Discovery report for chem

Research Objective

I can test a few thousands wells of compounds per week for an activity assay via a large batched experiment. I want to make the compounds using some kind of simple coupling reactions. What are computational ways to propose the compound blocks/reactions? What ways can I use to also update the predictions after each round of experiments?

Summary of Discoveries

Discovery 1: Synthesis-Aware Library Generation Using Robust Coupling Reactions

Synthesis-aware library generation that restricts exploration to buyable building blocks and a small set of robust coupling reactions is the most reliable way to populate weekly, thousand-compound activity assays. Encoding these couplings as reaction SMARTS in RDKit or KNIME enables forward enumeration at scale, and closed-loop active learning retrains activity and yield models after each batch to steadily improve selection quality.

Discovery 2: Filter-First Design and Predictive Prioritization for Activity and Yield

Filter-first library design paired with multi-objective predictive prioritization enables high-throughput testing of coupling products that are both synthetically tractable and likely active. Inexpensive 2D features augmented with explicit reaction-condition encodings support accurate yield prediction on fixed-protocol HTE, whereas extrapolative performance degrades and model complexity (e.g., GNNs) offers no clear advantage over robust baselines. Iterative, batched acquisitions guided by Pareto-efficient criteria integrate predicted activity, yield, and procurement constraints, with model updates after each experimental round.

Discovery 3: Closed-Loop Active Learning and Large-Batch Candidate Selection

Closed-loop active learning can turn large-batch assays into an adaptive design loop that reliably outperforms random screening while respecting synthetic constraints. For moderate batches, multi-objective Bayesian optimization with Pareto-based acquisitions selects compounds that are both active and makeable; for very large batches, scalable uncertainty-aware selection remains challenging, so diversification with greedy or scalarization portfolios is often preferred.

Discovery 4: Machine-Learned Recommendation of Reaction Conditions for Coupling Reactions

Machine learning models that only “see” the two coupling partners do not reliably recommend bases or solvents beyond a naïve popularity baseline, whereas models that encode the full reaction graph (reactants and products) deliver substantially better top-k condition recommendations. For high-throughput coupling campaigns, this points to using reaction-level encodings, balanced high-throughput datasets with negative outcomes, and iterative retraining after each experimental batch to debias and improve recommendations.

Synthesis-Aware Library Generation Using Robust Coupling Reactions

Summary

Synthesis-aware library generation that restricts exploration to buyable building blocks and a small set of robust coupling reactions is the most reliable way to populate weekly, thousand-compound activity assays. Encoding these couplings as reaction SMARTS in RDKit or KNIME enables forward enumeration at scale, and closed-loop active learning retrains activity and yield models after each batch to steadily improve selection quality.

Background

High-throughput discovery campaigns live at the intersection of chemical space exploration and reliable synthesis. While generative models can propose diverse structures, the bottleneck in weekly plate-based execution is delivery risk: compounds must be reachable rapidly from commercial building blocks via reactions that succeed without bespoke optimization. The medicinal chemistry toolkit is dominated by a handful of C–C and C–N couplings and amide formations that map cleanly to rule-based templates; modern toolchains can apply these templates to reagent catalogs to create make-on-demand libraries, while iterative modeling and active learning decide which candidates to test next.

Results & Discussion

The central finding is that constraining library design to commercially available reagents and a small set of robust couplings maximizes executable hit rate at weekly throughputs. Empirical high-throughput experimentation on a Chemspeed platform without per-reaction optimization found 100% one-shot detection for Suzuki–Miyaura and Schotten–Baumann amide couplings, 84% for Buchwald–Hartwig C–N formation, and 83% for peptidic couplings, whereas sulfonamide formation achieved 71% and Heck and Sonogashira couplings were approximately 67% in the same campaign; "one-shot success" was defined by product detection above a five percent UV area threshold in a single attempt, highlighting class-dependent baseline reliability under automation constraints [r1, vaillant2025]. These high-performing classes align with the reactions most widely deployed in plate-based arrays and miniaturized formats, where homogeneous, aerobic conditions in high-boiling solvents are favored and air-sensitive or heterogeneous systems frequently fail without careful engineering [r1, krska2017, shen2021, mahjour2021]. Because these couplings also map cleanly to rule-based forward synthesis, they offer the best bridge between computational proposal and practical synthesis for campaigns processing a few thousand wells per week [r0].

