A Systematic Informatics Approach to Synthetic Chemistry

by

Babak A. Mahjour

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Medicinal Chemistry)
in The University of Michigan
2023

Doctoral Committee:

Assistant Professor Tim Cernak, Chair
Professor Charles Brooks
Professor John Montgomery
Associate Professor Andrew White

Babak A. Mahjour
bmahjour@umich.edu
ORCID: 0000-0002-8225-6514
© Babak A. Mahjour 2023

DEDICATION

This work is dedicated to my family, whom I love dearly, and to my friends, with whom I don't interact or see as often as I would like.

ACKNOWLEDGEMENTS

This work would not have been possible without Tim Cernak’s thoughtful and considerate supervision. The Cernak lab was the ideal place to foster my scientific development. I owe much of my accomplishments to the support of my lab mates, who were always willing to help. I also acknowledge the College of Pharmacy and the Department of Medicinal Chemistry for their caring oversight, and the University of Michigan for being my home for the last 10 years. Finally, I thank the American Chemical Society’s Division of Medicinal Chemistry for my selection as one of their predoctoral fellows.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vi

LIST OF APPENDICES xxiv

ABSTRACT xxv

Chapter I. Introduction 1

Chapter II. Reaction Targeting 7

A map of the amine-carboxylic acid coupling system 11

Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing 39

Reinforcing the Supply Chain of Umifenovir and Other Antiviral Drugs with Retrosynthetic Software 47

Analyzing Conceivable Reactions from Commercially Available Chemical Building Blocks 68

One Step Retrosynthesis of Drugs by Cross-Coupling from Building Block Catalogs 80

Chapter III. High Throughput Experimentation 94

Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery 96

Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis 121

Miniaturization of Popular Reactions from the Medicinal Chemists' Toolbox for Ultrahigh-Throughput Experimentation 146

Development of copper-catalyzed deaminative esterification using high-throughput experimentation 154

Expansion of NanoSAR for the Deconvolution of Large Compound Libraries 171

Chapter IV. Reagent Discovery 175

Designing Chemical Reaction Arrays using phactor and ChatGPT 177

Weighted Reaction Fingerprints for Visualizing Reactivity Cliffs and Generality 198

Reaction Array Analysis and Design via Reagent Networks 218

Chapter V. Molecular Compression 230

Molecular sonification for molecule to music information transfer 232

Molecular Featurization via Musical Additive Synthesis 255

Chapter VI. Informatics Tools and Software for Education and Automated Research Workflows 259

Interactive Python Notebook Modules for Chemoinformatics in Medicinal Chemistry 260

LRphase: an efficient algorithm for haplotagging long reads 281

An Automated and Colorful PAMPA Assay To Investigate Drug Permeability 285

Chapter VII. Software Packages 289

phactor 291

phactor inventory 298

samples 304

retrosynthetic visualizer 309

Chapter VIII. Conclusion 316

APPENDICIES 318

Appendix A 319

BIBLIOGRAPHY 338

LIST OF FIGURES

1-Overview | The systematic informatics approach to synthetic chemistry consists of four main concepts. High value reactions are identified and can be invented with high throughput experimentation. Algorithms and computational representations of molecules enable reagent discovery to optimize reactivities and expand substrate compatibility. 1

2-Fig. 1 | Transformation enumeration strategy and notation. a, Ethylamine (**1**) and propanoic acid (**2**) can couple to form amide **3**, but can also couple to form 79 other products, including **4**–**9**. **b**, Enumerating all combinations of sp^2 or sp^3 hybridization for the 80 coupling patterns yields 320 product substructures. **c**, A notation system for classifying transformations; see also 6-Extended Data Fig. 1. 20

3-Fig. 2 | Substructure search of 320 amine–acid coupling transformations within 9,279 complex molecules from DrugBank. Each line represents the appearance of a product substructure of a transformation in a complex molecule, and the colour of the line represents the frequency of occurrence of that substructure in that molecule. The dots around the periphery denote which specific transformations appear in complex molecules (+)-noscapine (green dots), (−)-quinine (purple dots) and (−)-sitagliptin (blue dots), which connect to 112, 96 and 55 transformations, respectively. The numbers around the periphery can be matched to a full list of transformation notation labels found in Extended Data Table 1. 21

4-Fig. 3 | Experimental exploration of the sp^2–sp^2 amine–acid coupling transformation space. a, Products with different property profiles (for example, basic, acidic, neutral, lipophilic) can be produced from the same two starting materials, **10** and **11** (top), by varying the transformation and reaction conditions. NfOH, nonafluorobutanesulfonic acid; HFIP, hexafluoroisopropanol; Cp*, pentamethylcyclopentadiene; 1,10-phen, 1,10-phenanthroline. **b**. Kernel density estimation plots show the range of accessible molecular weight, partition coefficient (log*P*) and polar surface area by coupling **10** to **11** via various amine–acid coupling transformations. Grey lines denote the molecular weight, log*P* and polar surface area of **12**–**17**. The calculations use **13** and **15** in their charged protonation state. 23

5-Fig. 4 | Late-stage diversification. Various transformations enable the diversification of the complex molecules **18**, **22** and **27**. We performed a virtual enumeration of other complex molecules—shown in 14-Extended Data Fig. 9—wherein the full transformation set was enumerated for four complex molecule pairings to demonstrate that a wide range of properties can be accessed, depending on which transformation is selected. LiHMDS, lithium bis(trimethylsilyl)amide; HATU, 1-[bis(dimethylamino)methylene]-1*H*-1,2,3-triazolo[4,5-*b*]pyridinium 3-oxide hexafluorophosphate; DIPEA, *N*,*N*-diisopropylethylamine; NHPI, *N*-hydroxyphthalimide; PMHS, polymethylhydrosiloxane; LDA, lithium diisopropylamide; TMSCl, trimethylsilyl chloride; NBS, *N*-bromosuccinimide......... 24

6-Extended Data Fig. 1 | Workflow for enumeration of amine–acid transformations. For a pair of coupling partners, we consider a reaction at the functional group A (amine) and B (carboxylic acid oxygen, B[O] or carbon, B[C]). Deamination reactions are noted as −A and decarboxylation reactions are noted as −B. Enumeration following steps 1–3 produces 320 transformations. For the enumeration of all *syn*- and *anti*-diastereomers (step 4), consult also 11-Extended Data Fig. 6 25

7-Extended Data Fig. 2 | Kernel density estimate plots for 320 conceivable amine–acid coupling transformations. Distribution of common physical properties from the achiral amine–acid coupling of ethylamine, ethenamine, propanoic acid and acrylic acid. MW, molecular weight; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; PSA, polar surface area; FSP3, fraction sp^3; ROTB, rotatable bonds; FC, formal charge; QED, quantitative estimate of drug-likeness. 27

8-Extended Data Fig. 3 | Number of DrugBank hits per transformation. This bar chart shows how many times a transformation is found in the DrugBank database. Each number on the abscissa maps to a transformation listed in Extended Data Table 1. 28

9-Extended Data Fig. 4 | Decarboxylative transformations from the enumeration scheme. Decarboxylative reactions that produce an amine bound to an sp^3 or sp^2 carbon chain appear in high frequency. These reactions can be used to synthesize a large number of drugs contained in DrugBank. Each transformation can be found by its corresponding number in Extended Data Table 1. The colour scale is the same as in 3-Fig. 2. rxn, reaction. 29

10-Extended Data Fig. 5 | Transformations from the enumeration scheme found in specific drugs. The chord diagrams show connectivity of transformation substructures as retrosynthetic disconnections in target molecules, with red and blue dots highlighting the transformations shown at left in each panel. **a**, Noscapine connects to 112 of the transformations. **b**, Quinine connects to 96 transformations. **c**, Sitagliptin connects to 55 transformations. The colour scale is the same as in 3-Fig. 2. 31

11-Extended Data Fig. 6 | Enumeration of regioisomers and diastereomers. a, The transformation substructures enumerated in 4-Fig. 3 are from the 320 achiral bond arrangements available from coupling **1**, **2** and their sp^2 variants ethenamine and acrylic acid. **b**, To sample three-dimensional and regiochemical space, a β′ substituent was added as a differentiating substituent. The β′ substituent may be any substituent, but is enumerated as being distinct from the β substituent. Considering this regiochemical enumeration increases the 320 achiral coupling transformations to 588. **c**, Subsequent enumeration of all possible diastereomers leads to 1,005 chiral coupling transformations. These 1,005 three-dimensional substructures were used as inputs in the principal moment of inertia plot in 12-Extended Data Fig. 7. 32

12-Extended Data Fig. 7 | Principal moment of inertia plot of 1,005 amine–acid coupling transformations incorporating stereochemistry and regiochemistry. In this expanded three-dimensional space, regiochemistry and stereochemistry of the transformations were considered. A total of 1,005 ways to connect an amine to an acid were found. The products presented a diversity of properties and three-dimensional shapes. Each molecule is coloured by its quantitative estimate of drug-likeness. 33

13-Extended Data Fig. 8 | High-throughput experimentation for the discovery of a copper-promoted esterification reaction. a, An esterification reaction discovered through reaction screening of transition metals with ligands and additives. **b**, Recipe and well mapping. **c**, Calibration curve, for product **17** versus caffeine internal standard, used to convert the ultraperformance liquid chromatography with ultraviolet–visible spectrometry peak area to concentration, and thus to assay yield. Error bars show deviation among triplicate injections. **d**, Heat map depicting assay yield screening results. CuI with $AgNO_3$ and pyridine showed the most promising results, achieving 18.5% assay yield using 30 mol% CuI with $AgNO_3$........ 34

14-Extended Data Fig. 9 | Kernel density estimate plots of a series of complex molecules as substrates in the amine–acid coupling system. The amine–acid pair depicted was used as an input to combinatorial enumeration, and the number of valid products identified is noted for each pairing. Distributions of common physical properties are shown for each coupling set. Abbreviations are as in 7-Extended Data Fig. 2. 37

15-Extended Data Table 1 | Transformation labels. This table maps each transformation number from the periphery of the chord diagram in 3-Fig. 2 to a transformation label. 38

16-Fig. 1 | a. Given an amine **1** and carboxylic acid **2,** the most popular transformation to unite this pair of building blocks produces the amide **3**. **b.** Coupling products arising from a curated subset of chemical transformations charted by our prior work. **c.** Examples of transformations excluded from our prior work. 40

17-Fig. 2 | a. Demonstration of matrix-encoding of reaction from panel **1a**. In the molecular diagrams, white circles represent carbon, blue circles nitrogen, and red circles oxygen. The numbers in the adjacency matrix correspond to the atom indices in the cartoon atoms, while the color represents bond order. **b.** Workflow for exhaustive generation of amine–acid coupling products. **c.** The difference between the product matrix and starting material matrix is the transformation matrix. 41

