} &= \mathbb{1}[\sigma(u_{t,k}^s) \geq \tau_{t,k}^s]\, \widehat{m}_t. \end{aligned} \tag{6}$$

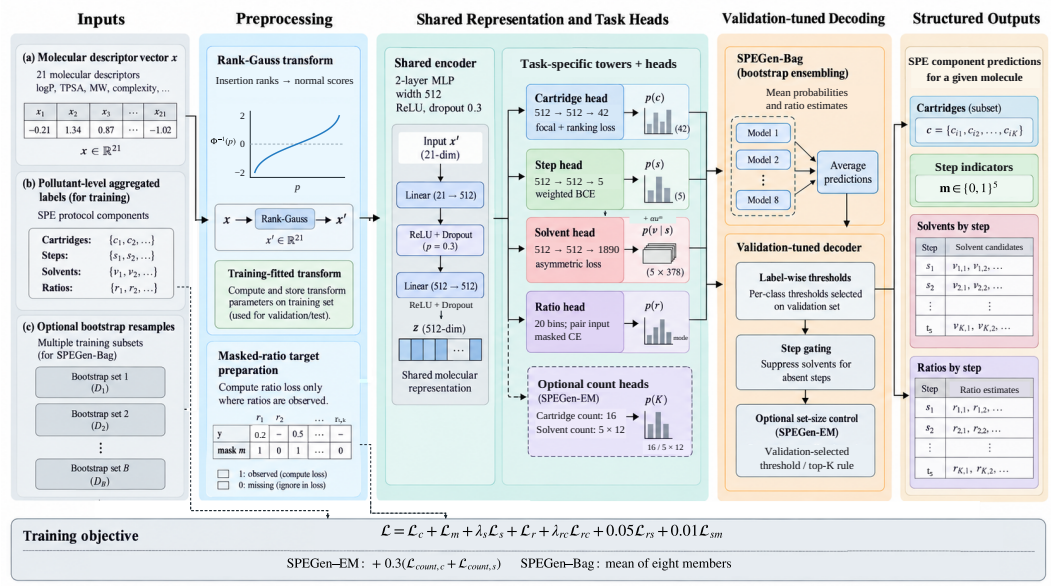


Figure 2: Overview of SPEGen. Rank-Gauss-normalised descriptors pass through a shared encoder and four task-specific towers and heads. Step logits guide solvent scores; observed step–solvent pairs supervise ratio prediction. Validation-tuned thresholds and step gating produce component sets and ratio estimates. SPEGen-EM adds count prediction, and SPEGen-Bag averages eight members' classification probabilities and ratio point estimates.

All SPEGen decoders use step gating to retain solvent predictions only for predicted-present steps. Within these steps, the selected solvent set may be empty. SPEGen-Bal uses $(\lambda_s, \lambda_{rc}) = (1, 0.5)$ and SPEGen-Chain uses $(0.1, 2)$, both with F1-oriented decoding. SPEGen-EM extends Bal with cartridge and per-step solvent count classifiers and searches threshold, top-k and threshold-plus-top-k rules under an EM/Jaccard objective. SPEGen-Bag averages eight bootstrap-trained Chain members (Breiman, 1996); classification probabilities are averaged before validation calibration, whereas ratios are averages of member modal-bucket point estimates. Appendix C specifies checkpoint selection and the complete decoder search.

5 Experiments

5.1 Experiment Setting

All models use the same 21 descriptors and fixed 9,152/1,144/1,145 training/validation/test split. Comparators are decision tree, random forest, SVM, MLP, XGBoost and LightGBM (Breiman et al., 1984; Breiman, 2001; Chen & Guestrin, 2016; Ke et al., 2017; Pedregosa et al., 2011). We report cartridge F1 (C-F1), exact match (C-EM), solvent F1 (S-F1), ratio error, cartridge top-1 (C-top1), component-triplet accuracy (Triplet) and strict ratio-verified Hit@1. Triplet and Hit@1 use the 1,043 protocol-annotated test pollutants; missing selected ratios are failures in Hit@1. SSS, HC and GUS are defined in Appendix D; HC measures agreement between predicted solvent occupancy and annotated step presence. Single networks report mean ± sample standard deviation ($n = 4$, ddof=1); baselines and Bag have one evaluation. Column-best cells are bold; only MAE is minimised. Baseline thresholds are fixed at 0.5; SPEGen thresholds are selected on validation data (Appendix G).

5.2 Chemical Signal

We first examine chemical associations in the descriptor space. Figure 3(a) shows that lipophilic, low-TPSA pollutants favour silica, whereas more polar regions favour mixed-mode ion exchangers and broader reversed-phase materials. The same descriptors also correlate with solvent choice: methanol becomes less common in the most lipophilic bins, while ethyl acetate and dichloromethane

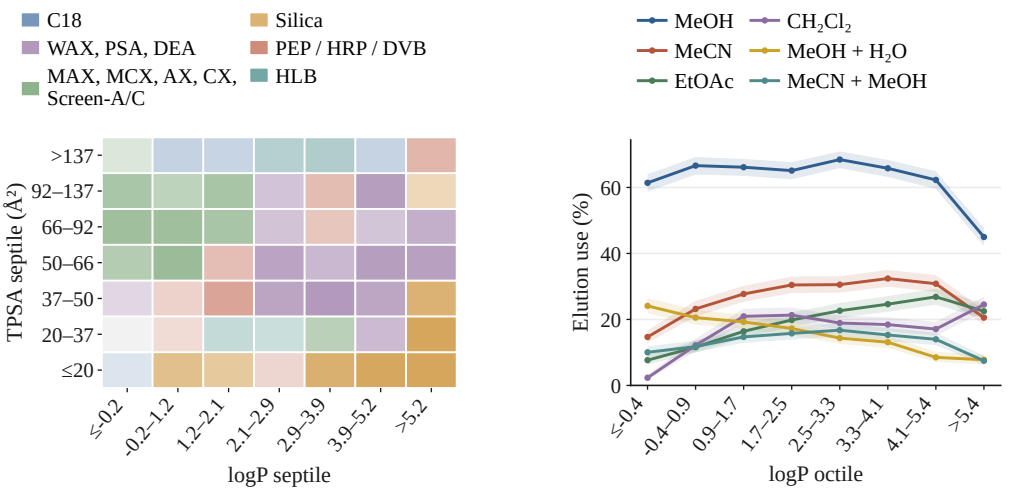


Figure 3: Chemical structure in the supervision. (a) Locally dominant cartridge classes across logP–TPSA quantile bins, with opacity proportional to enrichment over the corpus prior. (b) Prevalence of leading elution solvents across logP octiles with Wilson 95% confidence intervals. MeOH, methanol; MeCN, acetonitrile; EtOAc, ethyl acetate; CH_2Cl_2, dichloromethane.

Table 1: Per-task performance on the held-out test split.

Model	C-F1	C-EM	S-F1	Ratio MAE	Ratio≤0.10
Decision tree	0.4020	0.0786	0.3933	0.1740	0.5542
Random forest	0.4030	0.1293	0.4497	0.1524	0.4898
SVM	0.0587	0.0236	0.0069	0.2174	0.3734
MLP	0.4171	0.1092	0.4567	0.1611	0.4904
XGBoost	0.4115	0.1188	0.4563	0.1850	0.3915
LightGBM	0.4181	0.1153	0.4509	0.1764	0.4115
SPEGen-Bal	0.4799 ± 0.0089	0.0155 ± 0.0021	0.4660 ± 0.0009	0.1581 ± 0.0032	0.6373 ± 0.0057
SPEGen-Chain	0.4781 ± 0.0057	0.0179 ± 0.0023	0.4538 ± 0.0030	0.1570 ± 0.0011	$\mathbf{0.6395 \pm 0.0030}$
SPEGen-EM	0.3490 ± 0.0166	$\mathbf{0.1443 \pm 0.0096}$	0.4716 ± 0.0062	0.1594 ± 0.0022	0.6341 ± 0.0038
SPEGen-Bag	**0.4909**	0.0384	**0.4747**	**0.1464**	0.5916

become more common. These descriptive analyses characterise the full corpus; model selection uses the held-out validation split. Additional enrichment maps and cartridge–solvent associations appear in Appendix B.

5.3 Main Result

Tables 1 and 2 compare four operating points of SPEGen with six tabular baselines under the same split and descriptor input. The variants differ in loss weights, decoding objectives and the optional count heads; SPEGen-Bag ensembles eight SPEGen-Chain members (Appendix Table A2).

On per-task prediction, SPEGen-Bag improves cartridge F1 from 0.4181 for LightGBM to 0.4909 and solvent F1 from 0.4567 for the MLP to 0.4747. Its ratio MAE is 0.1464, compared with 0.1524 for the strongest baseline. SPEGen-Chain gives the highest ratio-within-0.10 rate, 0.6395 versus 0.5542 for the decision tree. SPEGen-EM raises cartridge exact match to 0.1443, above the random forest's 0.1293, but its smaller predicted sets reduce cartridge F1 to 0.3490.

On structured metrics, SPEGen-Bag gives the highest cartridge top-1 (0.6140 versus 0.5965 for XGBoost), but random forest leads Triplet (0.5168 versus 0.4995). Under strict ratio verification, MLP leads Hit@1 at 0.1016, followed by SPEGen-Chain at 0.0952 ± 0.0073; Bag obtains 0.0872. For Bag, 521 triplets match the recorded components. Of these, 379 lack an annotated ratio for

Table 2: Structured-output scores with strict ratio-verified Hit@1.