To propose compounds at scale, reaction-based enumeration from buyable reagents is the workhorse. Workflows such as PathFinder encode more than one hundred reaction SMARTS covering amide formation, Suzuki, Sonogashira, alkylation, and ether couplings, enabling retrosynthetic disconnections and forward combinatorial rebuilds from catalogs like eMolecules, MolPort, or ZINC; these libraries are routinely triaged with multi-parameter optimization, docking, or free energy methods before selection [r0, konze2019, walters2018virtualchemicallibraries]. Industry-scale forward enumeration of prevalidated two to three step sequences underpins on-demand REAL or MADE spaces with demonstrated greater than 80% experimental success per step, providing a template for robust, simple-coupling campaigns [r0, grygorenko2020]. Generative approaches can be made executable by constraining the search to products reachable from available reactants via a forward reaction oracle or synthetic graph walk, as in MoleculeChef, ChemBO, or DOGS, which mirror virtual-library workflows and have delivered high single-step delivery rates in practice [r0, gao2020, tropsha2024]. Hierarchical synthon strategies and chemical-space docking (for example, V-SYNTHES or BioSolveIT workflows) enumerate R-groups in stages or dock fragments and then enumerate products via validated couplings to traverse ultra-large on-demand spaces, though fragment-first docking requires subsequent validation [r0, sadybekov2023]. Fragmentation/recombination rules such as RECAP or BRICS provide buildable fragments and recombination logic keyed to common bond formations, and are typically combined with PAINS, lead-like, and diversity filters to maintain quality [r0, bon2022].

Executability hinges on robust reaction encodings and early pruning. Public, generalized reaction SMARTS are available for amide formation, Suzuki–Miyaura, Sonogashira, aromatic C–N formation, and reductive amination, with best practices including explicit atom mapping, mapped cores, and leaving-group lists that extend halides to pseudohalides where mechanistically supported; protection and conflict handling are implemented chiefly via reactive-handle-aware filters, masking, and workflow logic rather than embedded in the SMARTS [r8, cox2016, warr2021a]. Stereochemical integrity is preserved by using stereochemically annotated SMARTS and RDKit substructure matching augmented with RDChiral-style local rules to copy, invert, destroy, or create stereocenters and E/Z double-bond geometry, with post-reaction sanitization and symmetry-aware stereoisomer enumeration as needed [r9, coley2019, kochev2018]. Pre-enumeration filtering of reagents with reaction-aware rules is markedly more efficient than enumerate-then-filter, reducing combinatorial blowup and enabling throughputs on the order of tens to hundreds of thousands of products per minute on modest compute; anchor-then-enumerate strategies recover top-ranked chemotypes while evaluating only a small fraction of poses [r12, takacs2024, warr2021, muller2022]. Route fidelity and cost can be layered in by reserving computer-aided synthesis planning or expert checks for top-ranked subsets because per-molecule retrosynthesis is seconds to a minute and thus too slow for the full library; simple synthesizability heuristics are cheaper but can misalign with discovery objectives and are best used post hoc [r0, gao2020]. For the chosen simple couplings, machine learning yield models trained on high-throughput or electronic lab notebook data help prioritize building blocks and conditions, with hybrid quantum descriptors plus Morgan fingerprints outperforming one-hot encodings, although out-of-domain generalization remains modest and careful curation is essential [r0, raghavan2024].

After each assay round, the proposal engine is updated in a closed loop. Batch-aware active learning retrains quantitative structure–activity relationship models on the new outcomes and uses uncertainty or ensemble diversity to select the next set from the forward-reachable pool; integrating coupling-specific yield predictors enables multi-objective selection that jointly optimizes activity, yield, and cost [r0, tropsha2024, raghavan2024]. Bayesian optimization over categorical choices (for example, discrete building blocks, ligands, bases, and conditions) with platforms like Gryffin or NEXTorch supports single or multi-objective criteria, including differentiable expected hypervolume improvement, and fits naturally to plate-based iterations [r0, torres2022]. Physics-informed enrichment loops, as exemplified by PathFinder, iterate fast machine learning enrichment with higher-fidelity scoring such as docking or free energy perturbation to triage enumerated candidates before synthesis, updating surrogates each round to improve sample efficiency [r0, konze2019]. A practical template emerges: select four to eight robust coupling classes; curate compatible building blocks from commercial catalogs; enumerate with reaction SMARTS; filter by PAINS, lead-like, shape, and diversity; overlay activity models, docking, and yield predictors to prioritize a few thousand high-value, high-deliverability compounds per batch; then retrain and apply uncertainty-aware acquisition to pick the next batch, reserving retrosynthesis for promotion decisions [r0, suaygarcia2022, zoete2016]. Remaining gaps include fast, accurate multi-step fidelity estimators and broadly generalizable yield predictors, as well as acquisition functions that incorporate cost, availability, and delivery-time metadata to better serve high-throughput operations [r0, raghavan2024, torres2022].