18-Fig. 3 | a. Schematic of enumeration from amine **1** and acid **2** to yield 56 million unique transformation matrices, which are filtered first to 222,740 unique products assuming carbon and oxygen atoms are degenerate, and further to 80,941 unique products after eliminating highly improbable structures. **b.** 2-dimensional histogram showing distribution of ring count and bond edit distance of the initial 222,740 products. **c.** Kernel density estimate (KDE) plots of various physiochemical properties of the expanded amine-acid coupling system, along with selected products. The respective property of the classic amide is shown by the vertical grey line. HBD = hydrogen bond donor, PSA = polar surface area, FSP3 = fraction *sp*^{3}-atoms, MW = molecular weight, HBA = hydrogen bond acceptors, QED = quantitative estimate of drug-likedness, LogP = partition coefficient, ROTB = number of rotatable bonds, Rings = number of rings. **d.** Principal Moment of Inertia (PMI) ratio distributions of all products from the expanded enumeration........ 42

19-Fig. 4 | PMI ratio plots showing shape space distribution of reaction enumerated late-stage diversification of two drug-like molecules, **a.** showing an *sp*^{3}–*sp*^{3} coupling, and **b.** showing an *sp*^{2}–*sp*^{2} coupling. For each pair of plots, the left shows the distribution of full coupling products, while the right shows the distribution of only the atoms at the reaction center (*cf*. 18-Fig. 3d). **c.** Joint distribution plots of physicochemical property modulation within the system depicted in **a** (**20** + **21**). See Fig. S2 in Supporting Information for corresponding plots for **22** + **23**. The x-axes show modulation of the entire molecule relative to the amide, while the y-axes show modulation of only the atoms at the reaction center. HBD = hydrogen bond donor, PSA = polar surface area, FSP3 = fraction *sp*^{3}-atoms, MW = molecular weight, HBA = hydrogen bond acceptors, QED = quantitative estimate of drug-likedness, LogP = partition coefficient, ROTB = number of rotatable bonds, Rings = number of rings. 43

20-Fig. 5 | UMAP projection of reaction fingerprints from simple amine–acid starting materials onto 80,941 enumerated products onto DrugBank. Dots are colored by the number of product substructure matches in drugs with the majority of product substructures appearing as unexplored chemical space. 45

21-Fig. 6 | Chord diagram showcasing connectivity of two drugs to enumerated amine–acid coupling products. Drug targets lie on the black band, and the exhaustive matrix enumeration products lie on the checkered yellow band. The color of the chord represents the substructure's frequency of occurrence, ranging from one to greater than ten (see Figure S7 for further discussion). Two disconnections are shown for athamontanolide (**29**), both into acrylic acid (**30**) and an amine (**31, 32**) with new disconnections generated by the current enumeration workflow. Similarly, noscapine (**33**) can be disconnected to amine **34** and acid **35**, or to alkyl acid **36** and secondary amine **37**. The frequencies of the top 100 most abundant substructure matches in DrugBank are plotted as a histogram, categorized by their elemental makeup. The top three most abundant structures for each group are shown (**38**–**49**). 45

22-Fig. 1 | Workflow to identify alternate starting material supply chains. 1. Published starting materials are encoded as SMILES strings, which are 2. excluded from automated retrosynthetic searching. 3. Select routes are validated in the wet lab. 59

23-Fig. 2 | Therapeutics being evaluated for COVID-19 and their retrosynthetic analysis networks. The number of clinical trials is based on search results for all listed trials, completed, active, or planned, found on www.clinicaltrials.gov (accessed July 7th, 2020). For the route networks, the yellow dot at the center is the target molecule. Routes in grey and orange below the target are published, and the routes in purple and pink above the target are routes predicted. Intersection nodes in literature routes reflect common intermediates or starting materials. A version of this diagram is available as an interactive route visualizer. 60

24-Fig. 3 | Predicted synthetic sequence to 4. Reaction conditions are proposals from the SYNTHIA™ software. 63

25-Fig. 4 | Experimental demonstration of four routes to 2. Each route terminates in inexpensive starting materials. **26**, **41**, **45**, **48**, and **49** have not been used in a published synthesis of **2**. The conversion of **33** into **35**, and the use of starting material **49**, instead of 5-bromo-2-chloro-1-nitrobenzene, were manual modifications to the software's prediction. See main text for details. 64

26-**Fig. 5 | Predicted and realized routes to 3.** Dashed arrows are predicted, while the solid arrow is manually designed and experimentally validated, with the subsequent addition of the key reaction to the SYNTHIA™ database. 67
27-Graphical Abstract | Enumerative combinatorics is taken beyond amines and acids to all common building block types. 68
28-Fig. 1 | Analysis of building blocks from the MilliporeSigma catalog. (A) Popular reactions in medicinal chemistry and popular building block classes. (B) Comparison of scope between our prior work and current work. (-X = halides, -BX_n = boronates) (C) Distribution of heavy atoms in commercially available amines and acids...76
29-Fig. 2 | (A) Chord diagram linking 399 conceivable boronate–bromide transformations to 15 antiviral drugs. A transformation (yellow arc) to drug (black arc) linking lines indicate the corresponding substructure produced by the transformation exists in the respective drug. (B) All conceivable reactions (222,740 transformations) between an amine and acid provides considerable access into drug-like chemical space. (C) Distribution of transformations in examined drugs from all conceivable transformations between building blocks. (D) Number of transformations between building blocks that exist in examined drugs. 77
30-Fig. 3 | Scatter plot of the commercial availability of building blocks against the transformation frequency of substructures formed found when enumerating transformations for the respective building block. High value transformations identified on the Pareto front are exemplified. Isolated yields are shown. Conditions: (A) $NaBH_4$ (1 equiv.), *p*-toluenesulfonic acid monohydrate (1 equiv.), 25 ºC, 0.5 h. (B) 1) triphenylpyrylium (1.0 equiv.), ethanol (1.0 M), 80 ºC 2) benzaldehyde (1.0 equiv.), Cs_2CO_3 (1.5 equiv.), DMSO (0.1 M), 60 ºC 3) $InBr_3$ (1.0 equiv.), chloroform (0.1 M), 60 ºC. 4) $NaBH_4$ (1.0 equiv.) (C) $Pd(OAc)_2$ (0.1 equiv.), XantPhos (0.1 equiv.), KO^{t}Bu (2.0 equiv.), toluene (0.2 M), 80 ºC, 18 h. (D) KO^{t}Bu (2.0 equiv.), DMSO (0.2 M), r.t., 18 h. (E) ^{t}BuOO^{t}Bu, 130 ºC, 40 h. 78
31-Fig. 4 | Recently reported amine–acid chemistries analyzed for chemical space access. On the left are Venn analyses of commercially available sp^3 and sp^2 amines, bromines, boronates, and carboxylic acids indicating novel chemical space reachable by recently discovered amine–acid chemistry. The t-SNEs to the right compares the structural access provide when using acids or amines instead of bromides or boronates. (A) The analysis performed for an sp^3–sp^3 deaminative–decarboxylative C–C cross coupling. (B) The analysis performed for an sp^3–sp^2 deaminative–decarboxylative C–C cross coupling. 79
32-Fig. 1 | An analysis of commercial building blocks and reported drug structures. a) A Pareto front cartoon exemplifying the value of certain building blocks. Building blocks that are highly available and can be used to form drugs are of high value. **b)** Enumerative combinatorics can identify target bonds between building blocks. In this study we focus on cross-couplings that form one bond after the elimination of both building block handles. **c)** Retrosynthesis of target and its representation as a chord diagram. **d)** Prevalence of bond types in DrugBank compounds. The alkyl-alkyl C–C bond is the most common bond type found, followed by the alkyl-alkyl C–N and C–O bonds and the alkyl-aryl C–C bond. **e)** Prevalence of building block classes in the Sigma Aldrich commercial catalog. 86

33-Fig. 2 | C–H motifs can be viewed as valuable and versatile building blocks. a) Pareto front and building block analysis of the commercial dataset. Classes of building blocks that have many commercial substrates and can be found often in one step retrosyntheses of drug compounds lie on the Pareto front. Core structures of benzyl–H and benzyl–amine, benzyl-halide, and benzyl-alcohol compounds in the commercial dataset are compared in the Venn diagram. Many benzylic compounds do not exist as their amine/halide/alcohol functionalized counterpart. **b)** Examples of couplings using unconventional building blocks. 87

34-Fig. 3 | a) All hypothetical single bond C–H cross couplings between benzylic carbon atoms and nitrogen atoms to form DrugBank compounds from commercial substrates. A benzylic functionalization previously reported by the Stahl group is exemplified as a reaction identified in the analysis.[138] **b)** All hypothetical single bond C–H cross couplings between benzylic carbon atoms and oxygen atoms to form DrugBank compounds from commercial substrates. Another example from the Stahl group is showcased here as a reaction found by the targeting algorithm.[139]. 89

35-Fig. 4. **| One step cross coupling maps for various activated C–H motifs.** In the first row and column, synthon a is grouped by commercial compounds that react at a carbon alpha to a carbonyl group with another alkyl building block. In the third row and first column, synthon a is grouped by compounds that couple at a benzylic hydrogen with another building block, such as benzyl alcohol **13** reacting with **14** to form drug **15**. 90

36-Fig. 5 | a-g) Examples of one step syntheses found in the analysis. **f)** The synthesis of bromhexine identified in this analysis was previously validated. 91

37-Fig. 6 | Single step cross coupling maps between DrugBank and purchasable compounds in Sigma-Aldrich's catalog, trellised by synthon A building block and bond formed. In the first column, both synthons link at an alkyl carbon. In the second column, synthon A, which is the building block group for the row, is alkyl and the remaining synthon b is aryl. In the third column, the building block arc contains aryl synthons and synthon b contains alkyl synthons. In the final column, both synthons are aryl. Compound numbers refer to retrosynthetic reactions shown in 36-Fig. 5. An example of a one-step synthesis of bromhexine experimentally validated is shown above and indicated on the bromide aryl-alkyl C-c bond chord plot.[54] 93

38-Fig. 1 | Overview of the phactor™ software. a Anatomy of a reaction as encoded by phactor™. **b** High-level software workflow of phactor™. Reaction arrays are designed from chemical inventories and reaction informatics. Resultant data is stored in delimited text (CSV) or in a relational database (SQLite3). phactor™ can convert results to Open Reaction Database (ORD)[149] and Chemical Description Language (XDL)[178] and is readily compatible with optimization programs such as EDBO+[176] and LabMate.ML[177]. **c** Workflow of phactor™. Once the reaction array is designed, phactor™ provides human-readable or machine instructions to execute the dosing manually or robotically (UPLC ultra-performance liquid chromatography). **d** phactor™ supports custom volumes allowing for reaction arrays to be performed at any scale. At a minimum, the hardware needed to execute a reaction array is an autopipette and an array reactor block. **e** phactor™ facilitates the design and execution of ultraHTE in 1536 wellplates. 113