Model	C-top1	Triplet	Hit@1	SSS	HC	GUS
Decision tree	0.4760	0.2790	0.0518	0.4660	0.8855	55.31
Random forest	0.5939	**0.5168**	0.0594	0.4848	0.8023	56.97
SVM	0.4699	0.3471	0.0125	0.2573	0.9041	31.05
MLP	0.5642	0.4506	**0.1016**	0.4666	0.8309	57.39
XGBoost	0.5965	0.5091	0.0537	0.4695	0.8084	55.96
LightGBM	0.5921	0.4746	0.0738	0.4680	0.8242	56.20
SPEGen-Bal	0.5775 ± 0.0056	0.4859 ± 0.0039	0.0865 ± 0.0021	0.4810 ± 0.0009	0.9072 ± 0.0026	$\mathbf{61.66 \pm 0.17}$
SPEGen-Chain	0.5948 ± 0.0056	0.4952 ± 0.0057	0.0952 ± 0.0073	0.4750 ± 0.0031	$\mathbf{0.9083 \pm 0.0019}$	61.22 ± 0.26
SPEGen-EM	0.5753 ± 0.0037	0.4705 ± 0.0064	0.0887 ± 0.0025	$\mathbf{0.4914 \pm 0.0013}$	0.8989 ± 0.0069	58.52 ± 0.47
SPEGen-Bag	**0.6140**	0.4995	0.0872	0.4751	0.9070	61.55

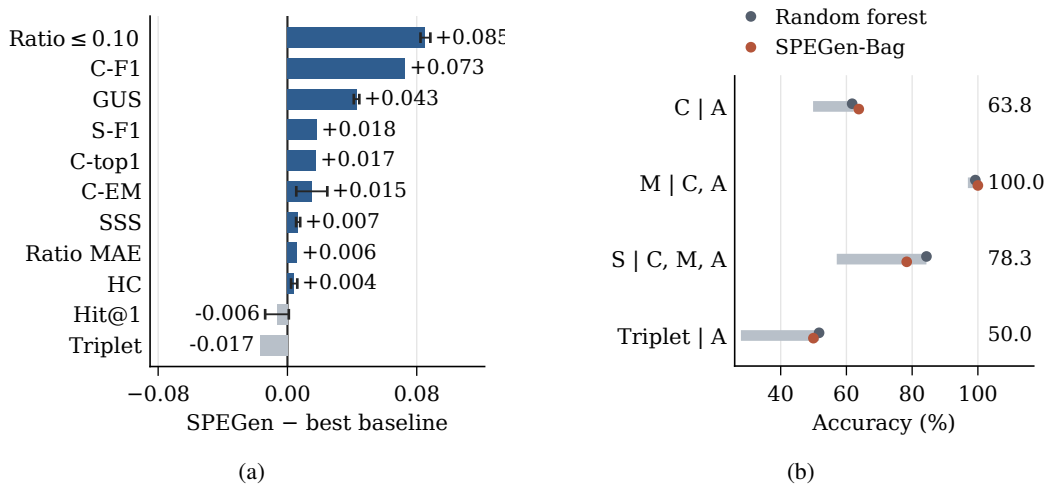


Figure 4: Held-out comparison with tabular baselines. (a) Margin between the best SPEGen variant and the strongest baseline for each metric; error bars show sample standard deviation for single networks; Hit@1 requires observed-ratio verification. GUS differences are divided by 100. (b) Component-triplet decomposition on 1,043 annotated pollutants: grey spans baselines, dark grey marks random forest, and orange marks SPEGen-Bag. C and M denote correct cartridge and step selections.

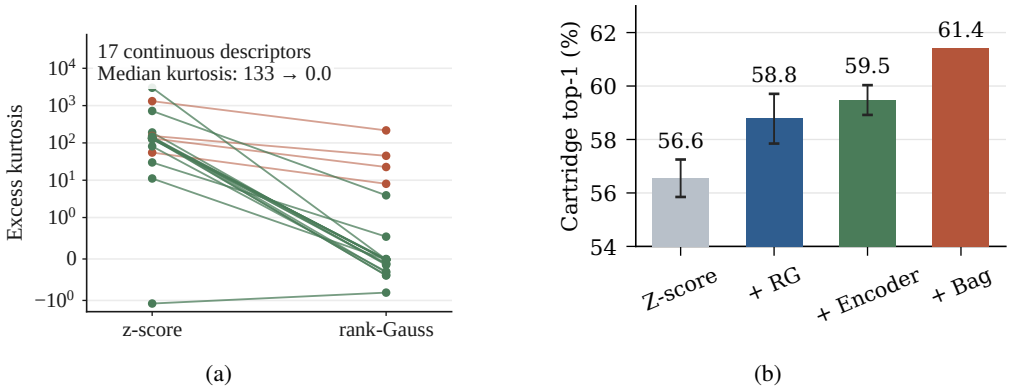
the selected pair and 51 exceed the error tolerance, leaving 91 verified hits out of 1,043 annotated pollutants. The strict score therefore reflects both component selection and the availability of ratio evidence.

Different operating points lead different metrics: Bal leads GUS, Chain leads HC, and EM leads cartridge exact match and SSS. The four variants collectively lead nine of eleven metrics, with random forest leading Triplet and MLP leading strict Hit@1. These column-wise maxima summarise complementary operating points. Single-network Chain reaches C-top1 0.5948, compared with XGBoost's 0.5965; the eight-member Bag ensemble reaches 0.6140. The comparison uses configuration-specific calibration and computational budgets, as detailed in Appendix G.

Figure 4(b) uses the same annotated population for all factors. Bag's cartridge prefix is $665/1043 = 0.6376$; step-given-cartridge is 1.000 and solvent-given-prefix is $521/665 = 0.7835$, whose product is 0.4995. Random forest is stronger on the last factor (0.8435). Thus solvent ranking remains a weakness even when cartridge ranking improves. The triplet analysis measures membership in the pollutant-level aggregate labels.

5.4 Ablation

Table 3 adds rank-Gauss normalisation, encoder tuning and bootstrap aggregation sequentially; n gives the number of repeated evaluations of each configuration. Rank-Gauss raises cartridge top-1



(a) (b)

Figure 5: Cumulative ablation. (a) Descriptor excess kurtosis under z-scoring and rank-Gauss normalisation; four near-constant counters remain resistant to a monotone transform. (b) Cumulative cartridge top-1 accuracy after rank-Gauss (RG) normalisation, encoder tuning and bootstrap bagging. Error bars apply to the four-seed single-network stages; the ensemble has one reported evaluation.

Table 3: Stage-wise ablation of SPEGen.

Configuration	C-top1	Triplet	Hit@1	C-F1	S-F1	GUS	n
Z-score, default encoder	0.5655 ± 0.0070	0.4626 ± 0.0106	0.0769 ± 0.0053	0.4574 ± 0.0074	0.4356 ± 0.0035	59.91 ± 0.29	4
+ rank-Gauss	0.5878 ± 0.0093	0.4851 ± 0.0095	0.0901 ± 0.0113	0.4658 ± 0.0062	0.4420 ± 0.0016	60.42 ± 0.17	4
+ tuned encoder	0.5948 ± 0.0056	0.4952 ± 0.0057	$\mathbf{0.0952 \pm 0.0073}$	0.4781 ± 0.0057	0.4538 ± 0.0030	61.22 ± 0.26	4
+ bootstrap bagging	**0.6140**	**0.4995**	0.0872	**0.4909**	**0.4747**	**61.55**	1

from 0.5655 ± 0.0070 to 0.5878 ± 0.0093, while also improving the other reported metrics; strict Hit@1 rises from 0.0769 to 0.0901. Figure 5(a) shows the accompanying distributional change: under z-scoring, many descriptor dimensions remain strongly heavy-tailed, whereas rank-Gauss reduces the median excess kurtosis of the 17 continuous dimensions from 133 to approximately zero.

Encoder tuning adds a further 0.0070 in cartridge top-1, and eight-member bootstrap aggregation adds 0.0192 more. Bagging also produces the best cartridge F1 and solvent F1, although strict Hit@1 changes from 0.0952 to 0.0872, illustrating a trade-off between broad set recovery and ratio-verified top-1 selection. Each increment measures the change relative to the preceding configuration in this development sequence. Additional single-run comparisons against configuration-specific controls, including numerical embeddings and uncertainty weighting (Gorishniy et al., 2022; Kendall et al., 2018), are summarised in Appendix E.

5.5 Decoding Analysis

The SPEGen variants exhibit a precision–recall trade-off. F1-oriented decoding predicts 6.2–6.5 cartridge labels against 3.3 reported labels, obtaining recall near 0.69 at precision near 0.38. Exact-match decoding predicts 1.4 labels, raises precision to 0.58 and obtains the best cartridge exact match in the study (0.1443), but reduces cartridge F1 to 0.3490.

In a separate diagnostic run, a fixed-ranking oracle-size intervention raises cartridge F1 from 0.4805 to 0.6532 and exact match from 0.0183 to 0.2498; solvent F1 rises from 0.4564 to 0.6470 (Appendix Table A5). A dedicated gradient-boosted count predictor obtains test R^2 values of 0.09 and 0.05 and Spearman correlations of 0.31 and 0.32 for cartridge and solvent counts, respectively. This predictor explains a small fraction of the observed count variation, leaving substantial scope for improved count modelling or additional context. The precision–recall and cardinality plots are in Figure A9; fixed-ranking oracle conventions are detailed in Appendix E.

5.6 Interpretability

A depth-three CART model (Breiman et al., 1984) on 2,924 single-cartridge pollutants reaches held-out balanced accuracy 0.31, against chance 0.17. Figure 6 shows silica enrichment of $5.2\times$ for

complexity above 374 and logP above 6.75. The solvent tree reaches 0.30 (Appendix Figure A7). Both trees are fitted directly to single-label subsets and describe descriptor–label associations. Balanced accuracy evaluates these restricted multiclass tasks; the main comparison evaluates multi-label recovery over the full target space.