Trajectory Sources

Trajectory r0: Overview and motivation For a program that can assay a few thousand wells per week and prefers simple couplings, synthesis-aware proposal methods that restrict search to available building blocks and robust reaction templates are the most reliable and scalable way to populate batched experiments whi...

Trajectory r1: The hypothesis is only partially supported: several named couplings (Suzuki–Miyaura, Buchwald–Hartwig, and amide formations) show high one-shot success rates on automated platforms, but consistent >80% performance across 5–10 reactions and corresponding SMARTS encodings for these classes are not dem...

Trajectory r8: Supported: public, openly documented workflows and toolkits provide generalized reaction SMARTS for amide coupling, Suzuki–Miyaura, Sonogashira, SNAr/C–N arylation, and reductive amination; protection/conflict handling is implemented primarily via filters, masking, and workflow logic rather than emb...

Trajectory r9: The hypothesis is supported: the most reliable large-scale RDKit-based enumeration workflows preserve stereochemical integrity by using explicitly atom-mapped, stereochemically annotated reaction SMARTS and leveraging RDKit-driven stereo detection/matching with additional local rules to copy, invert...

Trajectory r12: The literature and open-source workflows consistently support that applying property/substructure and reaction-aware filters to reactants before enumeration is markedly more computationally efficient and, when calibrated, yields final product libraries with quality/diversity comparable to those obta...

Filter-First Design and Predictive Prioritization for Activity and Yield

Summary

Filter-first library design paired with multi-objective predictive prioritization enables high-throughput testing of coupling products that are both synthetically tractable and likely active. Inexpensive 2D features augmented with explicit reaction-condition encodings support accurate yield prediction on fixed-protocol HTE, whereas extrapolative performance degrades and model complexity (e.g., GNNs) offers no clear advantage over robust baselines. Iterative, batched acquisitions guided by Pareto-efficient criteria integrate predicted activity, yield, and procurement constraints, with model updates after each experimental round.

Background

High-throughput experimentation and modern vendor-indexed catalogs make it feasible to explore tens of thousands of simple coupling products, but combinatorial explosion and uneven synthetic success challenge library construction and prioritization. Pre-enumeration filtering grounded in reactivity rules and pragmatic procurement constraints has emerged as a scalable way to assemble reaction-ready sets. At the same time, machine learning models can predict activity and yield, yet their generalization depends strongly on how reaction conditions are encoded and how extrapolative validation is defined. The convergence of these trends motivates closed-loop, multi-objective selection strategies that learn from each assay batch to efficiently traverse chemical and process design spaces.

Results & Discussion

A practical way to propose compound blocks and reactions at the scale of a few thousand wells per week is to adopt a filter-first workflow that curates building blocks before any enumeration. Programmatic access to vendor-indexed sources such as the ZINC/Arthor building-block index and the Chemspace API enables automated triage by price and stock status, for example by collapsing vendor “BB-50/40 in-stock” versus “BB-30/20/10 slower/more expensive” tiers into actionable “cheap” versus “expensive” pools to control procurement cost and delivery time [r2]. Structural quality is enforced up front using RDKit FilterCatalog (PAINS A/B/C, Brenk, NIH alerts) and ZINC20 reactive/unstable SMARTS, optionally complemented by reaction-specific inclusion/exclusion SMARTS, followed by Lipinski/Veber and QED filters to define drug-like subsets for enumeration and storage in an RDKit-enabled database [r2]. This pre-enumeration pruning is not only tractable at scale but also standard practice: open workflows report generating approximately 14.0 million two-step products from 1,000 curated reagents and achieving throughput around 134 thousand products per minute on 24 cores, while avoiding the storage and compute burden of full, blind enumeration [r12]. However, filters must be calibrated; aggressive rules can excise viable space (strict settings removed about 90% of approved drugs), and despite the intuitive appeal of reaction-specific SMARTS, there is no head-to-head evidence in the supplied literature that they outperform general-purpose alerts for predicting coupling failures [r12, r21].