39-Fig. 2 | Reaction arrays executed with the phactor™ software. The reaction array design and results are shown here as displayed on phactor™. Colour bars adjacent to compound numbers correspond to the colour bars in the reaction array design grid generated by phactor™. Product/internal standard ratios are calculated using the observed UV-derived peak area, while assay yields account for differences in product absorptivity by calibrating to authentic samples of products. **a** Preliminary esterification hit leading to publication[180]. **b** Optimized oxidative indolization conditions towards the synthesis of umifenovir[54]. **c** Allylation catalyst/ligand concentration ratio and base reaction array analysed by conversion and selectivity. 115

40-Fig. 3 | phactor™ has been utilized in a variety of synthetic campaigns. a–**i** Chemistries discovered via reaction arrays designed with phactor™. All input and output files used to produce reaction arrays (**a**–**f**) are provided via an online repository in addition to all compiled HTE results in a machine-readable format. Reaction schemes can be found in the “Selected screening examples” section of the Supplementary Information. 117

41-Fig. 4 | phactor™ facilitates ultraHTE direct-to-biology campaigns. a Event workflow for performing ultraHTE using phactor™ and a Mosquito robot. **b** Design of 1280 well amide coupling plate. 80 amines were selected to react with carboxylic acid **26**. Eight conditions were run in duplicate for each amine. **c** Results of the amide coupling are shown as a product/internal standard integration ratio from a 2-min LCMS injection of each well. The Mosquito robot is utilized to split the size 1536 plate into four sizes 384 plates for LCMS and bioassay analysis. **d** Percent inhibition of SARS-CoV-2 M^{Pro} when treated with a sample of the reaction mixture from the corresponding well. The 1280 plate is visually recreated. **e** IC_{50} curves for three scaled-up compounds chosen from the reaction array. Compounds **27**–**29** display a range of assay and inhibitory responses. 119

42-Fig. 5 | The six stages in the phactor™ workflow. Each stage is progressed sequentially. With an input reagent CSV, reaction arrays can be designed in seconds. Once the experiment has been executed, a standardized output can be downloaded on the report stage. 119

43-Fig. 6 | phactor™ enables rapid machine-learning analysis of multiple reaction arrays in tandem. Standardized output files can be rapidly merged to create massive datasets. Shown is a tSNE (t-distributed stochastic neighbour embedding) of all products made in the decarboxylative–deaminative sp^3–sp^3 C–C coupling detailed in ref. [181], coloured by average product/internal standard. 120

44-Fig. 1 | (a) HTE was used to develop reaction conditions for a key S_NAr coupling. (b) A first survey of four solvents and six bases revealed several productive conditions for the coupling of **1** and **2** to form **3**. Among the best conditions from this screen, the use of $NaHCO_3$ in NMP was revealed as the most robust set of conditions through subsequent parallel synthesis studies (not shown). (c) Registration of compounds with the 2-aminopyridine motif on the DGAT1 program accelerated after the day that the screen was run (red line). (d) The availability of these reaction conditions enabled hit expansion on a series of DGAT1 inhibitors, culminating in key compounds such as **3**–**7**. 139

45-Fig. 2 | (a) Miniaturized reaction exploration on a milligram scale was a key enabling tool in the late-stage functionalization of medicinal compounds. (b) A survey of ligands for the iridium-catalyzed C–H borylation enabled the selective functionalization of **8**. Abbreviations: dtbpy = 4,4′-di-*tert*-butyl-2,2′-bipyridine; $P(C_6F_5)_3$ = tris(pentafluorophenyl)phosphine; Me_4Phen = 3,4,7,8-tetramethyl-1,10-phenanthroline; dbfa = (*E*)-*N*,*N*-dibenzyl-*N*′-(pyridin-2-yl)formimidamide; MeDPA = *N*-methyl-*N*-(pyridin-2-yl)pyridin-2-amine; ^{t}dtbpp = tris(2,4-di-*tert*-butylphenyl) phosphite. 140
46-Fig. 3 | Development of ultraHTE. Chemical synthesis was performed in 1536-well plates. (A) Complex medicinally relevant molecules such as **14** could be synthesized on the gram scale using air-sensitive catalysts that were synthesized under conditions determined on the nanomole scale. (B) SPT Labtech mosquito. (C) Heat map of reaction performance for 1536 cross-coupling reactions on complex substrates. 141
47-Fig. 4 | (a) Use of ultraHTE to explore the substrate scope of cross-couplings on isolated palladium oxidative addition complexes (OACs). (b, c) Exploration of the substrate scope (b) on the nanoscale (c) highlights that OACs are robust coupling partners, with product observed by UPLC–MS for 54% of 384 nucleophiles attempted, compared with 10% using complementary catalytic conditions for the same library of nucleophiles. 142
48-Fig. 5 | Nanoscale synthesis and affinity ranking using ASMS analysis of crude ultraHTE reaction mixtures. A series of known CHK1 inhibitors were used as benchmarking compounds (right ordinate). Compound **25** was readily observable as a molecule with higher affinity for CHK1 than **23** and **24** based on protein titration affinity ranking. 143
49-Fig. 6 | phactor is a high-throughput experimentation design software. The reaction array design tool is shown. 144
50-Fig. 7 | HTE screen of transition metals and ligands, designed in phactor, to realize a deaminative esterification. 145
51-Fig. 8 | A systems chemistry approach that links chemical transformations and reaction conditions to structure, properties, and molecular function is enabled by ultraHTE. 145
52-Fig. 1 | Popularity of common reactions in the synthesis of pharmaceuticals. 147
53-Fig. 2 | Nanoscale ultraHTE Suzuki coupling reaction performance.a, Boronates (**1**–**10**) and halides (**11**–**22**) used in the study. b, Heatmap of conversion to product as determined by UPLC–MS relative to an internal standard at $\lambda = 254$ nm. c, Plate map. d, Box-plots comparing the effects of base, cosolvent and catalyst on reaction performance (black circles indicate outlying points). Bpin, 4,4,5,5-tetramethyl-1,3,2-dioxaborolane; Cbz, benzyloxycarbonyl; tAm, tert-amyl; r.t., room temperature; PROD/IS, product/internal standard; aq., aqueous. 148
54-Fig. 3 | Nanoscale reaction optimization for the reductive amination of staurosporine (**36**) with eight aldehydes, eight ketones and 48 reaction conditions. See Supplementary Section 5.1 for full list of aldehydes and ketones. Relative percentage conversion was determined by UPLC–MS at $\lambda = 254$ nm. Conc., concentration. 149

55-Fig. 4 | Nanoscale staurosporine analogue library synthesis via reductive amination. a, Heatmap of conversion to product for the reductive amination of staurosporine **36** with 48 aldehydes using eight reaction conditions. A representative example of aldehydes is shown (see Supplementary Section 5.2 for full list). Relative conversion was determined by UPLC–MS at λ = 254 nm. b,c, Selected reactions were repeated on 50 μmol scale giving **45**–**50** and S1 (b) or 0.30 mmol scale giving **51**–**53** (c). Yields are of purified products. 151

56-Fig. 5 | Two-step nanoscale synthesis of MK2-inhibitor analogues (**56**) and (**57**) via N-alkylation followed by Boc-deprotection. a, Electrophiles **58**–**69** were coupled to **54** and **55**. b, Heatmap of reaction performance determined by UPLC–MS at λ = 254 nm relative to standard curves created from isolated products. c, Selected examples **70**–**73** repeated on 40 μmol scale (see Supplementary Section 6.5 for full scope). Yields are of purified products. LiHMDS, lithium bis(trimethylsilyl)amide. 152

57-Fig. 1 | HTE enabled amine–acid esterification and its application in drug molecules. A The amine–acid esterification. **B** Esters are an important medicinal functionality as in drugs **1**–**3**. **C** Reaction discovery, optimization, and scope profiling is achieved using HTE. 156

58-Fig. 2 | Physicochemical properties of amides and esters and complementary chemical space of anilines to phenols. A An amine and carboxylic acid can be coupled in a variety of transformations beyond the traditional amide coupling, including the esterification reaction developed here, which gives a unique property footprint. **B** Venn diagram showing the complementarity of phenol to aniline building blocks available in the MilliporeSigma catalog. **C** tSNE analysis showing the complementarity of ester products generated by coupling aspirin to phenol (blue dots) versus aniline (yellow dots) building blocks available in the MilliporeSigma catalog. **D** Ester products have a higher predicted blood–brain permeability (CNS-pMPO score) than classic amide products produced from the same set of anilines when coupled to the drug metaclopramide. 156

59-Fig. 3 | Reaction discovery and profiling with HTE. A General deaminative esterification reaction. **B** Heatmap showing the discovery of the esterification reaction in 24 glass microvials. Dtbpy = 4,4′-di-*tert*-butyl-2,2′-dipyridyl, xantphos = 4,5-bis(diphenylphosphino)-9,9-dimethylxanthene. **C** Heatmap showing optimization of esterification in 96 glass microvials (see Supplementary Information for details). Pybox = 2,6-bis[(*4 S*)-(–)-isopropyl-2-oxazolin-2-yl]pyridine, box = 2,2-Bis((*4 S*)-(–)-4-isopropyloxazoline)propane, terpy = 2,2′:6′,2″-terpyridine, dmphen = 4,7-dimethoxy-1,10-phenanthroline, CuTC = copper(I) thiophene-2-carboxylate, CuPC = copper(II) phthalocyanine. **D** Heatmap showing the reproducibility of quadruplicate data for a library of 96 acids coupled to **7**–**10** in plastic 1536-well plates. The overall average standard deviation across the quadruplicate data was 5.5%. The wells with the gray color indicated that the dosage of the internal standard was missing. **E** UV-Vis absorbance data show an interaction between $Cu(MeCN)_4BF_4$, **4** and 2,4,6-collidine. **F** Reaction optimization. a ^{1}H-NMR yield with 1,3,5-trimethoxybenzene as internal standard, b Isolated yield. **G** Diazonium salts used in the 1536 screen shown in **D**. **H** Select reactions from those shown in **D** were repeated on a 0.300 mmol scale using $Cu(MeCN)_4BF_4$ (100 mol%) to produce **11**, **12**, and **13** in isolated yield shown. 158

60-Fig. 4 | Substrate scope. Reactions were run with carboxylic acids (0.300 mmol), diazonium salts (1.5 equiv.), $Cu(MeCN)_4BF_4$ (30 mol%), and 2,4,6-collidine (1.5 equiv.) in anhydrous acetonitrile (0.1 M) at room temperature, generally over 16 h. Reported yields are isolated yields of purified products. 160