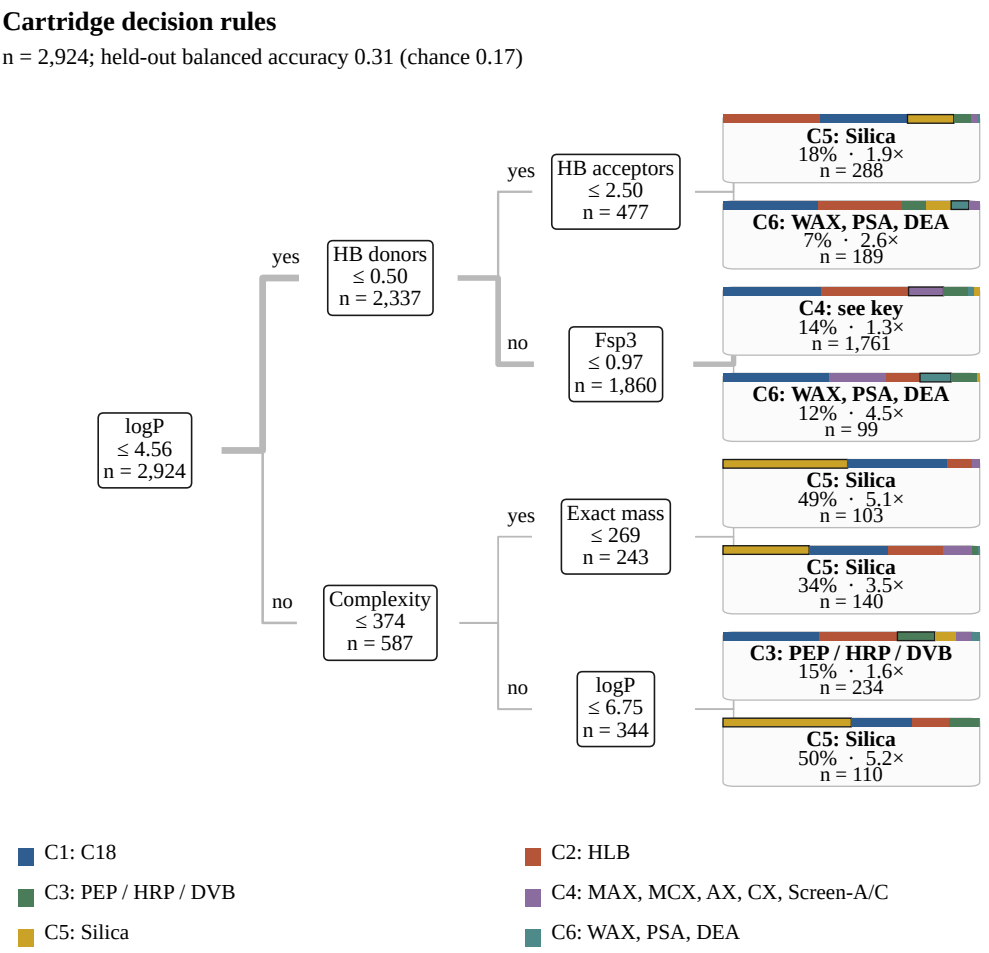


Figure 6: Descriptor-level cartridge rules on the single-cartridge subset. Leaves show the predicted class, observed share, enrichment and sample count; stacked bars show class composition, with the predicted class outlined. Edge width reflects sample flow. Class-balanced fitting determines the prediction; segment widths represent unweighted counts.

6 Conclusion

SPEGen learns SPE component recommendations from pollutant-level aggregation of published methods. On the fixed held-out split, its bagged variant improves cartridge and solvent set recovery and achieves the lowest ratio MAE among the reported methods. Cumulative ablations associate gains with rank-Gauss normalisation and bootstrap aggregation, while decoding analyses expose the trade-off between broad recovery and exact match. Descriptor-level trees provide complementary evidence of chemical associations in the corpus. The outputs remain candidate components: incomplete reporting and loss of source-method co-occurrence require further compatibility checks and prospective laboratory validation.

AI USE STATEMENT

Generative AI tools assisted the design and refinement of the research methodology and experiments in this work. We also used AI assistance to create and revise scientific figures and their plotting code, and to edit the manuscript for clarity and readability. The authors are responsible for evaluating AI suggestions, verifying scientific content and ensuring the accuracy and originality of all text, code and visual material in the submission.

ETHICS STATEMENT

We acknowledge the ICLR Code of Ethics and the responsibility to conduct and communicate this research transparently. This retrospective computational study uses literature-derived SPE records and molecular descriptors to support analytical method development. Uneven literature coverage and positive-only reporting can bias the recommendations. The supervision represents documented choices, with unreported outcomes remaining unknown. Pollutant-level aggregation yields candidate components whose joint compatibility requires method-level assessment. Any experimental application requires expert assessment of chemical compatibility, solvent hazards and waste handling, followed by laboratory validation. Use and redistribution of the source records and derived resources must respect the original sources' rights and access conditions. The implications of aggregation and the limits of generalisation are discussed in Appendices F and G.

REPRODUCIBILITY STATEMENT

All source code for data processing, model construction and training is included in the supplementary attachment, together with the decoding, evaluation, baseline and figure-generation routines. Sections 3 and 4 describe the task, model and decoder, and Section 5.1 specifies the split and experimental protocol. Appendix A documents target construction and missing annotations; Appendix C and Table A2 describe the objectives and training settings; Appendix D discusses evaluation conventions; and Appendix E presents additional diagnostics. The attached experiment scripts, saved configurations and outputs, and repository documentation provide the implementation record and figure-rebuild instructions. The supplementary archive supports source inspection and regeneration of the reported aggregates. Independent prediction-level replay requires trained checkpoints and full per-sample arrays, which are absent from the archive. Appendix G describes the implementation records and reproduction scope.

REFERENCES

Aurora Andrade-Eiroa, Moisès Canle, Valérie Leroy-Cancellieri, and Víctor Cerdà. Solid-phase extraction of organic compounds: A critical review (part i). *TrAC Trends in Analytical Chemistry*, 80:641–654, 2016.

Robbin Bouwmeester, Lennart Martens, and Sven Degroeve. Comprehensive and empirical evaluation of machine learning algorithms for small molecule LC retention time prediction. *Analytical Chemistry*, 91(5):3694–3703, 2019.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Bogusław Buszewski and Małgorzata Szultka. Past, present, and future of solid phase extraction: A review. *Critical Reviews in Analytical Chemistry*, 42(3):198–213, 2012.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Vladimir E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition, 1988.

Connor W. Coley, Dale A. Thomas, Justin A. M. Lummiss, Jonathan N. Jaworski, Christopher P. Breen, Victor Schultz, Thomas Hart, Joshua S. Fishman, Luke Rogers, Hua Gao, R. William Hicklin, Pascal P. Plehiers, Joshua Byington, John S. Piotti, William H. Green, A. John Hart, Timothy F. Jamison, and Klavs F. Jensen. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science*, 365(6453):eaax1566, 2019.

Xavier Domingo-Almenara, Carlos Guijas, Elizabeth Billings, J. Rafael Montenegro-Burke, Worawan Uritboonthai, Alex E. Aisporna, Emma Chen, H. Paul Benton, and Gary Siuzdak. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications*, 10:5811, 2019.

Núria Fontanals, Rosa M. Marcé, and Francesc Borrull. New materials in sorptive extraction techniques for polar compounds. *Journal of Chromatography A*, 1152(1–2):14–31, 2007.

Hua Gao, Thomas J. Struble, Connor W. Coley, Yuran Wang, William H. Green, and Klavs F. Jensen. Using machine learning to predict suitable conditions for organic reactions. *ACS Central Science*, 4(11):1465–1476, 2018.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18932–18943, 2021.

Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24991–25004, 2022.

Yury Gorishniy, Ivan Rubachev, Nikolay Kartashev, Daniil Shlenskii, Akim Kotelnikov, and Artem Babenko. TabR: Tabular deep learning meets nearest neighbors. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ef594af0d9a519db8fb292452c461fa-Abstract-Conference.html.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pp. 507–520, 2022.

Marie-Claude Hennion. Solid-phase extraction: Method development, sorbents, and coupling with liquid chromatography. *Journal of Chromatography A*, 856(1–2):3–54, 1999.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2207.01848.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3146–3154, 2017.

Steven M. Kearnes et al. The Open Reaction Database. *Journal of the American Chemical Society*, 2021. doi: 10.1021/jacs.1c09820.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. Thresholding classifiers to maximize F1 score. In *Machine Learning and Knowledge Discovery in Databases*, pp. 225–239. Springer, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Colin F. Poole. New trends in solid-phase extraction. *TrAC Trends in Analytical Chemistry*, 22(6): 362–373, 2003.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.

Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610, 2018.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Alain C. Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H. Nair, Philippe Schwaller, and Teodoro Laino. Automated extraction of chemical synthesis actions from experimental procedures. *Nature Communications*, 11:3601, 2020.

Alain C. Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H. Nair, Antonella Iuliano, and Teodoro Laino. Inferring experimental procedures from text-based representations of chemical reactions. *Nature Communications*, 12:2573, 2021.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Adrian Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

APPENDIX

A CORPUS CONSTRUCTION AND SUPERVISION STATISTICS

Table A1 summarises the target construction, learning signals and decoding rules. The protocol parser processes five steps—loading, conditioning, washing, elution and reconstitution—and converts every non-empty cell into one or more normalised solvent tokens. A numeric ratio, when present, is associated with its step–solvent pair. Cells without a ratio still supervise solvent classification but are masked from ratio prediction. Pollutant-level aggregation takes unions over cartridge, step and solvent labels and averages repeated ratios, as defined in Equation 2.

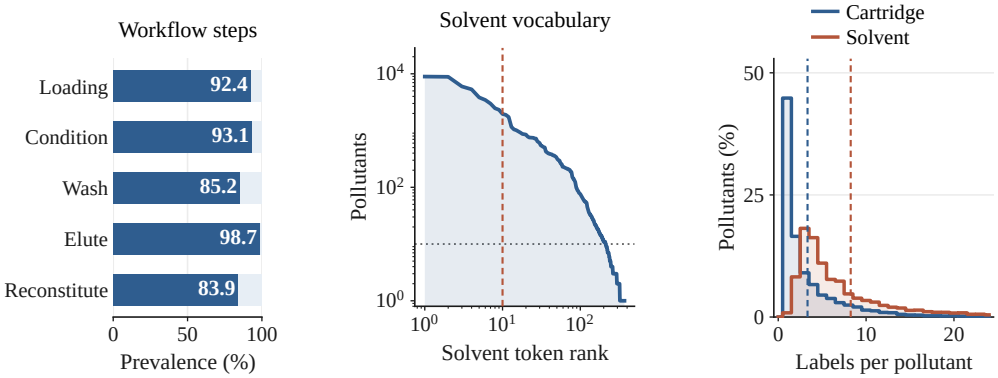


Figure A1: Shape of the supervision signal. (a) Workflow-step prevalence among annotated pollutants. (b) Solvent-token rank frequency. (c) Distributions of reported cartridge and solvent set size; dashed lines mark their means.

Figure A1 characterises step prevalence, solvent frequency and reported set sizes. Elution appears in 98.7% of annotated pollutants and reconstitution in 83.9%, so the step mask contains relatively little discriminative information. The solvent vocabulary is strongly long-tailed: the ten most frequent tokens account for 54% of all positive uses and 157 of 378 tokens occur fewer than ten times. Cartridge and solvent set sizes are right-skewed, with means 3.34 and 8.25. Numeric ratios are observed for 39.7% of positive step–solvent cells. Descriptor kurtosis before and after rank-Gauss normalisation is shown in Figure 5(a).