Within this curated space, simple C–N and C–C couplings such as Buchwald–Hartwig and Suzuki–Miyaura provide robust reaction templates for enumeration and yield prediction. On well-controlled HTE datasets, models that combine inexpensive 2D features with explicit encodings of reaction conditions (e.g., one-hot ligands, bases, solvents, temperatures) routinely achieve high accuracy, with reported held-out R^2 values exceeding 0.7; for instance, XGBoost with one-hot condition features has reached roughly 0.75–0.80 [r42]. Transformer models that consume reaction SMILES (implicitly encoding structure and conditions) have reported R^2 up to 0.956 on Buchwald–Hartwig HTE, whereas performance on heterogeneous literature data is substantially lower ($R^2 \approx 0.388$), underscoring the importance of consistent condition representation [r42]. When conditions are fixed, even models using only reactant 2D structure can perform strongly on HTE, with R^2

well above 0.7 and up to approximately 0.95 on Buchwald–Hartwig and around 0.81 on Suzuki–Miyaura datasets, making structure-only baselines a reasonable choice for standardized protocols [r56]. Together, these results motivate encoding condition variables wherever they vary and defaulting to structure-only surrogates under fixed, single-protocol screens [r42, r56].

Generalization beyond the immediate training regime remains a key risk, and extrapolative validation should guide model choice. In a head-to-head study on a 3,955-reaction Buchwald–Hartwig HTE dataset under fixed conditions, a graph neural network (MPNN with Mol2Vec embeddings) did not show a clear advantage over a descriptor-based Random Forest when evaluated under additive-wise extrapolative splits: reported held-out R^2 for the Random Forest were 0.93 (random split), 0.81 (Reference), and 0.59 (Out-of-AD), versus 0.96, 0.77, and 0.60 for the MPNN-Mol2Vec, respectively [r67]. The collapse in R^2 under extrapolation highlights narrow applicability domains in HTE and suggests prioritizing robust baselines with explicit condition encodings over more complex architectures unless validated gains are demonstrated on series-wise holdouts [r42, r67].

To prioritize which enumerated products to test each week, multi-objective acquisition should integrate predicted biological activity with a surrogate for synthetic success (predicted yield or synthesizability), while respecting procurement constraints learned during prefiltering. Pareto-based strategies such as Expected Hypervolume Improvement (EHVI) select batches that are expected to most increase the dominated volume in the two-objective space, avoiding the need to predefine weights; alternatively, weighted-sum scalarization can reflect project-specific trade-offs [r7]. Across studies, these approaches have efficiently converged toward Pareto-optimal regions when assessed by the hypervolume indicator, and the same machinery readily supports diversity-promoting constraints in batched acquisitions to better cover chemical space [r7]. In practice, the workflow initializes with a seed set, fits activity and yield predictors (using 2D features with condition encodings as appropriate), computes uncertainty-aware acquisition values for each candidate, and selects a batch that jointly optimizes activity, yield, diversity, and cost/lead-time constraints, then refits models after each round to incorporate new outcomes [r2, r7, r42].

This filter-first, predictive-prioritization loop aligns with the experimental cadence of a few thousand wells per week. Vendor-linked availability and cost metadata keep the reaction plan executable; reaction-aware and general structural alerts maintain library quality without over-pruning; and yield models that encode conditions stabilize prioritization under varied protocols. Because fixed-condition HTE can inflate apparent accuracy, the selection and retraining steps should be evaluated with series-wise holdouts and extrapolative splits to avoid overconfidence, and model families should be chosen for demonstrated performance rather than complexity per se [r2, r12, r42, r56, r67]. Finally, given the lack of head-to-head evidence that reaction-specific SMARTS outperform general-purpose filters in predicting coupling failures, rule sets should be treated as hypotheses to be tested and refined as the closed loop accrues mechanistically annotated outcomes [r21].