61-Fig. 5 | The amine–acid esterification applied directly to over-the-counter pills. Ester products **42**–**45** are derived from both reagents and OTC pills (See Supplementary Information for details).**The Formal Cross-Coupling of Amines and Carboxylic Acids to Form *sp*^{3}–*sp*^{2} Carbon–Carbon Bonds** 163

62-Fig. 1 | (A) Cross-coupling of amines and carboxylic acids to form C(sp^3)–C(sp^2) bonds as a complement to amide coupling. (B) Substructure search of 320 possible products from amine–acid coupling in the DrugBank database. (C) Analysis of building block availability from MilliporeSigma. Unions represent exact substructures differing only by the indicated functionality. 165

63-Fig. 2 | (A) HTE optimization on 10 μmol scale in 24- and 96-reaction arrays. Assay yields were determined by UPLC-MS. Selected conditions were repeated on the 0.25 mmol scale and analyzed by ^{1}H NMR with an internal standard. Isolated yield in parentheses. (B) Key compounds. 166

64-Fig. 3 | Substrate scope of C(sp^3)–C(sp^2) coupling of *N*-acyl-glutarimides with primary alkylpyridinium salts on 0.10–0.25 mmol scale. Isolated yields are shown. 168

65-Fig. 4 | (A) Effect of $RuCl_3$ on selectivity. Isolated yield on 0.25 mmol shown. Pie charts determined from ^{1}H NMR analysis of crude reaction with reference to an internal standard. (B) Effect of $RuCl_3$ in the absence of $NiBr_2{\cdot}DME$. TPP = triphenylpyridinium. (C) Coupling of **1** and **2** in the presence of ketone **37** and $RuCl_3$. ^{a}Not detected by UPLC-MS. ^{b}Yield from ^{1}H NMR. 168

66-Fig. 5 | Substrate scope of C(sp^3)–C(sp^2) coupling of *N*-acyl-glutarimides with secondary and benzyl alkylpyridinium salts on 0.25 mmol scale. Isolated yields are shown. 169

67-Fig. 6 | (A) Replacement of **1** with *N*-acyl phthalimide **45**, ^{1}H NMR yields shown. (B) DFT comparison of oxidative addition of **1** and **45**. (C) Proposed catalytic cycle. 170

68-Fig. 1 | Approaches to nanoscale synthesis and affinity ranking (NanoSAR); (a) Previously, crude nanoscale reactions were assayed by ASMS with affinity ranking achieved by decreasing the concentration of target protein to induce competition among compounds; (b) introduction of a Caliper bioassay into the workflow allows larger libraries of thousands of compounds to be evaluated and reduces the need for mass encoding. 172

69-Fig. 2 | Chemical assay results of ultrahigh throughput experiment performed for the project. Heatmaps represent the UV, MS, and ELSD integrations of the product over internal standard respectively. 172

70-Fig. 3 | Library design and synthesis (left) tSNE comparing targeted compounds to virtual compounds generated from enumeration of the enamine acid catalog. The targeted compounds cover a large portion of chemical space. (right) Relative conversion to product for the ultraHTE amide coupling of pharmacophore amines with 128 different acids as determined by UPLC-MS. Three detection methods, UV-absorbance, MSTIC(+) and ELSD were utilized. Each point represents one

reaction, colored by percent inhibition observed when subjected to Caliper biochemical assay. ... 173
71-Fig. 4 | Affinity Selection Mass Spectroscopy chord diagrams for each pool of reactions. All reactions were initially split into four pools, based on reagents, resulting in 768 reactions per pool. Chords connect reactions to compound masses found in that reaction. Each reaction may be connected to multiple masses, and heatmaps representing chemical and biological assay results are shown on the periphery of the reaction arc. Chords are colored by the lowest ALIS protein concentration that a compound in that reaction bound to. In the second iteration, 16 pools were made where reactions were split by conversion. ... 174
72-Fig. 1 | Overview of the ChatGPT to phactor workflow. Verbal input is given by a human to have ChatGPT generate a reaction array design for a particular coupling and substrate pair. The output can be fed directly into phactor, creating an assay recipe to be executed robotically or manually. ... 192
73-Fig. 2 | The amide coupling between **1** and **2** ChatGPT was asked to optimize. The reagent classes are specified in the prompt, but the specific species are generated by ChatGPT. ... 193
74-Fig. 3 | Executed reaction array and UPLC assay results of the screen designed by ChatGPT to perform the amide coupling between **1** and **2**. The top performing reaction condition was repeated on 0.2 mmol scale to yield **3** in 94% isolated yield. ... 193
75-Fig. 4 | The amide coupling between **4** and **1** ChatGPT was asked to optimize. The resultant design is different than with the original simpler substrates. ... 193
76-Fig. 5 | Results of the amide coupling to produce **5** designed by ChatGPT and phactor when executed experimentally. The top performing reaction condition was repeated on 0.4 mmol scale to yield **5** in 62% isolated yield. ... 194
77-Fig. 6 | The Suzuki coupling between **6** and **7** ChatGPT was asked to optimize. ... 194
78-Fig. 7 | Results of the Suzuki coupling reaction array designed by ChatGPT and executed with phactor. The top performing reaction condition was repeated on 0.2 mmol scale to yield **8** in 66% isolated yield. ... 195
79-Fig. 8 | Results of the repeated Suzuki coupling reaction array designed by ChatGPT when 20% water is added. ... 195
80-Fig. 9 | The Buchwald-Hartwig coupling ChatGPT is asked to optimize. ... 195
81-Fig. 10 | (A) 0.2 mmol scale-up result of the best hit from the reaction array proposed by ChatGPT-phactor. The combination of palladium catalyst complex Pd(dppf)Cl_2 and RuPhos gave the best result. (B) The reaction array design and results. (C) Crude UPLC trace for selected well B5. ... 196
82-Fig. 11 | Integrated GPT interface in phactor. Based on the substrates selected by the chemist and a description of the desired transformation, chemicals suggested by the AI can be automatically added into the reaction array design by phactor. ... 197
83-Fig. 1 | A) Anatomy of a reaction. Electrophiles and nucleophiles are arbitrarily defined as the reacting partners that form the product. B) The reaction fingerprint can be calculated by taking the product between a matrix of component fingerprints and a vector of weight arrays. Any fingerprint or feature vector can be utilized as long as the vector lengths for each reaction components are equal. Reaction component weights (w_n) are initialized at 1 and can be set to any value by the user.

A component * weight product is added to the sum for each reaction component *n*. 208

84-Fig. 2 | 2,786 Suzuki reactions performed in a high-throughput format plotted using a t-SNE trained on *reaction fingerprints*, each of which formed through the sum or concatenation of a reactant fingerprint for all components involved a particular reaction. (A) The template of the reaction performed across all experiments. (B) The 2,048-bit radius four Morgan Fingerprint of each reaction's electrophile, nucleophile, base, catalyst, and solvent were summed, plotted, and colored by output group. (C) The same manifold colored by exact product/internal standard values and reagent components (electrophile, nucleophile, catalyst, base, solvent.) (D) A specific reaction cluster containing all reactions between substrates **1** and **2**. (E) t-SNE of the reactions in the selected cluster when the catalyst fingerprint is multiplied by three (top) and when the base fingerprint is multiplied by three (bottom). The three manifolds for each result are colored by product/internal standard values, catalyst, and base. 210

85-Fig. 3 | 1,296 reactions performed in an HTE format towards the development of a novel sp^3–sp^3 deaminative-decarboxylative carbon–carbon cross coupling (ref. 13). The manifold splits into three regimes, each representing a different stage of the discovery campaign. (A) The manifold points are colored by the acid used in the experiment. (B) The manifold points are colored by the recorded PROD/IS value of the reaction. (C) All acids used in the experiment, with average PROD/IS values and the number of experiments they were used in. 212

86-Fig. 4 | Reactivity cliffs are identified when focusing on specific substrate pairs. (A) Compounds **27** and **28** cluster together with high and poor performing reactions. (B) Repeating the analysis on the cluster that form **29** from **27** and **28** reveals reagents that cause this reactivity to flip on. Reactivity cliffs are readily identified between solvent regimes. The 1:3 water:NMP regime sits between two reactivity cliffs, separating it from the two failing solvent systems 1:9 ^{t}AmOH:NMP and pure NMP. 213

87-Fig. 5 | Reaction array manifolds can be procedurally generated from the output files produced by phactor. (A)-(F) reaction array results as reported in ref. 8 and their corresponding manifolds colored by output value. Perplexity and weights were modified as described in the Supporting Information to optimize the latent space for visualization. 214

88-Fig. 6 | (a) Standard conditions for catalyst-free light-enabled decarboxylative–deaminative sp^2–sp^2 C–C coupling. (b) Sampling of substrates explored. (c) Screen design and execution. (d) Trellised reaction array fingerprint of experimental campaign. Each box contains all reactions performed for the given light and solvent regime. Reactions that perform well in one regime but not others are quickly identified. 217

89-Fig. 1 | A) A standardized template for chemical reactions. B) A reagent network containing 817 high throughput reactions performed in an optimization campaign of an alkyl-aryl deaminative decarboxylative carbon-carbon coupling method. Each vertex corresponds to a different class of reagents used in the reaction. These vertices are reduced manifolds of reagents corresponding to that class and can be formed from different featurizations depending on the class. Points are colored by mean yields of the reagent, and links between two points are colored by mean yield of reactions containing both reagents, indicating the effectiveness of their co-use. Black points indicate reagents not utilized in the campaign and serve as potential design space. 226

90-Fig. 2 | A) Graphical and reagent network analysis of an alkyl-aryl deaminative decarboxylative carbon-carbon coupling methodology optimization campaign. B) The analysis repeated on an alkyl-alkyl deaminative decarboxylative carbon-carbon coupling methodology optimization campaign. 228

91-Fig. 3. A) Best reaction from initial reaction array and reagent network analysis of ester to ether methodology campaign. The initial reaction array design was derived from a literature report. B) The second reaction array in the campaign. Boxplot trends reveal gallium iodide as the most robust Lewis acid for the transformation. C) Expected improvement manifolds calculated from the featurized silane and borane chemical spaces. New silanes and boranes were automatically selected to be tested in the follow up experiment. 228

92-Fig. 1 | (A) Workflow for SAMPLES. Molecules are first assigned a musical key based on aggregate chemical properties, then converted into a sequence of notes based on SELFIES encoding. MusicVAE is trained on a collection of sonified molecules to formulate the chemical/musical latent space. The latent embedding of molecular music can then be sampled, for instance through the interpolation between two embedded molecules, decoded by the MusicVAE decoder, then converted back into a molecular structure by SAMPLES. (B) Detailed workflow schematic of SAMPLES. Molecules are binned into base keys based on their physicochemical properties. More druglike molecules correspond to melodies that are more popular as reflected in the distribution of songs listed in Spotify. Specific SELFIES tokens are mapped to MIDI shifts, that result in a final MIDI value when summed with the base key of the molecule, which is hashed from aggregate physicochemical properties. The MIDI shifts correspond to SELFIES tokens as ranked by their popularity in DrugBank. As such, lower MIDI shifts away from the encoded key of the molecule indicate the more popular the SELFIES token. Melodies can then be decoded back into molecules. 245