Table A1: Structured label space, learning signals and decoding

Component	Symbol	Size	Construction	Learning and decoding
Molecular descriptors	$\mathbf{x}_i$	21	PubChem descriptors from Complexity to logP; rank-Gauss normalised on the training split	Shared encoder input
Cartridge set	$\mathbf{c}_i$	42	Union of cartridge classes reported for the pollutant; 3.34 labels on average	Positive-weighted focal loss plus ranking; per-label thresholds
Step mask	$\mathbf{m}_i$	5	Union over loading, condition, wash, elute and reconstitute	Masked weighted BCE; per-step thresholds
Step-wise solvents	$\mathbf{S}_i$	5×378	Union of normalised tokens conditioned on the step; 8.25 distinct solvents per protocol	Asymmetric loss; per-label thresholds gated by steps
Mixture ratios	$\mathbf{R}_i$	5×378	Values in $[0, 1]$; mean of reported ratios; missing cells are masked	Masked 20-bucket cross-entropy; modal bucket centre
Corpus	–	11,441	10,335 pollutants with protocol annotation and 1,106 with cartridge labels only	–

B CHEMICAL ANALYSES

For each cartridge class, prevalence within a descriptor quantile bin is divided by prevalence in the full corpus and reported on a base-2 logarithmic scale. Figure A3 shows the one-dimensional enrichment patterns; Figure 3(a) gives the joint logP–TPSA map. Silica is enriched 2.6-fold in the highest

Published methods for Ketobemidone

4 independent literature protocols, one pollutant

	Load	Condition	Wash	Elute	Reconstitute
method 1	MeOH+phosphate	MeOH H_2O phosphate	phosphate MeOH	EtOAc+hexane NH_4OH+EtOAc	MeCN+HCOOH
method 2	K-phosphate	—	—	—	NH_4HCO_2+MeOH
method 3	NH_4OAc+MeOH	MeOH H_2O	MeOH+H_2O	MeCN	NH_4OAc+MeOH
method 4	HCOOH+H_2O	MeOH H_2O	H_2O MeOH+H_2O MeOH	MeCN+H_2O	MeOH+H_2O
	▼ Union over methods; ratios averaged				
target	MeOH+phosphate K-phosphate NH_4OAc+MeOH HCOOH+H_2O	MeOH H_2O phosphate	phosphate MeOH MeOH+H_2O H_2O	EtOAc+hexane NH_4OH+EtOAc MeCN MeCN+H_2O	MeCN+HCOOH NH_4HCO_2+MeOH NH_4OAc+MeOH MeOH+H_2O

Published alternatives are retained as component-wise targets; joint method assignments are not retained.

Figure A2: Complete step-wise example for ketobemidone. Each row retains the solvents reported in one source method. The bottom row takes a union separately within each step; blank entries indicate missing reports. Repeated ratios, where available, are averaged.

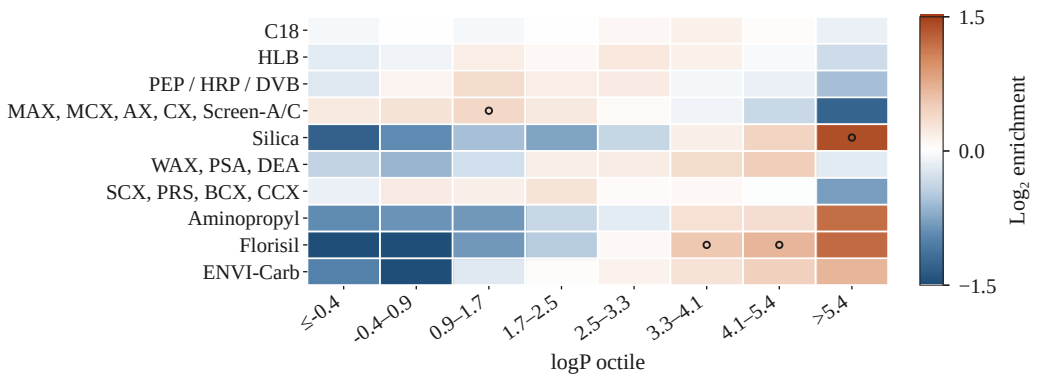
logP octile and depleted 2.4-fold in the lowest; Florisil and aminopropyl follow the same direction. Ion-exchange materials favour the polar, high-TPSA end, whereas C18 and HLB are relatively flat. Quantile bins contain equal numbers of molecules, cells with fewer than 40 observations are blank, and aggregate "Unknown" and "Null cartridge" classes are excluded.

Recorded solvent choice is associated with the workflow step, the analyte and the cartridge (Figures 3(b) and A4–A6). Conditioning is dominated by methanol and water, washing is predominantly aqueous and elution is predominantly organic. Across logP octiles, methanol decreases from roughly 66% in the middle of the range to 45% in the most lipophilic octile, while hexane rises from 2% to 30%, ethyl acetate from 8% to 23% and dichloromethane from 2% to 25%. Standardised mean differences give the same ordering: hexane-eluted pollutants are 0.8 pooled standard deviations higher in logP than the remainder, whereas water-eluted pollutants are 0.6 lower. The cartridge–solvent panel is centred by row because a pollutant may carry several cartridge labels, making raw conditional probabilities difficult to interpret.

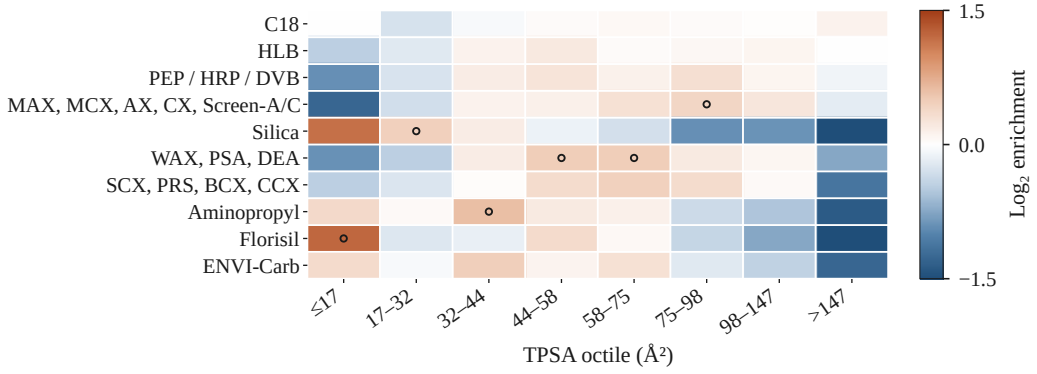
Depth-three CART models (Breiman et al., 1984) provide interpretable descriptor rules on unambiguous subsets: 2,924 pollutants with one cartridge and 2,868 protocols with one elution solvent. Held-out balanced accuracy is 0.31 for cartridge and 0.30 for solvent, compared with chance 0.17. Figures 6 and A7 show the cartridge and solvent trees, respectively. In the latter, complexity below 144 and monoisotopic weight below 150 select dichloromethane at 6.8 times its corpus rate. Together they show that the descriptor signal is chemically coherent but incomplete, motivating a model that can use descriptor interactions and correlations among outputs.

C Implementation and optimisation

Descriptor order and preprocessing. The input columns, in their stored order, are Complexity, CovalentUnitCount, DefinedAtomStereoCount, DefinedBondStereoCount, ExactMass, FormalCharge, HBondAcceptor, HBondDonor, HeavyAtomCount, MolecularWeight, MonoIsotopicWeight, RotatableBond, TPSA, UndefinedAtomStereoCount, UndefinedBondStereoCount, FCSP3, AtomCount, Kappa1, Kappa2, Kappa3 and LogP (each with the CSV suffix `_value`). All 11,441 rows in the



(a)



(b)

Figure A3: Cartridge selection across lipophilicity and polarity. (a,b) Enrichment across logP and TPSA octiles, expressed as $\log_2[P(\text{cartridge} \mid \text{bin})/P(\text{cartridge})]$. Open circles mark the most enriched class in a bin when enrichment exceeds $1.3\times$.

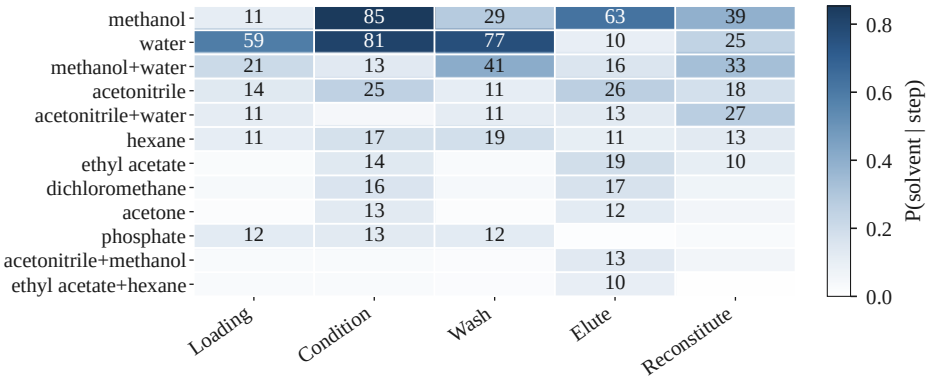


Figure A4: Solvent usage by workflow step, conditional on step presence. Cell values are percentages.

processed molecular table contain finite numeric values in these columns. The loader parses the values as float32, with imputation disabled.

For rank-Gauss preprocessing, the float32 descriptors are cast to float64, sorted using the training indices, and mapped by the left/right insertion counts in Equation 3. A tie occupies its count-interval midpoint; a new value between training observations receives the shared insertion count of that interval. Below the minimum, the probability is clipped to 10^{-6}; above the maximum, it is $n_{\text{tr}}/(n_{\text{tr}}+1)$ before clipping. The map uses $\sqrt{2}\,\text{erf}^{-1}(2p-1)$. Training means and population standard deviations of the mapped values are then fitted, with a 10^{-6} scale floor, and the final array is

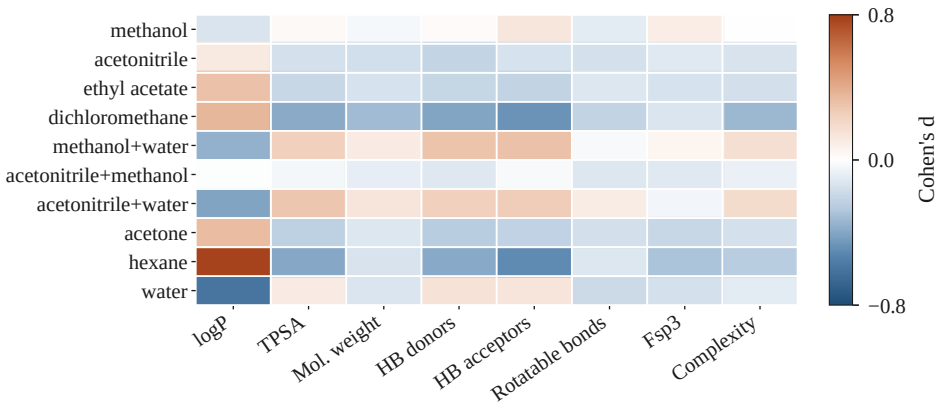


Figure A5: Descriptor associations with elution solvents. Values are standardised descriptor differences between solvent-associated and other pollutants.