Trajectory Sources

Trajectory r2: The hypothesis is supported: effective workflows use programmatic vendor/catalog access (e.g., Arthor/ZINC, Chemspace API) combined with price/stock filters and SMARTS/FilterCatalog structural-alert screening to yield high-quality building-block sets for enumeration (bedart2023 pages 6-8...

Trajectory r7: The surveyed literature supports the hypothesis that integrating predicted reaction yield as a second objective—using either Pareto-based methods such as EHVI or weighted-sum scalarization—enhances the selection of compounds that are both active and synthetically accessible (du2025advancinggeneticen...

Trajectory r12: The literature and open-source workflows consistently support that applying property/substructure and reaction-aware filters to reactants before enumeration is markedly more computationally efficient and, when calibrated, yields final product libraries with quality/diversity comparable to those obta...

Trajectory r21: I cannot answer: none of the provided sources report head-to-head benchmarks comparing reaction-specific substructure filters (e.g., SMARTS-RX/PathFinder/SAVI) to general-purpose filters (e.g., PAINS/Brenk/GSK) on HTE or ELN coupling datasets with precision/recall/F1/accuracy, so the hypothesis cann...

Trajectory r42: The literature supports that machine learning models combining 2D structural features with explicit encodings of reaction conditions can achieve high predictive performance ($R^2 > 0.7$) for common couplings in well-controlled HTE datasets (schwaller2021 pages 3-5, zhu2021prediction...

Trajectory r56: The literature supports that machine learning models using only reactant structure-derived features can accurately predict reaction yield (with R^2 values often exceeding 0.7) on HTE datasets generated under fixed reaction conditions, as exemplified by studies on Buchwald–Hartwig and Suzuki–Miyaura p...

Trajectory r67: The available head-to-head extrapolative evaluation on a large fixed-condition HTE dataset shows no statistically demonstrable advantage of GNNs over fingerprint/descriptor-based Random Forests; the observed differences are small or even favor the RF baseline, contradicting the hypothesis (sato2022p...

Closed-Loop Active Learning and Large-Batch Candidate Selection

Summary

Closed-loop active learning can turn large-batch assays into an adaptive design loop that reliably outperforms random screening while respecting synthetic constraints. For moderate batches, multi-objective Bayesian optimization with Pareto-based acquisitions selects compounds that are both active and makeable; for very large batches, scalable uncertainty-aware selection remains challenging, so diversification with greedy or scalarization portfolios is often preferred.

Background

High-throughput biology increasingly allows testing thousands of compounds per batch, but the bottleneck shifts to which molecules to make and assay next, especially when synthesis is limited to simple coupling reactions. The design space is discrete (reagents, building blocks, conditions) and objectives are multi-faceted (activity, yield/synthesizability), making it natural to combine categorical-aware Bayesian optimization, uncertainty-guided active learning, and diversity-aware selection. The central question is how to propose reaction components and products to synthesize and how to update models after each round so that the next batch most effectively advances both discovery and feasibility.

Results & Discussion

Iterative retraining on batched outcomes enables adaptive libraries that beat random selection across molecular property-learning tasks. In large-batch settings, uncertainty-prioritized policies such as query-by-committee (ensemble variance) reduce error substantially—on the 4M-compound Molecules3D corpus, a top-N ensemble-variance policy achieved ≈5× lower mean squared error than random or simple heuristics—while global variance-reduction methods (NIPV) distribute selections across clusters and accelerate mean absolute error reductions, reaching near-optimal accuracy with ~50% of labels on materials tasks [r5]. However, model-guided metrics degrade in very high-dimensional descriptor spaces and can be unstable early, at times allowing random to compete; safeguards include explicit diversity/density controls, global variance objectives, and dimensionality reduction before batch selection [r5]. From a computational standpoint, pool-wide scoring with ensembles is practical at scale, whereas classical expected-improvement or expected-error-reduction objectives are often too expensive for thousand-compound batches, limiting their use in practice [r5].