93-Fig. 2 | SAMPLES translates molecules into music. (A) The generation of increasingly complex molecules from **1** through **4** corresponds to increasing musical complexity. Each line shows the molecular structure, the corresponding musical score, and a waveform of the MIDI output. Audio recordings are available in the ESI and can be quickly retrieved by scanning the QR code with a mobile device. (B) In the reverse direction, the song “Twinkle, Twinkle Little Star” produces molecule **5**. 246

94-Fig. 3 | SAMPLES is amenable to encoding complex molecules. (A) The pair of similar molecules **6** and **7** have SAMPLES compositions that are distinct from another similar pair of molecules **8** and **9**. (B) QR codes linking to SAMPLES encodings of select large complex molecules. 247

95-Fig. 4 | tSNE embedding of 11 159 drugs from DrugBank (2048 bit Morgan Fingerprints of radius 2), coloured by their SAMPLES musical key. Similar molecules **10** and **11** have SAMPLES outputs that are distinct from other similar molecules **12** and **13**. The music encoded by the score adjacent to the molecules can be listened to by following the QR code. It can be heard that these similar molecules have similar musical encodings when using our algorithm. 249

96-**Fig. 5 |** Molecular editing in SAMPLES generates distinct but related molecules. The manual editing of single notes in the SAMPLES of **11** leads to **14**, **15**, or **16**. 250

97-**Fig. 6 |** CROSSFADE merges SAMPLES with the melody mixing function of MusicVAE to create interpolated molecules based on two input molecules. 252

98-**Fig. 7 |** A human created music composition leading to **29**. 252

99-**Fig. 8 |** (A) Survey results from 75 participants. Each participant was given the SAMPLES encoded melody of four survey molecules. For each survey molecule, without knowledge of the name or structure of any molecule, each participant was asked to choose the most similar melody from a selection of four other SAMPLES encoded drugs. Survey responses are cross examined against the Tanimoto similarity between each test molecule and survey molecule for each question. (B) Superimposed waveforms, for each question, of the survey molecule, the molecule most structurally similar to survey molecule, and the most popular answer if the majority response in incorrect. (C) Structures of survey and test molecules for each question. One structure for each set of test molecules was chosen to have high similarity to the respective question's survey molecule to serve as the ‘correct answer’. 253

100-Fig. 1 | The interface when caffeine is encoded. SMILES can be input in the bottom right to bootstrap a song. The number of notes corresponds to the size of the molecule, and additional notes and synths can be added at will. Notes can be changed using the interactive piano, and oscillators and filters can be modified on the left hand side. 256

101-Fig. 2 | The encoding of bromhexine produces a different musical output. 257

102-Fig. 3 | An abstractable data structure dictates the settings and notes for each synth. This data structure can be recreated in external scripts and dropped into the program to create the music. 258

103-Fig. 1 | (a) A self-contained code template generates several plots and is provided to students. Several lines are deactivated, as indicated by green text. Lines of code can be reactivated by removing the leading “#”, resulting the code to produce a different visual output when executed. (b) The plot generated by the code in (a) when run as is. The three plotted points are defined in the code template, connecting the code to its visual output. (c) The plot generated when line 8 of (a) is activated, resulting in larger points. (d) The plot generated when line 9 is also activated, resulting in different point colors. (e) The plot generated when line 11 is activated, creating the same scatter plot with a blue background. (f) A potential plot generated after students are instructed to include an additional parameter to the scatter function. In this case, the shape of the points are changed. 277

104-Fig. 2 | Students are provided with the code to import any tabular JSON file. The utility of the package Pandas is used in reading tables programmatically. The “alldrugsprops.json” file is provided to students, and provided with a template to import the data into Python, as seen in the input block. Students then inspect the contents of the file with Python, revealing a datafile of over 9,000 drugs and their properties downloaded from DrugBank. 278

105-Fig. 3 | Four chemoinformatic experiments run by students during the first session. In each graph, over 9,000 drugs are graphed onto scatter plots as point, where their x and y locations correspond to calculated properties of the drug. Students learn how variables can be used to rapidly change the axes and colors of the plot, exposing them to a strategy to rapidly investigate chemical space and generate reports. 278

106-Fig. 4 | (a) The template plotting code is incrementally improved until it can be used effectively to make manuscript-ready graphics. This code creates a scatter plot of the Moonshot compounds provided in the CSV, where the x axis represents a compound's number of aromatic rings, and the y axis represents its LogP. (b) By coloring the points by the compound's recorded solubility, a trend is revealed where compounds with fewer aromatic rings and lower LogP are more soluble. Validating GSK's Solubility Forecast Index is a simple experiment to allow students to build confidence in their ability to manipulate and analyze datasets. 279

107-Fig. 5 | By the end of the modules, students have implemented a data filter and a principal component analysis on a dataset of SARS-COV-2 Main Protease inhibitors. (a) the PCA before the data filter is applied. Points represent DrugBank compounds and are colored by their molecular weight. By changing the color of the points, the correlation of the reduced axes can be mapped to specific features of the dataset. In this instance, molecular weight tracks with the first principal component, increasing as the x value increases. (b) the PCA after the Lipinski filter is applied. Outliers are removed by the filter, and the distribution of molecular weights in the first principal component becomes wider. 279

108-Fig. 6 | Student feedback to the lesson plan collected after the lab. Questions were intended to gauge the student's perspective on their learning experience. In the first year, students felt strongly that their understanding of Python, chemical space, and data science were improved. In the second year, students felt strongly that their understanding of data science improved but were neutral on other learning objectives. The difference between the two years may be accounted for by the level of involvement the graduate student teaching assistance had in developing the course. In the first year, the class was taught by the writer of the lesson plan; the second year was taught by assistants who were given the lesson plan to teach. 280

109-Fig. 1 | Overview of LRphase. A. Long sequencing reads are first prepared from genomic DNA fragments isolated from cells with available phased variant data for all parental haplotypes. Reads are mapped to the reference genome, either within LRphase with minimap2, or externally using any desired mapping/filtering workflow, with mapped reads supplied in BAM format. B. Haplotagging begins by intersecting individual mapped reads with known phased, heterozygous single-nucleotide variants (SNVs) supplied as a VCF file. The number of matches and mismatches are counted for both maternal and paternal phases and resulting counts are subsequently used to choose the best match to phase using one of the two scoring modes. C. Reads may be haplotagged using either the scoring model (Mode 1, orange box), or by simply assigning to the phase with the greatest number of matching alleles across all heterozygous sites (Mode 2). In Mode 1, match and mismatch counts are used to calculate log-likelihoods for matching to maternal and paternal phases. Likelihoods are computed as multinomial probabilities ($P_{maternal}$

and $P_{paternal}$ respectively), representing the aggregate probability of obtaining the observed number of matches and mismatches to either phase given the observed sequencing error rate, ϵ. $P_{maternal}$ and $P_{paternal}$ are adjusted by a Bayesian prior ($ß_m$ and $ß_p$) and log-likelihood ratios (LLRs) $L_{mat/pat}$ and $L_{mat/pat}$, are calculated to determine which phase has the most evidence for matching a given read. Reads are assigned to the phase corresponding to the greater of $L_{mat/pat}$ and $L_{mat/pat}$, are labeled “unphased” if there is a tie, or “nonphasable” if there are no intersecting heterozygous variants. D. After all reads are haplotagged, the false-discovery rate (FDR) is controlled by calculating the expected number of incorrect haplotags, $N\varepsilon$, as the mean of the negative-binomial distribution with N = the number of phaseable reads and $P_{success} = 1-(\epsilon/3)$. Phaseable reads are sorted by LLR and the $N\varepsilon$ * (1-FDR) lowest-scoring reads are relabeled as unphased. The remaining haplotagged reads are expected to include errors at a rate corresponding to the specified FDR. Finally, haplotagged results are written to output file(s) in BAM format, with the assigned haplotype stored in the “HP” tag, while custom tags are used to store all variables used in LLR calculations. 282

110-Fig. 2 | Haplotagging performance of LRphase and WhatsHap. Precision-Recall (PR) curves are used to assess the ability of LRphase (A) and WhatsHap (B) to identify the correct haplotag. Curves describe the fraction of correct and incorrect haplotag assignments across varying score thresholds. Individual points along each curve are color coded to illustrate score ranges at each combination of precision and recall. A. PR curve for LRphase haplotag assignments. B. PR curve for WhatsHap haplotag assignments. 284

111-Fig. 1| A representative cartoon of how PAMPA works. 286

112-Fig. 2 | Structures of phenazopyridine and methylene blue. 286

113-Fig. 3 | A PAMPA plate donor plate (left) and the acceptor plate (right) after a one-hour incubation. Note only phenazopyridine is permeable. 286

114-Fig. 4 | The setup of the Opentrons OT-2 layout. 287

115-Fig. 5 | Top: Opentrons transferring from PAMPA plate to plate reader plate. Right: PAMPA (white, source) and plate reader (black, destination) plates for Opentrons dosing. Blue represents methylene blue and yellow represents phenazopyridine. The height of the color represents relative concentration. The source plate shows what was placed into the donor plate. Row A in acceptor plate is methylene blue calibration curve, Row B is phenazopyridine calibration curve, and row C is what permeated through the donor plate. 287

116-Fig. 6 | Distribution of students correctly answering each part of the lab report. N=75 288

117-Fig. 1 | A visualization of phactor’s codebase. 291

118-Fig. 2 | The settings page of phactor. 292

119-Fig. 3 | The factors page of phactor. 292

120-Fig. 4 | The product input pop-up page of phactor on the factors stage. 293

121-Fig. 5 | The chemicals stage of phactor with compounds input. 293

122-Fig. 6 | The connected inventory input pop-up page of phactor on the chemicals stage. 294

123-Fig. 7 | The connected ask GPT input pop-up page of phactor on the chemicals stage. 294