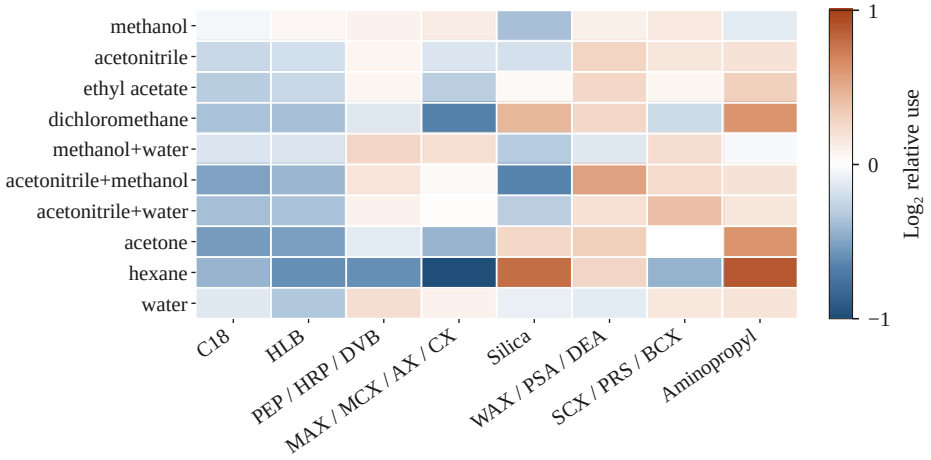


Figure A6: Cartridge associations with elution solvents. Cartridge preference is expressed as $\log_2[P(\text{solvent} \mid \text{cartridge})/\text{row mean}]$.

float32. The fitted training statistics are reused for validation and test data. For bagging, preprocessing is fitted on the original training partition before member indices are resampled. The implementation is `spe/data.py::apply_feature_transform`.

Architecture. The shared encoder consists of two repetitions of Linear–ReLU–Dropout with width 512 and dropout 0.3. Four independent task towers each apply two more 512-to-512 Linear–ReLU–Dropout blocks. Cartridge and step projections have dimensions $512 \to 42$ and $512 \to 5$. The solvent projection $512 \to 1890$ is reshaped to 5×378 and receives αu_t^m at every solvent position in step t. The scalar α is an unrestricted learned parameter initialised at 1; its separate $(\alpha - 1)^2$ penalty has weight zero. This additive connection guides solvent scores before the hard step gate is applied during decoding.

The conditional ratio head applies the ratio tower to the shared representation, concatenates the result with a 32-dimensional step embedding and a 64-dimensional solvent embedding, and applies Linear$(608, 256)$–ReLU–Dropout(0.3)–Linear$(256, 20)$. With bucket index $b = 0, \ldots, 19$, the target class is $b^* = \min(19, \max(0, \lfloor 20r \rfloor))$ for non-negative ratios; $r = 1$ belongs to bucket 19. Inference returns $(b^*_{\text{pred}} + 0.5)/20$ with $b^*_{\text{pred}} = \arg\max_b q_b$. The implementation also allocates a continuous-regression head, which remains inactive in bucket mode. All four variants use the descriptor-only input and bucketed ratio configuration summarised in Table A2.

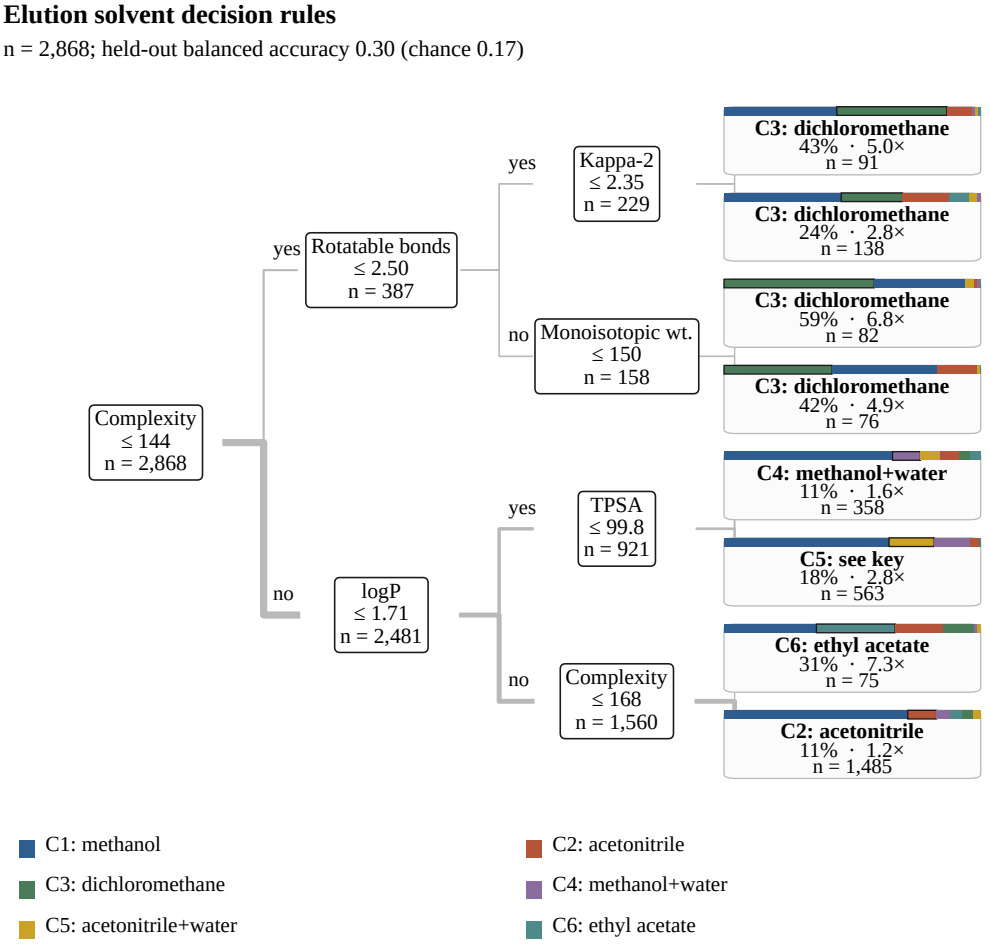


Figure A7: Descriptor-level decision rules for the elution solvent, using the same construction as Figure 6; C1–C6 identify the solvents in this figure's key.

Availability masks and task losses. Let $a_i \in \{0, 1\}$ indicate protocol availability and let $o_{i,t,k}$ mark a finite observed ratio. Cartridge targets exist for all samples. For cartridge and step label j, positive weights are $\omega_j = \mathrm{clip}((N_j^- + 10^{-6})/(N_j^+ + 10^{-6}), 1, 50)$. They are computed over all training rows, including zero-filled step labels on cartridge-only samples. These samples therefore affect step class-weight statistics even though their direct step loss is masked. Define $B_\omega(y, p) = -\omega y \log p - (1 - y) \log(1 - p)$ and $p_y = yp + (1 - y)(1 - p)$. For a batch $\mathcal{B}$,

$$\mathcal{L}_c = \frac{1}{42|\mathcal{B}|} \sum_{i \in \mathcal{B}, j} (1 - p_{y,ij})^{1.5} B_{\omega_j}(c_{ij}, p^c_{ij}), \qquad \mathcal{L}_m = \frac{\sum_{i,t} a_i B_{\omega_t}(m_{i,t}, p^m_{i,t})}{5 \sum_i a_i + 10^{-6}}. \tag{A1}$$

The cartridge loss preserves the positive weights when applying focal modulation (Lin et al., 2017). For solvents, the asymmetric element loss (Ridnik et al., 2021) and step-weighted reduction are

$$\ell_s(y, p) = -y \log p - (1 - y)(p^-)^4 \log(1 - p^-), \qquad p^- = \max(p - 0.05, 0),$$
$$w_{i,t} = a_i[2.2 m_{i,t} + (1 - m_{i,t})], \qquad \mathcal{L}_s = \frac{\sum_{i,t,k} w_{i,t} \ell_s(s_{i,t,k}, p^s_{i,t,k})}{378 \sum_{i,t} w_{i,t} + 10^{-6}}. \tag{A2}$$

Logarithm arguments are clipped below at 10^{-8}. The solvent objective uses the asymmetric element weights in Equation A2. Its reduction assigns weight 2.2 to present steps and weight 1 to absent steps on annotated samples; $a_i = 0$ removes direct solvent supervision for samples lacking protocol annotations.

Only finite ratio pairs on annotated samples contribute to $\mathcal{L}_r$. Their one-hot bucket targets are smoothed with 0.1 mass on adjacent bins: weight 0.9 on the target and 0.05 on each available

neighbour, followed by renormalisation at the boundaries. The loss averages cross-entropy over all such pairs in a batch and is zero when none are present. The observed-pair mask excludes missing ratio cells from this average.

Ranking and neighbourhood terms. The ranking denominators span all 42 cartridge labels and all 378 solvent labels within each eligible step. For positives P_i and vocabulary $\mathcal{V}$,

$$\ell_{\mathrm{rank}}(\mathbf{u}_i, P_i) = -\frac{1}{|P_i| + 10^{-6}} \sum_{j \in P_i} \log \frac{\exp(u_{ij})}{\sum_{k \in \mathcal{V}} \exp(u_{ik})}. \tag{A3}$$

Rows without positives are excluded from the batch ranking mean. Cartridge ranking uses all eligible rows; solvent ranking uses only (i, t) with $a_i m_{i,t} = 1$ and at least one positive solvent. The weighted batch sum is divided by the sum of valid-row weights plus 10^{-6}. Cartridge rank weight is 0.5 for Bal/EM and 2 for Chain/Bag; solvent rank weight is 0.05 throughout.