For moderate batch sizes (roughly q 50–100), multi-objective Bayesian optimization (MOBO) with Pareto-based acquisition functions provides an effective default. Expected hypervolume improvement (EHVI) and its noisy/parallel variants (NEHVI/qNEHVI) are designed to maximize the hypervolume—the volume in objective space dominated by the Pareto set—and inherently balance exploration and exploitation via the Gaussian process posterior; NEHVI is one-step Bayes-optimal for hypervolume and empirically outperforms scalarized EI baselines on chemistry/materials tasks without needing ad-hoc mean-weighted exploitation bonuses [r14]. When the design goal is “active and makeable,” integrating predicted reaction yield or synthesizability with predicted activity as a second objective, using EHVI or weighted-sum scalarization, improves convergence toward favorable trade-offs as measured by hypervolume gains and Pareto-front coverage [r7]. Open-source stacks support these workflows: EDBO (GP+qEHVI), BoTorch/GPyTorch (flexible GP MOBO, including qNEHVI), NEXTorch, and Gryffin (categorical/mixed-variable BO), enabling direct optimization over discrete reaction components with proper uncertainty modeling [r6].

In very large-batch regimes (q 500), acquisition-side scaling becomes the bottleneck. Exact Pareto-based qEHVI stagnates or fails at surprisingly small q, and even qNEHVI incurs steep wall-clock growth by a few hundred candidates; under fixed time budgets typical of high-throughput campaigns, portfolios of parallel weighted-sum scalarizations (e.g., qNParEGO) are more likely to achieve higher attained hypervolume than a single Pareto-based batch when $q > 1000$ [r15]. In prospective large-batch campaigns, groups often default to greedy ranking due to unreliable regression uncertainty and training-time constraints; when uncertainty is used, it commonly comes from MC-dropout for feed-forward networks, mean–variance heads for graph networks, or GP-based Thompson sampling—not a uniform standard of fingerprint DNN ensembles—underscoring operational heterogeneity at scale [r48]. To mitigate redundancy and maintain chemical coverage when selecting hundreds to thousands at once, scalable diversification layers are effective: Tanimoto-penalized greedy selection over ECFP fingerprints (e.g., an additive min-distance reward) has been applied to generate batches well above 500, and cluster-then-select pipelines using GPU k-means on 2048-bit Morgan fingerprints have selected thousands to tens of thousands of diverse leaders for procurement; notably, head-to-head large-q benchmarks against determinantal point processes are lacking, so method choice remains empirical at this scale [r41, r51, r63].

To propose coupling-compatible compounds and reaction components, categorical-aware BO frameworks directly optimize over discrete building blocks, ligands, catalysts, and conditions. Gryffin performs BO on categories using smooth approximations and descriptor-informed similarity, showing competitive or superior performance on discrete reagent selection, including ligand+process optimization for Suzuki–Miyaura reactions; GP-based EDBO similarly excels in mixed spaces, with Matérn-5/2 kernels preferred for continuous variables and Hamming kernels for purely categorical encodings when categories are exchangeable [r6]. Practical guidance includes replacing naive one-hot encodings with physicochemical descriptors or latent embeddings when chemical similarity matters and using ARD/SAAS priors to manage high-dimensional descriptor spaces; toolchains such as Gryffin, EDBO, BoTorch/GPyTorch, NEXTorch, Phoenics, GAUCHE, Hyperopt/TPE, and benchmarking suites (Olympus, poli/poli-baselines) provide ready implementations for enumerated coupling libraries and reaction optimization under constraints [r6]. These platforms support MOBO over activity and predicted reaction yield/synthesis success via EHVI/qEHVI or weighted sums, allowing the algorithm to discover favorable trade-offs without fixing weights a priori [r6, r7].

Updating predictions after each round follows a closed-loop protocol: retrain or warm-start the surrogate on accumulated labels, quantify uncertainty with scalable methods that have seen practical use at large batch sizes (ensembles, cross-validation, MC-dropout, or mean–variance heads), and choose the next acquisition policy appropriate to q and compute budget [r5, r48]. When redundancy emerges or the design space is high-dimensional, add explicit diversity or global-variance controls (e.g., NIPV, covariance-aware sampling, Tanimoto penalties, or cluster-based selection) and consider representation learning or dimensionality reduction to stabilize early rounds [r5]. For moderate batches, qNEHVI provides strong Pareto-front progress without ad-hoc exploitation terms; for extreme batches, parallel scalarization portfolios plus diversity filters and, if desired, bandit-style acquisition portfolios can improve coverage and robustness, consistent with evidence that Pareto-based acquisitions already encode exploration–exploitation and that scaling, not missing exploitation terms, is the dominant barrier at q   100 [r14, r15].