124-Fig. 8 | The grid stage of phactor displaying a 24-well reaction array. 295
125-Fig. 9 | The grid stage of phactor displaying a 1536-well reaction array with 1,280 reactions..295
126-Fig. 10 | The connected internal standard calculator pop-up page of phactor on the grid stage. 296
127-Fig. 11 | The analysis stage of phactor displaying the results of a 24-well reaction array...296
128-Fig. 12 | The UPLC-MS analytics popup on the analysis stage of phactor displaying UPLC-MS trace of a completed reaction. 297
129-Fig. 13 | The report stage of phactor displaying the report of a completed 24-well reaction array. 297
130-Fig. 1 | Visualization of the phactor inventory codebase. 298
131-Fig. 2 | Request table of the phactor inventory..299
132-Fig. 3 | Confirm arrival pop-up of the request page. 299
133-Fig. 4 | Compounds table of the phactor inventory. 299
134-Fig. 5 | BarCodak instance pop-up table of the phactor inventory.......................300
135-Fig. 6 | PCA embedding page of the phactor inventory. 300
136-Fig. 7 | Consumables table of the phactor inventory. 300
137-Fig. 8 | Edit consumable form of the phactor inventory. 301
138-Fig. 9 | Consumable shopping cart table of the phactor inventory. 301
139-Fig. 10 | Request compound form of the phactor inventory.................................302
140-Fig. 11 | Register compound form of the phactor inventory.................................302
141-Fig. 12 | Search database form of the phactor inventory.....................................302
142-Fig. 13 | Admin panel of the phactor inventory. 303
143-Fig. 1 | Samples webapp playing taxol. 304
144-Fig. 2 | Prototype of the SAMPLES additive synthesizer webapp. 304
145-Fig. 3 | 8-bit sequencing prototype implementation of samples. 305
146-Fig. 4 | 8-bit sequencing prototype implementation of samples with filters and low frequency oscillators. 306
147-Fig 5. | iPhone implementation of SAMPLES algorithm 308
148-Fig. 1 | Retrosynthetic analysis of remdesivir on the visualizer webapp. 309
149-Fig. 2 | Retrosynthetic analysis of bromhexine on the visualizer webapp. 309
150-Fig. 3 | Retrosynthetic analysis of umifenovir on the visualizer webapp. 310
151-Fig. 4 | Retrosynthetic analysis of favipiravir on the visualizer webapp...............310
152-Fig. 5 | Retrosynthetic analysis of ritonavir on the visualizer webapp. 311
153-Fig. 6 | Retrosynthetic analysis of cobicistat on the visualizer webapp. 311
154-Fig. 7 | Retrosynthetic analysis of ribavirin on the visualizer webapp. 312
155-Fig. 8 | Retrosynthetic analysis of camostat on the visualizer webapp. 312
156-Fig. 9 | Retrosynthetic analysis of darunavir on the visualizer webapp. 313
157-Fig. 10 | Retrosynthetic analysis of nelfinavir on the visualizer webapp..............313
158-Fig. 11 | Retrosynthetic analysis of galidesivir on the visualizer webapp. 314
159-Fig. 12 | Retrosynthetic analysis of gelsemine on the visualizer webapp.............314
160-Fig. 13 | Retrosynthetic analysis of baricitinib on the visualizer webapp. 315
161-Fig. A1 | Proton NMR of 2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxylic acid. 319
162-Fig. A2 | Carbon NMR of 2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxylic acid. ...320

163-Fig. A3 | Proton NMR of N-(benzo[d]thiazol-2-yl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide. 321
164-Fig. A4 | Carbon NMR of N-(benzo[d]thiazol-2-yl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide. 322
165-Fig. A5 | Proton NMR of N-(4-(tert-butyl)phenyl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide. 323
166-Fig. A6 | Proton NMR of N-(4-(tert-butyl)phenyl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide. 324
167-Fig. A7 | Proton NMR of 2-oxo-N-(p-tolyl)-1,2,3,4-tetrahydroquinoline-4-carboxamide. 325
168-Fig. A8 | Carbon NMR of 2-oxo-N-(p-tolyl)-1,2,3,4-tetrahydroquinoline-4-carboxamide. 326
169-Fig. A9 | Proton NMR of 1-(*p*-tolyl)isoquinoline. 327
170-Fig. A10 | Proton NMR of 1-(4-ethylphenyl)isoquinoline. 328
171-Fig. A11 | Proton NMR of 1-([1,1'-biphenyl]-4-yl)isoquinoline. 329
172-Fig. A12 | Proton NMR of 2-methyl-N-(p-tolyl)benzamide. ^{1}H NMR (500 MHz, CDCl$_3$) δ 7.63 (s, 1H), 7.46 (dd, *J* = 24.8, 7.7 Hz, 3H), 7.34 (t, *J* = 7.5 Hz, 1H), 7.26 – 7.20 (m, 2H), 7.16 (d, *J* = 7.9 Hz, 2H), 2.47 (s, 3H), 2.34 (s, 3H). 330
173-Fig. A13 | Carbon NMR of 2-methyl-N-(p-tolyl)benzamide. ^{13}C NMR (126 MHz, cdcl$_3$) δ 168.11, 136.57, 136.35, 135.50, 134.16, 131.17, 130.14, 129.54, 126.67, 125.83, 120.04, 26.38, 20.92, 19.80. 331
174-Fig. A14 | Proton NMR of (R)-2-methyl-N-(4-oxo-4-(3-(trifluoromethyl)-5,6-dihydro-[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl)-1-(2,4,5-trifluorophenyl)butan-2-yl)benzamide. ^{1}H NMR (500 MHz, dmso) δ 8.14 (t, *J* = 8.8 Hz, 1H), 7.53 – 7.38 (m, 2H), 7.27 (td, *J* = 7.6, 1.4 Hz, 1H), 7.16 (t, *J* = 6.8 Hz, 2H), 7.06 (dd, *J* = 7.8, 1.5 Hz, 1H), 5.05 (d, *J* = 3.9 Hz, 1H), 4.88 (d, *J* = 3.7 Hz, 1H), 4.60 (d, *J* = 4.8 Hz, 1H), 4.29 (dq, *J* = 38.5, 6.5 Hz, 1H), 4.01 (dddd, *J* = 38.2, 20.6, 14.5, 7.8 Hz, 3H), 3.03 – 2.94 (m, 1H), 2.77 (ddd, *J* = 23.4, 14.8, 8.4 Hz, 3H), 2.07 (s, 1H), 2.01 (s, 1H), 1.90 (s, 1H). 332
175-Fig. A15 | Carbon NMR of (R)-2-methyl-N-(4-oxo-4-(3-(trifluoromethyl)-5,6-dihydro-[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl)-1-(2,4,5-trifluorophenyl)butan-2-yl)benzamide. ^{13}C NMR (126 MHz, dmso) δ 169.76, 168.91, 151.45, 137.58, 135.78, 135.08, 130.73, 129.63, 127.04, 125.79, 119.99, 117.98, 106.53, 104.88, 46.53, 43.98, 42.55, 41.71, 38.63, 38.23, 33.16, 21.48, 19.23, 19.10. 333
176-Fig. A16 | Proton NMR of 2,4'-dimethyl-1,1'-biphenyl. ^{1}H NMR (499 MHz, cdcl$_3$) δ 7.28 – 7.23 (m, 8H), 2.42 (s, 3H), 2.29 (s, 3H). 334
177-Fig. A17 | Carbon NMR of 2,4'-dimethyl-1,1'-biphenyl. ^{13}C NMR (126 MHz, cdcl$_3$) δ 139.02, 135.39, 130.26, 129.84, 129.06, 128.76, 127.05, 125.73, 21.17, 20.50. .. 335
178-Fig. A18 | Proton NMR of 2-methyl-N-(p-tolyl)aniline. ^{1}H NMR (499 MHz, cdcl$_3$) δ 7.22 (dd, *J* = 8.0, 1.3 Hz, 2H), 7.18 – 7.11 (m, 3H), 6.98 – 6.90 (m, 3H), 2.35 (s, 3H), 2.30 (s, 3H) 336
179-Fig. A19 | Carbon NMR of 2-methyl-N-(p-tolyl)aniline. ^{13}C NMR (126 MHz, cdcl$_3$) δ 142.02, 141.01, 130.89, 130.56, 129.87, 127.10, 126.79, 121.19, 118.72, 117.36, 29.76, 20.70, 17.90. 337

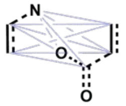
LIST OF APPENDICES

Appendix A: NMR Spectra of Isolated Compounds 319

ABSTRACT

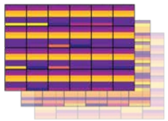
The navigation of chemical space through synthesis is a core principle behind drug discovery. Traditional medicinal chemistry programs rely on a few, well-developed reactivities to create pharmacophore derivatives and synthetic routes to target scaffolds. While reactions such as the amide coupling and Suzuki coupling are robust and tolerate many substrates, they introduce bias into chemical properties by limiting synthesis to a handful of bond formations, narrowing the scope and flexibility of chemical structures that can be formed from available building block compounds. Through computational enumeration, specific transformations can be targeted that direct access to specific and desirable chemical space. With high throughput experimentation and data science, these transformations can be invented and optimized experimentally to realize high value chemistries and accelerate drug discovery. This thesis describes an informatics-based systemic approach to synthetic chemistry, centered around reaction discovery to better access drug-like chemical space.

Chapter I. Introduction



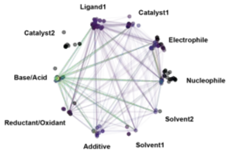
Reaction Targeting
Chapter II

enumeration
retrosynthesis
building block availablity
chemical space access



High Throughput Experimentation
Chapter III

phactor
automation
UPLC-MS analytics



Reagent Discovery
Chapter IV

reaction fingerprints
condition networks
recommender systems

Molecular Compression
Chapter V

sonification
additive embedding
feature optimization

1-Overview | The systematic informatics approach to synthetic chemistry consists of four main concepts. High value reactions are identified and can be invented with high throughput experimentation. Algorithms and computational representations of molecules enable reagent discovery to optimize reactivities and expand substrate compatibility.

This thesis is broken into four main chapters, each consisting of papers corresponding to a specific theme of the systematic informatics approach to synthetic chemistry. Another chapter is included consisting of additional projects I completed or contributed to, as well as a final chapter providing a brief overview of several software packages that were completed during my graduate studies. Each subsection of each chapter represents a self-contained project or paper, and compound numbers are reset at **1** at each subchapter. Each chapter begins with an introduction, and each subchapter begins with a prologue describing the work, which is then followed by text of the manuscript that corresponded with the project as well as its respective figures.