For the descriptor vectors in a batch of size B, define $\bar{\mathbf{x}}_i = \widetilde{\mathbf{x}}_i / (\|\widetilde{\mathbf{x}}_i\|_2 + 10^{-8})$. The similarity matrix has its diagonal set to zero before selecting $k = \min(8, \max(1, B - 1))$ entries per row. The implemented regulariser is

$$\mathcal{L}_{\mathrm{sm}} = \frac{\sum_i \sum_{j \in \mathcal{N}_k(i)} \max(\bar{\mathbf{x}}_i^{\mathsf{T}} \bar{\mathbf{x}}_j, 0) \|\mathbf{z}_i - \mathbf{z}_j\|_2^2}{Bk + 10^{-8}}. \tag{A4}$$

Clipping the cosine weights makes this regulariser non-negative. A zero diagonal may be selected ahead of negative neighbours, but its self-distance and weight are zero; a singleton batch contributes zero. The regulariser uses descriptors and shared features from every batch row, including cartridge-only samples.

Count heads in the EM variant. The cartridge tower feeds a linear 16-class count head (classes 0–15), and the solvent tower feeds 5×12 count logits (0–11 per step). Counts exceeding the last class are clipped to that class for supervision. Both heads use the same adjacent-class smoothing rule as the ratio head. Cartridge count cross-entropy averages over all batch rows; solvent count cross-entropy averages over the five steps of annotated rows only, including zero counts at absent steps. Each loss has coefficient 0.3. The default scalar estimate is the modal class; the decoder also supports a validation-selected distribution quantile. These auxiliary count heads are specific to the EM variant.

Table A2: Architecture and active training configuration.

Item	Specification
Inputs and split	21 descriptors; split seed 42; 9,152/1,144/1,145 rows, including 8,248/1,044/1,043 with protocol annotations.
Shared encoder / towers	Shared $21 \to 512 \to 512$; four separate $512 \to 512 \to 512$ towers; ReLU and dropout 0.3 after each hidden layer.
Classification projections	$512 \to 42$, $512 \to 5$, $512 \to 5 \times 378$; additive learned step-logit guidance for solvent scores.
Ratio head	Step/solvent embeddings 32/64; $608 \to 256 \to 20$; mode-of-buckets point prediction.
Bal / EM weights	Task weights $(1, 1, 1, 1)$; rank weights $(0.5, 0.05)$; smoothness 0.01. EM additionally uses two count losses, each 0.3.
Chain / Bag weights	Task weights $(1, 1, 0.1, 1)$; rank weights $(2, 0.05)$; smoothness 0.01.
Optimisation	AdamW, learning rate 10^{-3}, weight decay 0, batch 64, 60 epochs, gradient-norm clipping 1.0.
Initialisation seeds	Single-network comparisons: 42, 101, 102, 103. Bag members: 301–308, with replacement resampling to the original training size.
Inactive options	Functional groups, base-solvent decomposition, numeric embeddings, uncertainty weighting and the separate gate penalty.

Checkpoint selection. AdamW (Loshchilov & Hutter, 2019) is used with zero weight decay. Every epoch is evaluated using initial thresholds 0.5 before the later decoder calibration, and the earliest

epoch attaining the maximum score is retained. The implemented score is

$$J_{\text{val}} = 3\,\text{Exact}_{\mathcal{A}} + 1.5\,\text{SSS}_{\text{all}} + 1.2\,\text{HC}_{\text{all}} + 4\,\text{Triplet}_{\mathcal{A}} + \text{GUS}/100. \tag{A5}$$

$\mathcal{A}$ denotes annotated validation rows. $\text{Exact}_{\mathcal{A}}$ requires exact cartridge, step and solvent sets and all available ratio errors within 0.10; an empty ratio-record list passes that last condition. Checkpoint selection uses all-row SSS and HC: unavailable protocol components receive score 1, and cartridge Jaccard is evaluated for every row. Table 2 reports SSS and HC on annotated test rows, as defined in Appendix D. The coefficients of cartridge top-1 and strict Hit@1 in the checkpoint objective are zero. Saved configurations and epoch histories record the selected checkpoint for each run.

Threshold calibration and decoder search. After checkpoint selection, each cartridge label is calibrated on all validation rows; steps and solvents use only annotated validation rows. For a label with at least five positives, binary F1 is maximised over 41 equally spaced thresholds from 0.2 to 0.9. Otherwise the threshold remains 0.5. Predictions use $p \geq \tau$; equal F1 retains the first threshold in the ascending grid. A second search adds a task-wide shift from $\{-0.05, 0, 0.05, 0.10\}$ to every per-label threshold, clipping the result to $[0, 1]$.

Bal, Chain and Bag use threshold decoding. EM evaluates threshold, clamp and cardinality modes, enumerated in that order. Cardinality mode returns the K highest-scoring labels. Clamp intersects the threshold set with top-K only when its size exceeds K. The allowed count is $K = \text{clip}(\lceil \eta \widehat{k} \rceil, K_{\min}, K_{\max})$, with $\eta \in \{0.8, 1, 1.2, 1.5\}$, cartridge bounds $[1, 42]$ and solvent bounds $[0, 378]$. The count bounds apply to clamp and cardinality modes. Threshold mode can return an empty set, with the top-1 fallback disabled.

In EM runs, $\widehat{k}$ is either the modal class or the first class at which the normalised cumulative count distribution reaches $q \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. Count-irrelevant axes are removed from threshold candidates and shift-irrelevant axes from cardinality candidates. The resulting search has 4 threshold, 112 clamp and 28 cardinality configurations; solvent candidates additionally consider gating off/on. Cartridge and solvent decoders optimise separate objectives: micro-F1 for Bal/Chain/Bag and $0.5\,\text{EM} + 0.5\,J$ for EM, where the solvent labels are flattened across steps. Empty unions have $J = 1$. Equal objectives retain the first candidate enumerated. Top-K uses NumPy's default `argsort(-prob)` ordering, so tie ordering follows the installed library version.

All saved Bal/Chain runs and Bag select threshold mode, zero shift and solvent gating. EM selections are listed in Table A3; their full per-label arrays remain in the supplementary run files. The decoder selects component sets within the five fixed workflow-step categories.

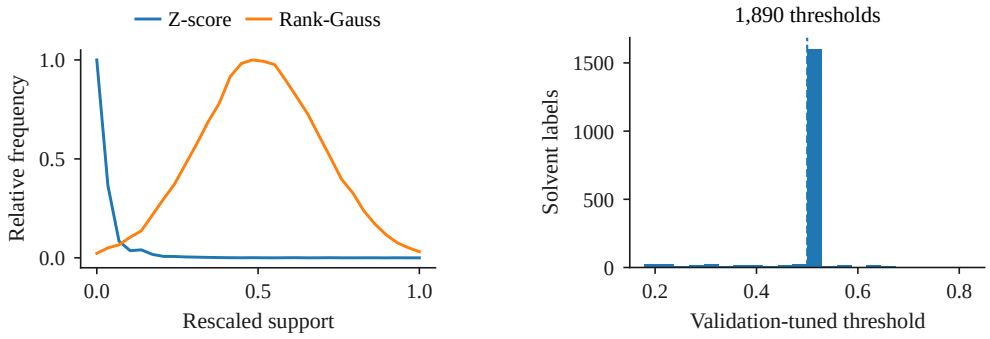
Table A3: Validation-selected EM decoding; count scale is 1 and solvent gating is on.

Seed	Cartridge mode	Quantile	Shift	Solvent mode	Quantile
42	clamp	0.3	−0.05	clamp	0.5
101	cardinality	0.4	0	clamp	0.6
102	clamp	0.3	0	clamp	0.6
103	cardinality	0.3	0	cardinality	0.5

Bagging. Eight independently initialised Chain models use replacement samples of the original training partition. The validation and test partitions remain fixed. Classification sigmoid probabilities are averaged across members. Ratio records are aligned by (i, t, k), and the member modal-bucket point estimates are averaged for each pair. Per-label calibration and F1 decoder search are then repeated on the averaged validation predictions. The saved ensemble selects zero shifts and hard step gating. The implementation checks member label/order agreement and missing ratio keys; the aggregated dataset contains one mean target per observed step–solvent key.

D Evaluation definitions

Populations and data types. Let $\mathcal{I}$ contain the 1,145 test pollutants and $\mathcal{A} = \{i : a_i = 1\}$ the 1,043 with protocol annotations. Cartridge measurements use $\mathcal{I}$; step, solvent, Triplet, strict Hit@1,



(a) (b)

Figure A8: Input and decoder distributions. (a) The heaviest-tailed descriptor before and after rank-Gauss, using the central 0.3–99.7 percentile interval and independently rescaled supports. (b) Validation-tuned thresholds for 1,890 step–solvent labels; the dashed line is 0.5.

SSS and HC use $\mathcal{A}$. Ratio measurements use 6,258 observed step–solvent records, regardless of whether the decoder selected their labels. There are 904, 100 and 102 cartridge-only rows in the training, validation and test partitions. Protocol-dependent metrics are restricted to $\mathcal{A}$ through the availability indicator.

Set and ratio scores. Micro counts pool TP, FP and FN over the appropriate population and all output labels. Cartridge exact match compares the whole predicted cartridge set with its target:

$$F_{1,\text{micro}} = \frac{2\,\text{TP}}{2\,\text{TP} + \text{FP} + \text{FN}}, \qquad \text{EM}_c = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbb{1}[\widehat{\mathbf{c}}_i = \mathbf{c}_i]. \tag{A6}$$

The source implements F1 through precision and recall with 10^{-12} denominators and returns zero for an all-zero micro denominator. Let $\mathcal{R}$ be the observed ratio-record list. The ratio columns are

$$\text{MAE}_r = \frac{1}{|\mathcal{R}|} \sum_{(i,t,k) \in \mathcal{R}} |\widehat{r}_{i,t,k} - r_{i,t,k}|, \qquad A_r(\epsilon) = \frac{1}{|\mathcal{R}|} \sum_{(i,t,k) \in \mathcal{R}} \mathbb{1}[|\widehat{r}_{i,t,k} - r_{i,t,k}| \le \epsilon]. \tag{A7}$$

The displayed tolerance is $\epsilon = 0.10$. Both metrics compare predictions with the aggregated archive ratios.