Trajectory Sources

Trajectory r5: Evidence supports uncertainty/committee-based and diversity-aware batch acquisition outperforming random and naïve exploitative orderings at thousand-scale selection, but the literature provided does not yet establish that adding explicit exploitative scoring (e.g., expected improvement/high predict...

Trajectory r6: The literature supports the hypothesis that categorical-aware BO frameworks such as Gryffin are well-suited for optimizing discrete reaction components and are especially effective when augmented with physicochemical descriptors and mixed-variable integration. (hase2021 pages 1-2, ...

Trajectory r7: The surveyed literature supports the hypothesis that integrating predicted reaction yield as a second objective—using either Pareto-based methods such as EHVI or weighted-sum scalarization—enhances the selection of compounds that are both active and synthetically accessible (du2025advancinggeneticen...

Trajectory r14: The hypothesis is supported: in chemistry/materials MOBO, Pareto-based acquisitions (EHVI/NEHVI, ParEGO variants) already encode the exploration–exploitation trade-off, and practice improves batch performance via noisy/parallel variants, portfolio/DPP/EA hybrids, and lookahead/cost-aware designs—not...

Trajectory r15: Given the available evidence on scaling and wall-clock behavior, the hypothesis is supported: under fixed wall-clock time with q 1000, a portfolio of parallelized weighted-sum scalarizations is more likely to attain higher Pareto-front hypervolume than a single batch Pareto-based acquisition such ...

Trajectory r41: The hypothesis is supported by at least one explicit large-scale pipeline that selects >1000 compounds via clustering-based representative selection using Morgan fingerprints, though other large-scale screens in the provided context prioritized large batches without clustering-based diversification ...

Trajectory r48: The hypothesis is not supported: in large-batch ($q > 500$) prospective closed-loop campaigns, published studies do not converge on DNN ensembles over fingerprints with ensemble-variance UQ as a standard; instead they use a mix of RF/FFNN/MPNN surrogates with fingerprints or learned graph embeddings a...

Trajectory r51: The hypothesis is supported: at least one peer-reviewed study explicitly implements and applies a greedy sequential diversification algorithm with Tanimoto/ECFP similarity penalties to select very large batches ($q > 500$), and multiple contemporaneous works describe the same greedy, iterative re-scor...

Trajectory r63: The provided sources do not document any study that benchmarks a DPP-based selection (exact or greedy) against Butina/leader-follower or Tanimoto-based MaxMin/MaxSum for large-batch ($q > 1000$) molecular library selection.

Machine-Learned Recommendation of Reaction Conditions for Coupling Reactions

Summary

Machine learning models that only “see” the two coupling partners do not reliably recommend bases or solvents beyond a naïve popularity baseline, whereas models that encode the full reaction graph (reactants and products) deliver substantially better top-k condition recommendations. For high-throughput coupling campaigns, this points to using reaction-level encodings, balanced high-throughput datasets with negative outcomes, and iterative retraining after each experimental batch to debias and improve recommendations.

Background

Carbon–carbon and carbon–heteroatom coupling reactions underpin rapid library synthesis for activity screening, but their success hinges on selecting compatible catalysts, bases, and solvents. As screens scale to thousands of wells per week, manual condition selection is a bottleneck; data-driven recommenders promise to prioritize viable conditions at scale. However, reaction data mined from the literature are biased toward positive results and popular conditions, and most reports omit failed experiments, which complicates learning causal structure–condition relationships. Aligning models, data representations, and evaluation metrics with the realities of many-to-many reaction–condition mappings is therefore central to achieving practical gains in coupling campaigns.