In the chapter following this introduction, *reaction targeting* is introduced. This chapter consists of five papers detailing enumerative combinatorics, the computational ideation of synthetic transformations, and corresponding analyses. It is detailed how enumeration can be used to explore reaction space and traverse physicochemical property profiles by changing the transformation between building blocks. It is also shown how reaction agnostic and/or building block aware retrosynthetic algorithms aware can be developed to invent creative and potentially superior routes to target molecules. The following publications are described in this chapter:

1. Mahjour, B.; Shen, Y.; Liu, W.; Cernak, T. A map of the amine-carboxylic acid coupling system. *Nature* **2020**, 580, 71–75.
2. Zhang, R.; Mahjour, B.; Cernak, T. Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing. *ChemRxiv*, **2023**.
3. Lin, Y.; Zhang, Z.; Mahjour, B.; Wang, D.; Zhang, R.; Shim, E.; McGrath, A.; Shen, Y.; Brugger, N.; Turnbull, R.; Jasty, S.; Trice, S.; Cernak, T. Reinforcing the Supply Chain of Umifenovir and Other COVID-19 Therapeutics with Retrosynthetic Software. *Nature Communications* **2021**, 12 (1), 1-8.
4. Mahjour, B.; Zhang, R.; Outlaw, A.; Zhang, Z.; Harmata, A.S.; Cernak, T. Analyzing Conceivable Reactions from Commercially Available Chemical Building Blocks. *This is under review at Organic Letters.*
5. Mahjour, B.; Flynn, K. M.; Stahl, S.; Cernak, T. One Step Retrosynthesis of Drugs by Cross-Coupling from Building Block Catalogs. *In collaboration with the Stahl Group at University of Wisconsin. Manuscript in Preparation.*

The next chapter includes work detailing the development and use of *high throughput experimentation*. Central to this chapter is the development of the high throughput experimentation manager phactor, a full-stack software that manages the data workflow for all high throughput experiments in the lab. Several papers are included where high throughput experimentation is used to invent and optimize high value reactions, miniaturize popular reactions, and complete ultrahigh throughput medicinal chemistry campaigns. The use of phactor in the lab led to the generation of a wealth of standardized and machine-readable data. The following publications are described in this chapter:

1. Mahjour, B.; Zhang, R.; Shen, Y.; McGrath, A; Zhao, R; Mohamed, O. G.; Lin, Y; Zhang, Z; Douthwaite, J. L.; Tripathi, A.; Cernak, T. Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery. *Nature Communications*, **2023**, 14, 3924.
2. Mahjour, B**.**; Shen, Y.; Cernak, T. Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. *Accounts Of Chemical Research* **2021**, 54 (10), 2337-2346.
3. Shen, Y.; Mahjour, B.; Cernak, T. Development of Deaminative Esterification Using High-Throughput Experimentation. *Communications Chemistry* **2022**, 5 (1), 1-8.
4. Douthwaite, J. L.; Zhao, R.; Shim, E.; Mahjour, B.; Zimmerman, P.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp^3–sp^2 Carbon–Carbon Bonds. *Journal of the American Chemical Society*, **2023**, 145, 20, 10930-10937.

5. Douthwaite, J.; Mahjour, B**.** et al. Expansion of NanoSAR for the deconvolution of large compound libraries. *In collaboration with Relay Therapeutics. Manuscript in Preparation.*

The following chapter consists of works using machine readable reaction data in exploring reaction space to analyze reaction campaigns. In *reagent discovery*, various visualizations and concepts are discussed to understand and predict reactivity trends. In early 2023, the advent of considerably effective artificial intelligence that understands human speech was unleashed. One work in this chapter details the use of such an intelligence to design high throughput experiments. The following publications are described in this chapter:

1. Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays using phactor and ChatGPT. *ChemRxiv*, **2023**. *OPRD,* accepted*.*
2. Mahjour, B.; Hoffstadt, J.; Schorin, D.; Cernak, T. Weighted Reaction Fingerprints for Visualizing Reactivity Cliffs and Generality. *Manuscript in Preparation.*
3. Mahjour, B.; McGrath, A.; Douthwaite, J.; Zhang, Z.; Cernak, T. Reaction Array Analysis and Design via Reagent Networks. *Manuscript in Preparation.*

The next chapter consists of creative projects where molecules are ‘embedded’ or ‘compressed’ into musical scores. These scores, encoding structural and physicochemical properties of the molecules, can be represented as waveforms – in essence one dimensional arrays that can be used for data science studies. In the initial study, the concept of *molecule compression* is introduced and explored. Encoded molecular music is used to train neural networks and the scores can be modified and decoded into new molecules. An ongoing project is further optimizing the encoding of the molecule into the highly dimensional domain of music. The following publication is described in this chapter:

1. Mahjour, B.; Bench, J; Zhang, R.; Frazier, J.; Cernak, T. Molecular Sonification for Molecule to Music Information Transfer. *Digital Discovery*, **2023**, 2, 520-530.
2. Mahjour, B.; Cernak, T.; Molecular Compression via Musical Additive Synthesis. *Manuscript In Preparation.*

In the subsequent chapter, a list of software and research workflow projects are included that I worked on or contributed to and are directly related to the core thesis of this manuscript but do not fall into other categories. This chapter contains efforts I made in developing a lesson plan for teaching undergraduates Python and cheminformatics in medicinal chemistry, as well as a software I wrote that phased long-read DNA samples by statistically matching them to the maternal or paternal allele based on the number of associated single nucleotide polymorphism. The following publications are described in this chapter:

1. Mahjour, B.; McGrath, A.; Outlaw, A.; Zhao, R.; Zhang, C.; Cernak, T. Interactive Python Notebook Modules for Chemoinformatics in Medicinal Chemistry. *This is under review at the Journal of Chemical Education.*

2. Holmes, M. J.; Mahjour, B.; Castro, C. P.; Farnum, G. A.; Diehl, A. G.; Boyle, A. P. LRphase: an efficient method for assigning haplotype identity to long reads. *BioRxiv,* **2023.**
3. McGrath, A.; Montgomery, D.; Mahjour, B; Outlaw, A.; Zhao, Ruheng.; Cernak, T. An Automated and Colorful PAMPA Assay To Investigate Drug Permeability. *Manuscript in Preparation.*

The final chapter before the conclusion lists several software packages that I have written that are used in the lab or are publicly available. Throughout my studies I wrote countless lines of code. The benefit of being a full-stack developer is that many of my projects are accessible through the internet and can be easily shared without any technical expertise required by the user. The details and infrastructure of these software packages are discussed within this chapter.

Chapter II. Reaction Targeting

Chemical space is the potentially infinite set of molecules that can exist. Every drug, natural product, substance, or material that physically exists is composed of molecules contained within chemical space. As such, physical inventions of drugs and materials leading to advancements in science, technology, and humanity requires the ideation, synthesis, and evaluation of chemical matter within this space. While constraints such as size and atomic composition provide tenable boundaries to the space, brute-force exploration is currently impossible computationally because of both time- and space-complexity. Furthermore, while computational enumeration through graph theory allows exhaustive ideation of chemical matter, only a small fraction of these molecules are accessible synthetically. Reaction targeting is a paradigm that intelligently uses computational enumeration to generate and evaluate hypothetical transformations while avoiding combinatorial explosion and maximizing synthetic feasibility.[1]

Molecular enumeration from the perspective of graph theory has been studied as early as 1875 to enumerate molecule classes, such as branched alkanes,[2, 3] alkyl alcohols,[4, 5] and cyclic carbon skeletons.[6] Work by Ugi, Dugundji, and others eventually led to matrix-based computer-assisted discovery of pericyclic reactions.[7-15] More systemic reaction cataloging methods based on redox state changes, size of reaction center,[16] and identity of reactant atoms,[17-21] as well as fingerprint methods[22] have also been reported since these initial studies. In contemporary studies, computing power has enabled exhaustive graphical enumeration of small molecules up to a handful of atoms and bond types and the development of generative algorithms can navigate embeddings of samples of chemical space.[23, 24] These efforts have cumulated in large datasets that attempt to catalog conceivable molecules.[25] Commercial catalogs and reaction databases are used in conjunction to design synthetic pathways to valuable chemical spaces.[26-28]

In contrast to molecular enumeration, reaction enumeration uses similar means to ideate potential transformations between molecular motifs. These motifs may be single atoms such as carbon or nitrogen in the most basic form, simple functional groups such as amines or acids, commercial building blocks, or complex drugs or natural products in the most complicated form. As opposed to connecting these substrates using traditional or known reactivities to build virtual libraries, a single pair of substrates can generate a library of isosteres when enumerating the transformation between them. For instance, the formation of a C–N bond after the elimination of the –OH group of a carboxylic acid and a hydrogen on an amine is recognized as the amide transformation. Depending on the size of the substrates being analyzed, an exhaustive enumeration is generally computationally feasible, with the most extreme case resulting in the decomposition and reformation of every bond between every atom.

Reaction targeting aims to use reaction enumeration to identify the most valuable transformations between classes of substrates. Transformation utility can be gauged by a hypothetical reaction’s ability to easily access novel or useful chemical space, optimize retrosynthetic routes, and exploit abundant feedstock molecules. As such, the paradigm of reaction targeting can be viewed from two perspectives.

In the first, transformations can be ideated to provide streamlined access to chemical space, accelerating the number of compounds that are able to be synthesized and subsequently evaluated. Drug compounds often require a careful balance of physicochemical properties. Modifying the transformation between substrates as opposed to entire substrates with the same transformation provides greater control over the property profile of the potential drug compound. For instance, a molecule containing an amide bond formed from an acid and amine substrate may be indicated to perform better with a C–C isostere. Indeed, this was the case for many HIV inhibitor design campaigns.[29] Creating an identical molecule with the C–C isostere can be synthetically cumbersome unless a direct method from the original substrates existed: the deaminative-decarboxylative C–C coupling. Isosteric modifications lessen the reliance on alternative building blocks that share structural cores and allow for fine-tuned modification of the molecule as opposed to a coarse modification through substrate replacement.

Alternatively, reaction targeting can be used to identify reactions that can achieve syntheses of specific molecules more optimally. There is some overlap with the chemical space perspective: a reaction with good substrate scope that creates enhanced access to a scaffold from common building blocks provides the ability to make many analogs. Whether it is a specific molecule or a scaffold being targeted, the disconnection of bonds can be enumerated. The resulting synthons can then be matched to feedstock molecules that are commercially available to create retrosynthetic routes.

In this chapter, initial reports of reaction targeting are shared, with a focus on exploration of chemical space and retrosynthesis. Many works are accompanied with synthetic realization of high value reactions. When analyzed in bulk format, generic transformations between generalized substrates are revealed, such as those that form C–C bonds between common substrate classes such as amines and carboxylic acids. Since the initial analyses of reaction targeting, many of these high value reactions have been realized synthetically through the use of high throughput experimentation and data science. Further analyses reveal valuable reactions that can form drug structures in few steps from specific or constrained starting materials. In conjunction with modern synthetic and data science techniques, reaction targeting will continue to prove to be a valuable philosophy in chemical synthesis and medicinal chemistry.