Component-triplet accuracy. Component-triplet accuracy is stored as `chain_acc_no_ratio`. Let $j_i^* = \arg\max_j p_{ij}^c$, $t_i^* = \arg\max_t p_{it}^m$ and $k_i^* = \arg\max_k p_{i,t_i^*,k}^s$. Ties take the first maximum in the stored label order. The highest-scoring indices are evaluated independently of the threshold-selected sets. Define

$$C_i = c_{i,j_i^*}, \quad M_i = m_{i,t_i^*}, \quad S_i = s_{i,t_i^*,k_i^*}, \qquad \text{Triplet} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} C_i M_i S_i. \tag{A8}$$

The event tests one cartridge–step–solvent triple against the pollutant-level aggregate labels. Its scope is component membership; assessing method-level compatibility additionally requires the original joint assignments, and complete-set recovery is evaluated through exact-match metrics.

All factors in Figure 4(b) use the same $\mathcal{A}$:

$$\text{Triplet} = P(C \mid \mathcal{A}) P(M \mid C, \mathcal{A}) P(S \mid C, M, \mathcal{A}). \tag{A9}$$

For Bag these factors are $665/1043$, $665/665$ and $521/665$, giving $521/1043 = 0.4995206$. The main-table cartridge top-1 uses the full test population, yielding $703/1145 = 0.6139738$. Random forest uses $644/1043$, $639/644$ and $539/639$, giving $539/1043 = 0.5167785$.

Strict ratio-verified Hit@1. Let A_i^r indicate an available ratio record for the selected (t_i^*, k_i^*), and e_i^* be its absolute error (the mean absolute error if multiple records share the key). Then

$$\text{Hit@1}_{\text{strict}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} C_i M_i S_i A_i^r \mathbb{1}[e_i^* \le 0.10]. \tag{A10}$$

All models use the `hit_at_1_strict` field with the common denominator $|\mathcal{A}|$. A hit requires all three component memberships, an observed ratio for the selected pair and an absolute ratio error within 0.10. For Bag, 521 triples satisfy component membership; 379 lack a selected-pair ratio and 51 exceed the tolerance, leaving $91/1043 = 0.0872483$ verified hits. The score measures evidence-supported component and ratio prediction within the archive. Keeping the denominator fixed across models allows direct comparison despite differences in their selected pairs.

Scheme similarity score (SSS). Write $J(Y, \widehat{Y}) = |Y \cap \widehat{Y}|/|Y \cup \widehat{Y}|$, with $J(\varnothing, \varnothing) = 1$. For annotated row i, let J_i^c compare cartridge sets, J_i^m compare step masks, and $J_{i,t}^s$ compare solvents in step t. All five steps enter the solvent average, including empty steps. Let

$$R_i^{\mathrm{sim}} = \begin{cases} \dfrac{1}{|\mathcal{R}_i|} \displaystyle\sum_{(t,k)\in\mathcal{R}_i} \max\left(0, 1 - \frac{|\widehat{r}_{i,t,k} - r_{i,t,k}|}{0.10}\right), & |\mathcal{R}_i| > 0, \\ 1, & |\mathcal{R}_i| = 0, \end{cases} \tag{A11}$$

$$\mathrm{SSS} = \frac{1}{|\mathcal{A}|} \sum_{i\in\mathcal{A}} \left(0.2 J_i^c + 0.2 J_i^m + \frac{0.4}{5} \sum_{t=1}^{5} J_{i,t}^s + 0.2 R_i^{\mathrm{sim}}\right).$$

The four component weights remain fixed, and samples without ratio observations receive $R_i^{\mathrm{sim}} = 1$. This convention assigns favourable ratio similarity to incomplete records. We therefore report SSS as an auxiliary summary alongside strict Hit@1, which explicitly requires an observed ratio. The code field is `scheme_similarity_info_only`. The all-row version used for checkpoint selection and the archived oracle table assigns $0.2J_i^c + 0.8$ to unannotated rows and averages over $\mathcal{I}$.

Step–solvent occupancy agreement (HC). Define $e_{i,t} = \mathbb{1}[\sum_k \widehat{s}_{i,t,k} > 0]$. The stored field `hierarchical_consistency_info_only` is

$$\mathrm{HC} = \frac{1}{5|\mathcal{A}|} \sum_{i\in\mathcal{A},t} \mathbb{1}[e_{i,t} = m_{i,t}]. \tag{A12}$$

HC uses the annotated step indicator $m_{i,t}$ as its reference. A step contributes one agreement when its predicted solvent occupancy matches the annotation. Errors arise when a recorded step is assigned an empty set or an unrecorded step receives solvents. Step gating couples occupancy to $\widehat{m}$ through Equation 6, so errors in step prediction also affect HC. This score assesses step–solvent occupancy; chemical compatibility requires separate evaluation. The all-row checkpoint version assigns value 1 to each unannotated row.

Global utility score (GUS). The implementation pools task counts before combining scores:

$$\mathrm{GUS} = 100\left(0.25F_{1,c}^{\mathcal{I}} + 0.25F_{1,m}^{\mathcal{A}} + 0.35F_{1,s}^{\mathcal{A}} + 0.15A_r^{\mathrm{records}}(0.10)\right). \tag{A13}$$

A_r^{records} pools errors from the ratio-record list; a wholly empty list contributes zero. Cartridge F1 uses all test pollutants; step and solvent F1 use the annotated subset. Each F1 term pools the task-level counts before weighting. GUS is reported as an auxiliary score on a 100-point scale.

The saved GUS fields use record-precision ratio errors, whereas the separate neural ratio arrays are cast to float32 before ratio metrics convert them to float64. Tolerance boundaries can therefore differ between code paths. For Bag, the saved GUS is 61.5504679; recombining its saved array-level submetrics gives 61.5408802. This difference is equivalent to four tolerance decisions among 6,258 records. Tables use the stored record-level aggregate. Resolving the four individual tolerance decisions requires the full per-sample prediction arrays, which are absent from the supplementary archive.

Variance and descriptive analyses. For each repeated configuration, the four saved seed scores $v_1, \ldots, v_4$ are summarised by their arithmetic mean and sample standard deviation $\sqrt{\sum_i (v_i - \bar{v})^2/3}$. The same arrays and ddof=1 are used in every table and error bar. Bold marks the best point estimate in each column. Baselines and the eight-member ensemble each have a single evaluation; the ensemble's eight constituent networks contribute to that one aggregate prediction. GUS is on a 100-point scale; plots marked percent multiply decimal scores by 100. Full-corpus enrichment

analyses use quantile bins, omit cells with fewer than 40 molecules, use Wilson intervals (Wilson, 1927) and Cohen's d (Cohen, 1988). CART visualisations use balanced class weights and a held-out 20% partition of the single-label subsets.

E ADDITIONAL QUANTITATIVE ANALYSES

Figure A9 compares the decoding trade-offs. Reported cartridge count averages 3.27 over all test rows; reported solvent count averages 15.24 positive step–solvent cells over annotated rows. For context, collapsing repeated tokens across workflow steps gives a full-corpus mean of 8.25 distinct solvent tokens. The cardinality comparison uses the step–solvent-cell count, matching the solvent output tensor.

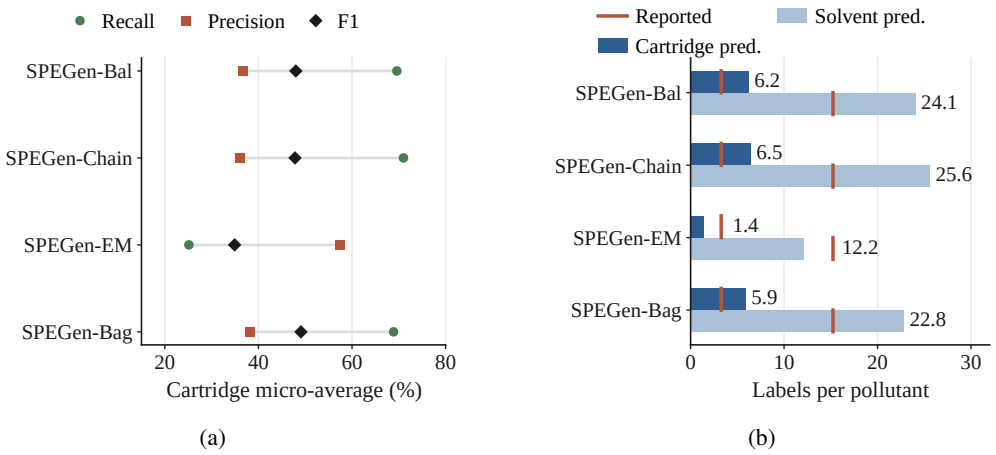


Figure A9: Decoder trade-offs. (a) Cartridge precision, recall and F1. (b) Mean predicted and reported cartridge-label and step–solvent-cell counts. F1-oriented variants return larger sets than exact-match decoding.

Paired intervention results. Additional interventions are organised into two seed-42 control families. R1 is `output/screen21/R0_control_seed42`, with C-top1 0.5091703. R2 is `output/improve21/s1_C12_rank_gauss_init42`, with 0.5868996. The former is the source configuration for numerical embeddings and uncertainty weighting; the latter is the rank-Gauss configuration for alternative transforms and ranking objectives. Table A4 reports each control score, intervention score and their difference, computed at full precision before rounding. The control configuration is fixed separately for each experimental family.

Table A4: Paired single-run interventions against recorded controls.

Intervention	**Control**	**Control score**	**C-top1**	**Δ (pp)**
Top-1 hinge	R2	0.5869	0.4961	−9.083
Uncertainty weighting	R1	0.5092	0.4262	−8.297
RankNet	R2	0.5869	0.5502	−3.668
Piecewise-linear embedding	R1	0.5092	0.4891	−2.009
Median/IQR transform	R2	0.5869	0.5668	−2.009
Signed log1p transform	R2	0.5869	0.5860	−0.087
Periodic embedding	R1	0.5092	0.5153	+0.611
Any-positive ranking	R2	0.5869	0.5948	+0.786

Top-1 hinge and uncertainty weighting reduce C-top1 by 9.083 and 8.297 percentage points in their respective paired runs. Any-positive ranking and periodic embedding increase it by 0.786 and 0.611 points. Each contrast is based on one paired run and serves as an exploratory comparison; repeat runs would be needed to estimate its variability. Figure A10 distinguishes the two control groups without applying one shared noise band to differently configured experiments.