Results & Discussion

The central finding is that condition recommenders trained only on 2D encodings of the two coupling partners do not surpass a naïve popularity baseline on heteroaryl Suzuki–Miyaura reactions, whereas richer reaction-level encodings do. In a direct test, Beker et al. trained feed-forward neural networks, graph convolutional networks, and a positive-unlabeled (PU)-corrected network on >10k heteroaryl/aryl–heteroaryl Suzuki–Miyaura reactions using only substrate features (Morgan fingerprints, RDKit descriptors, or autoencoder embeddings) to classify bases and solvents (e.g., 7×13 or 7×6 label grids) with five-fold cross-validation and ranked outputs; the models “fail to perform significantly better than [a] naïve [popularity] baseline,” with solvent top-1 accuracies often <50% and top-3 close to baseline, and base accuracy inflated by label skew toward carbonates [r53, beker2022]. These results implicate literature bias and the absence of negative examples as dominant factors that limit reactant-only models to recapitulating popularity rather than causative chemistry [r53, beker2022].

In contrast, condition recommendation improves when models encode the full reaction graph (reactants plus products) or reaction fingerprints and treat condition selection as a contextual, multi-component prediction problem. A reaction-graph neural network (AR-GCN) that embeds reactants and products reports 31–42% top-1 improvements over a popularity baseline for catalyst and solvent recommendations across Suzuki, Negishi, and C–N couplings, indicating that reaction-level context helps recover actionable condition–structure relationships beyond frequency effects [r53, ball2025]. A complementary approach that uses reaction and product fingerprints to sequentially predict catalyst, solvent(s), reagents, and temperature achieves 69.6% top-10 “close-match” for full reaction contexts over Reaxys (a benchmark that includes coupling classes), further supporting sequential, context-aware modeling for many-to-many mappings [r53, gao2018]. Related work using a pretrained graph neural network over reactants and products reports per-component accuracies of 59% (catalyst) and 42% (solvent), again exceeding naïve baselines while reflecting the relative difficulty of solvent recovery [r53, ball2025].

The reported metrics compare ranked recommendations against observed conditions under standardized baselines. The popularity baseline selects the most frequent condition(s) in the training data irrespective of structure; “top-k accuracy” is the fraction of test reactions whose true label appears in the model's top-k recommendations, and "top-1 improvement over baseline" quantifies the absolute gain relative to that frequency-only strategy [r53, ball2025]. Fivefold cross-validation with held-out reactions was used to estimate generalization for substrate-only models; outputs were ranked lists of bases and solvents for each test reaction [r53, beker2022]. The "top-10 close-match for full contexts" metric evaluates whether all context components (e.g., catalyst, solvent(s), reagent, temperature) are simultaneously retrieved within the top-10 lists, reflecting the joint compatibility requirement of multi-component condition sets [r53, gao2018]. Two practical caveats recur: (i) label skew (e.g., carbonate bases) can inflate base accuracy without reflecting broad generalization, and (ii) solvent recovery is harder because labels are diverse and literature reports under-sample negative outcomes, encouraging models to learn popularity rather than causality; PU-corrected training can partially address missing negatives but did not overcome these limitations in the reactant-only setting [r53, beker2022].

For high-throughput coupling campaigns that must deliver thousands of compounds per week, three implications follow. First, condition proposal should use reaction-level encodings—reaction SMILES or reactant+product graphs—and model the context sequentially (e.g., catalyst → base → solvent), approaches that have demonstrated substantive gains across Suzuki, Negishi, and C–N couplings relative to popularity baselines [r53, gao2018]. Second, to avoid learning publication bias, each weekly batch should log both successes and failures and aim for balanced coverage across condition families; assembling a robotized HTE dataset with negative examples is explicitly highlighted as necessary to unlock gains even for substrate-only models [r53, beker2022]. Third, after each round, retrain or fine-tune the recommender on the cumulative HTE data, track top-k versus the popularity baseline, and consider coarsening solvent/base labels by physicochemical descriptor groupings early on to stabilize learning—an approach hypothesized to raise solvent top-1 by 10% on Suzuki–Miyaura couplings that can be tested with fivefold cross-validation and bootstrap confidence intervals as data accrue [r53]. In data-sparse phases, PU-corrected training can mitigate missing negatives, but as HTE data accumulate, reaction-graph models with sequential prediction should be prioritized for robust, generalizable condition selection to drive coupling-based library synthesis [r53, gao2018].

Trajectory Sources

Trajectory r53: The hypothesis is not supported: the only published models that strictly use only the 2D structures of the two reactants for coupling reactions (heteroaryl Suzuki–Miyaura) yield base/solvent recommendations that do not meaningfully outperform popularity baselines, whereas higher-performing coupling-...