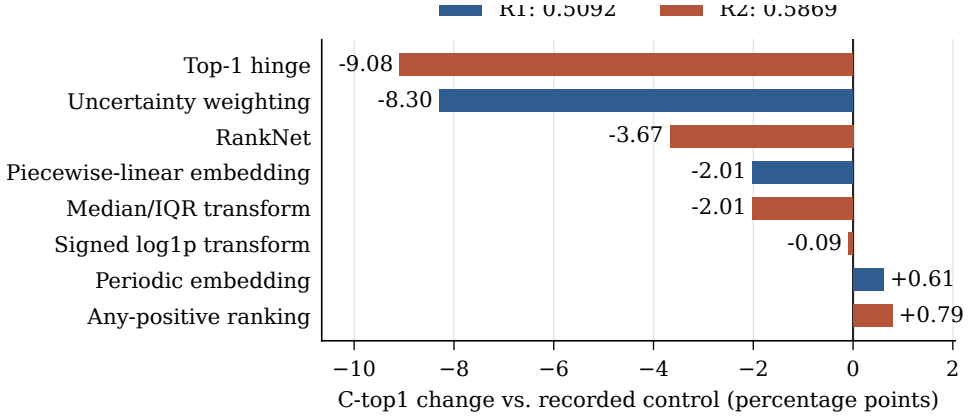


Figure A10: Single-run changes in cartridge top-1 against identified controls R1 and R2. Control IDs and scores are given in Table A4; colour denotes the control family.

Fixed-ranking oracle size. The oracle diagnostic uses the seed-42 Chain checkpoint with its own threshold-only reference configuration. Table A5 reports this individual run and its associated oracle intervention. The diagnostic replaces cartridge and per-step solvent counts with true counts, selects the corresponding top-K scores and gates solvents using predicted steps. Score arrays, ratio predictions and step scores remain unchanged. Oracle solvent set size can therefore remain below the true value when a true-present step was predicted absent.

Table A5: Fixed-score oracle-size diagnostic for the seed-42 Chain run.

Metric	Threshold reference	Oracle size
Cartridge F1	0.4805	**0.6532**
Cartridge exact match	0.0183	**0.2498**
Solvent F1	0.4564	**0.6470**
Component-triplet accuracy	**0.4909**	**0.4909**
$\mathrm{SSS}_{\mathrm{all}}$	0.5085	**0.6286**
GUS	61.35	**72.34**
Mean cartridge set size	6.43	3.27

Table A5 reports all-row similarity $\mathrm{SSS}_{\mathrm{all}}$; Table 2 uses annotated-only SSS. Appendix D defines both aggregation populations. GUS uses the same record-level definition, and the source is `output/improve21/oracle_cardinality.json`. A fixed-ranking size replacement changes the selected sets while leaving Triplet, the score ranking and the ratio estimates unchanged. With an imperfect ranking, a selection larger than the true count can sometimes achieve higher F1, so the oracle-size result is a diagnostic of one decoding intervention.

The dedicated count predictor obtains test $R^2 = 0.0933/0.0513$ and Spearman correlations $0.3074/0.3245$ for cartridge/solvent counts. These modest values show that this fitted predictor captures only part of the variation in recorded set sizes, motivating further study of count estimation and context-dependent supervision.

F Interpretation of aggregated supervision

Source-method associations. Pollutant-level aggregation preserves the alternatives documented across source methods. For example, methods pairing cartridge c_1 with solvent system s_1 and cartridge c_2 with system s_2 produce the target sets $\{c_1, c_2\}$ and $\{s_1, s_2\}$. The archive directly supports the pairs (c_1, s_1) and (c_2, s_2); using the cross-pair (c_1, s_2) requires additional compatibility evidence. Similar considerations apply to choices collected across different steps, sample matrices and concentrations. A shared encoder learns descriptor associations under this supervision. To

construct an executable method from its predictions, an analyst must select mutually compatible components, recover their method-specific context and verify the resulting procedure.

Ratio summaries and mixture composition. Averaging repeated ratios produces one target for each step–solvent pair. Fractions 0.5 and 0.9, for example, yield a mean of 0.7. This illustrative case shows how aggregation reduces multiple source observations to a single value; the corpus-wide prevalence of such conflicts remains unquantified. Single-network inference selects a modal bucket centre through Equation 4. Bagging then averages the member point estimates, potentially producing values between bucket centres. Both predictions are evaluated against the aggregated target. Preserving multiple method-specific ratio modes would require corresponding method-level supervision.

A solvent token can represent a neat solvent or a composite system such as methanol+water. A union of tokens therefore contains alternative solvent systems, while mixture fractions refer to the constituents of one selected system. Physically constrained composition prediction would first identify the method and its constituent list, then impose non-negativity and a sum-to-one condition within that mixture. The current bounded ratio estimates describe individual step–solvent records; joint mixture validation remains an additional stage of experimental method development.

Documentation and feasibility. The archive supplies observed positives, and absent labels enter the classification objectives as computational negatives. Consider an idealised model in which each documented choice is feasible in its recorded context. Let V indicate feasibility, D indicate selection and documentation, and $O = VD$ indicate an observed positive. Conditional on molecular descriptors x and analytical context a,

$$P(O = 1 \mid x, a) = P(D = 1 \mid V = 1, x, a)P(V = 1 \mid x, a). \tag{A14}$$

The expression follows the probability chain rule. Inferring feasibility from observed positives requires information about the documentation process and analytical context. Focal and asymmetric losses handle label imbalance under the observed supervision, and decoder calibration targets recovery of recorded choices. Experimental failures, contextual annotations and prospective laboratory measurements would provide the additional evidence needed to assess extraction success and recommendations for unfamiliar compounds.

G Reproducibility and comparison scope

Run-level provenance. Tables and performance plots use the same saved JSON records. The supplementary reporting manifest contains field mappings, source hashes and full-precision per-seed scores. The four non-bootstrap seeds are 42, 101, 102 and 103. Their run families are `s5_C2t08_init*` (Bal), `hp/t08_h512_l2_d0.3_lr0.001_wd0_bs64_init*` (Chain) and `s8_EMt08_init*` (EM), under `output/improve21/`. Bag uses `s6_bagt08_init301` through `init308` and the combined file `ens_bagt08.json`. These run identifiers link each configuration to its saved results. The manifest also lists the paired intervention and control paths.

Baseline implementation. The baseline entry point uses training-fitted descriptor z-scoring and one-vs-rest classification with fixed threshold 0.5. All baseline records use this fixed decision rule. Decision trees use library defaults apart from seed 42; random forests use 300 trees for classification and 500 for ratios. SVM denotes LinearSVC (sigmoid-transformed decision scores) and LinearSVR. MLP uses hidden sizes $(256, 128)$, ReLU, L_2 coefficient 10^{-4}, learning rate 10^{-3}, 300 classifier or 400 regressor iterations, and no early stopping. XGBoost uses 400 classifier or 500 regressor trees, depth 8, learning rate 0.05, and row/column subsampling 0.8. LightGBM uses 500 classifier or 700 regressor trees, 63 leaves, learning rate 0.05, and row/column fractions 0.8. Each ratio regressor receives descriptors concatenated with one-hot step and solvent IDs, with a second training-fitted standardisation; predictions are clipped to $[0, 1]$. The baseline configurations and remaining library defaults are specified by `baseline.py`. Separate baseline hyperparameter-search records are absent from the archive.

Neural search and computational comparison. The neural search script considers widths $\{128, 256, 384, 512\}$, depths $\{2, 3, 4\}$, dropout $\{0.1, 0.15, 0.2, 0.3\}$, learning rates $\{3 \times 10^{-4}, 5 \times$

Table A6: Implementation sources and available reproduction records.

Component	Source or record
Target construction	`utils/dataloader.py`, `utils/parsing.py`; processed molecular/protocol tables and exact descriptor order.
Input transform	`spe/data.py`; insertion-count transform, clipping and post-transform training statistics (Appendix C).
Network and objectives	`model.py`, `spe/model.py`, `spe/losses.py`, `spe/train.py`; task towers, count/ratio heads, masks and non-negative neighbour weights.
Selection and decoding	`spe/evaluate.py`, `spe/decode.py`; saved configuration, epoch history, per-label thresholds and decoder selection for each run.
Metrics and aggregation	`spe/metrics_strict.py`; recorded strict Hit@1 and annotated-only SSS/HC; sample standard deviation and paired-control reporting manifest.
Prediction-level replay	The archive supports source inspection and saved-score aggregation; prediction-level replay additionally requires checkpoint weights and full per-sample arrays, which are absent.

$10^{-4}, 10^{-3}, 2\times10^{-3}\}$, weight decays $\{0, 10^{-5}, 10^{-4}, 10^{-3}\}$ and batches $\{64, 128, 256\}$. Its default search budget is 20 sampled distinct configurations with sample seed 7, followed by recorded repeat runs. Per-run files list the configurations evaluated and their selected checkpoints. Checkpoints use Equation A5; decoder selection uses validation data. The model families share descriptor information and a common evaluator, with different preprocessing, calibration and optimisation procedures. Bag combines eight independently trained networks. Matching search effort and outer-loop ensemble size across model families would provide a more controlled computational comparison.

Reproducibility scope. The archive includes data-processing, model, training, decoding, evaluation and figure scripts, together with 121 run directories containing saved test summaries and associated records. The reporting manifest regenerates numeric tables from these summaries. Synthetic tests cover gating, HC, non-negative regularisation and modal-bucket decoding. Full prediction-level replay requires the trained checkpoints and raw per-sample arrays, including the records needed to resolve the small GUS tolerance discrepancy in Appendix D. These files are absent from the archive, which limits independent verification to implementation behaviour and saved-score aggregation.

The fixed random split supports retrospective comparison among pollutant identifiers within this corpus. Chemical scaffolds and source publications may overlap across partitions. Evaluating scaffold- and source-disjoint holdouts, dependency-aware and modern tabular baselines, matched outer-loop ensembles, and prospective laboratory performance would extend the present evidence. Public resources such as the Open Reaction Database (Kearnes et al., 2021) offer opportunities for separately defined chemical-procedure tasks. The cumulative ablation quantifies the gains along one ordered sequence of interventions; factorial or leave-one-out experiments would help isolate component interactions.