A map of the amine-carboxylic acid coupling system

In this work, enumerative combinatorics is introduced. All possible reactions between a simple amine and simple carboxylic acid are enumerated, resulting in 320 potential transformations. When computationally performing these transformations between amines and carboxylic acids, a range of physicochemical, structural, and spatial properties can be achieved. Several of these transformations are experimentally realized with simple and complex substrates. This work was published in *Nature* in 2020.[1]

Abstract

Chemical transformations determine the structure of a product, and therefore its properties, which in turn affect complex macroscopic functions such as the metabolic stability of pharmaceuticals or the volatility of perfumes. Therefore, reaction selection can influence the success or failure of a candidate molecule to meet a functional objective. The coupling of an amine with a carboxylic acid to form an amide bond is the most popular chemical reaction used for drug discovery.[30] However, there are many other ways to connect these two common functional groups together. Here we show computationally that amines and acids can couple via hundreds of hypothetical yet plausible transformations, and we demonstrate experimentally the application of a dozen such reactions. To investigate the contribution of chemical transformations to properties, we developed a string-based notation and used an enumerative combinatorics approach to produce a map of conceivable amine–acid coupling transformations, which can be charted using chemoinformatic techniques. We find that critical physicochemical parameters of the products, such as partition coefficient and polar surface area, vary considerably depending on the transformation chosen. Data mining the amine–acid coupling system produced here should enable reaction discovery, which we demonstrate by developing an esterification reaction found within the mapped space. Complex molecules with distinct property profiles can also be discovered within the amine–acid coupling system, as we show here via the late-stage diversification of drugs and natural products.

Main

The amide coupling is a robust and popular reaction used frequently in chemical synthesis. The transformation couples an amine (**1**) and a carboxylic acid (**2**) to form an amide (**3**) (1-Fig. 1a). Viewed in the context of physicochemical properties, the transformation unites a hydrophilic basic moiety (**1**) bearing two hydrogen bond donors, with a hydrophilic acidic moiety (**2**) bearing one hydrogen bond donor and two hydrogen bond acceptors, to generate a neutral product, **3**. The amide product is more lipophilic than the starting reagents, and has one hydrogen bond donor and one hydrogen bond acceptor. Chemoinformatic studies have linked physicochemical properties to functions as complex as toxicity[31] and even successful market launch[32], and so the ability to modulate the numbers of hydrogen bond donors, hydrogen bond acceptors, the partition coefficient logP, the molecular weight, and other properties of a molecule via chemical synthesis is of high importance. Control over physicochemical properties using chemical synthesis is typically achieved by varying starting materials iteratively or in a combinatorial manner[33], or by varying build–couple–pair reaction sequences to introduce skeletal diversity[34]. We hypothesized that physicochemical properties could be varied simply by switching the chemical transformation while holding the building blocks constant. In our view, transformations describes the mapping of atoms and bonds from starting materials to products[11], and can be described as reactions only when accompanied by experimental reaction conditions. We reasoned that a map of conceivable transformations would provide opportunities in reaction discovery, especially given contemporary developments in robotic[35-39] and algorithmic[40-44] techniques for predicting reaction conditions, in addition to presenting a strategy for chemical-space exploration.

The amide coupling is used in one quarter of the reactions reported in small-molecule pharmaceutical patents[30]. As a result, there is an abundance of available amine and acid building blocks. We questioned how many other transformations exist for the amine–acid coupling pair. Considering amine–acid couplings at the transformation level reveals opportunities for reaction discovery. For example, instead of coupling **1** and **2** to form **3** (2-Fig. 1a), a decarboxylation could occur to give **4**, or a deamination could occur to give **5**; likewise, a tandem decarboxylation–deamination could occur to forge a carbon–carbon bond as in **6** (2-Fig. 1b). Compounds **7**–**9** are also possible, and the set of compounds **3**–**9** collectively reveals that **1** and **2** could couple to form acidic products, basic products, neutral products and zwitterionic products. We used enumerative combinatorics (6-Extended Data Fig. 1) to create simplified molecular-input line-entry system (SMILES) strings for all products arising from the coupling of two generic functional groups, A and B.

A notation was developed (2-Fig. 1c) to describe how functional groups A or B can couple at the atoms of the functional group (A, B[C], or B[O], when B is CO_2H), or at the α or β carbon atoms. The notation also describes how functional groups A, B or both may appear in or be absent from the product. The transformation notation is written in the form ${}^{h}A^{XY}/{}^{h}B^{XY}$, where h is the hybridization (2 = sp^2, 3 = sp^3), X is the reacting atom and Y is any additional modification including loss of A or B (−A, −B), dehydration (−OH), or reduction (+H, +H2). All combinations of sp^2–sp^2, sp^2–sp^3, sp^3–sp^2 and sp^3–sp^3 transformations from **1**, **2** and their sp^2 variants ethenamine and acrylic acid (3-Fig. 2) were included, leading to 320 product substructures. Four of the transformations produced the amides resulting from coupling sp^2–sp^2, sp^2–sp^3, sp^3–sp^2 or sp^3–sp^3 amines and acids, respectively, but the vast majority of the enumerated transformations are currently unknown as reactions. By charting the amine–acid cross-coupling space, we aim to understand how chemical transformations affect physicochemical properties.

The enumerated SMILES strings were used as inputs to a series of chemoinformatic calculations. First, SMILES strings of the products were computationally ionized at pH 7.4 (Supplementary Information), and then used to calculate a range of physicochemical properties (7-Extended Data Fig. 2). The full set of 320 products spans a range of molecular weights from 54.1 to 120.2 g mol^{-1}, log$P$ = −2.29 to 2.19 units, hydrogen bond acceptors from 0 to 3, hydrogen bond donors from 0 to 2, polar surface area = 0 to 67.8 Å^2, fraction sp^3 = 0 to 1, number of rotatable bonds from 1 to 4, and a formal charge between –1 and 1. A composite function of drug-like properties, the quantitative estimate of drug-likeness[45], ranged from 0.27 to 0.54. These findings demonstrate that the choice of transformation can have a sizeable effect on properties. In the context of drug discovery, it may be necessary to decrease the number of hydrogen bond donors when optimizing a molecule for the ability to cross the blood–brain barrier[46], whereas it may be necessary to increase the number of hydrogen bond donors to improve aqueous solubility[47]. In this way, transformation mapping can enable studies in property optimization.

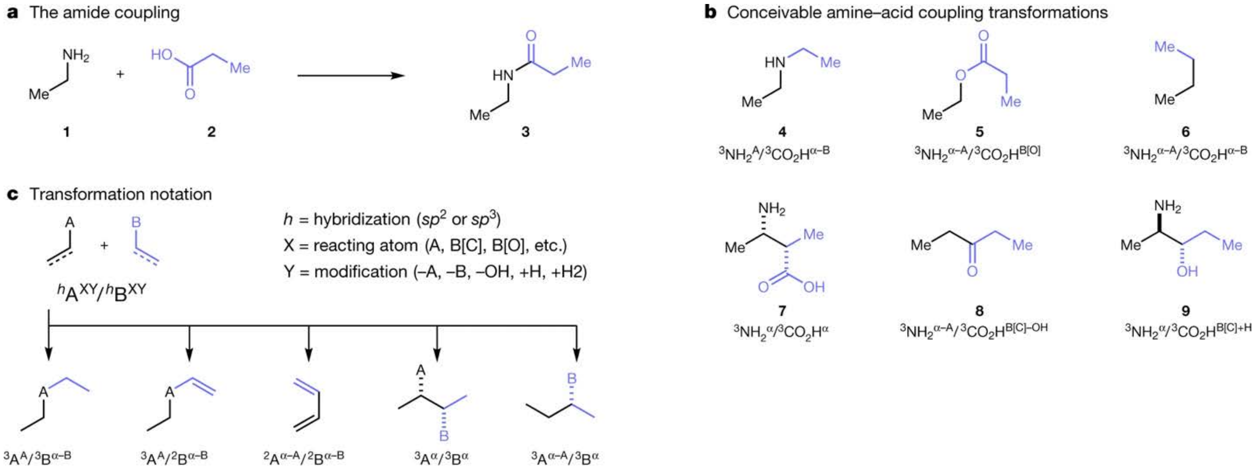
The 320 product molecules from the combinatorial enumeration were next used as substructures to search 9,279 pharmaceuticals and natural products from the DrugBank database[48]. As can be seen in 3-Fig. 2, there is a high degree of connectivity between the products of nearly every amine–acid coupling transformation with diverse pharmaceuticals and natural products. Each connecting line represents the successful identification of an enumerated product substructure within a drug, and the colour of the line depicts the frequency that a substructure occurs in that molecule. The density of connections in this system suggests that nearly every one of the 320 transformations depicted on the periphery of 3-Fig. 2 could find use in the synthesis of complex molecules. As expected, the simple alkyl chain **6**, formed by coupling **1** to **2** ($^{3}NH_{2}^{\alpha-A}/^{3}CO_{2}H^{\alpha-B}$), occurs frequently as a product substructure: 59,432 times among the DrugBank molecules (8-Extended Data Fig. 3). Likewise, decarboxylative transformations to produce an amine bound to an sp^{3}- or sp^{2}-carbon chain (such as $^{2}NH_{2}^{A}/^{3}CO_{2}H^{\alpha-B}$ to produce amines) appear in high frequency (9-Extended Data Fig. 4). Some transformations, such as $^{2}NH_{2}^{A}/^{3}CO_{2}H^{B[O]}$, do not appear as substructures in pharmaceuticals or natural products at all. This finding can be rationalized because, in this case, the transformation produces a hydroxyl amine ester, which is probably too reactive a functionality to persist in any of the complex molecules found in DrugBank. Analysing the system in the other direction, novel retrosynthetic strategies emerge by using amine–acid coupling transformations. For instance, (+)-noscapine connects to 112 transformations, (−)-quinine connects to 96 transformations and (−)-sitagliptin connects to 55 transformations (10-Extended Data Fig. 5), providing strategies for total synthesis. Our analysis until this point focused solely on achiral bond connectivities. In three-dimensional space, there are many more possible transformations, because some transformations produce *syn*-diastereomers whereas others produce *anti*-diastereomers (11-Extended Data Fig. 6). These chiral coupling transformations sample a substantial assortment of three-dimensional shapes (12-Extended Data Fig. 7).

To demonstrate our ability to control properties with atom-level precision, several transformations were selected and realized experimentally (4-Fig. 3a). We selected the amide coupling as well as four known reactions, which maximized the diversity of properties attainable from the coupling of *p*-toluidine (**10**) to *o*-toluic acid (**11**). Using the free amine and acid directly, we executed the amide coupling (${}^{2}NH_2{}^{A}/{}^{2}CO_2H^{B[C]–OH}$) under Schotten–Baumann conditions to give **12** in 91% yield. A $B(C_6F_5)_3$-catalysed reductive *N*-alkylation[49] was used to realize the ${}^{2}NH_2{}^{A}/{}^{2}CO_2H^{B[C]+H2}$ transformation, giving amine **13** in 60% yield. A cyclized benzoxazole (**14**) was also generated from **12** under oxidative conditions[50]. Through activation of the amine as the diazonium salt, an *ortho*-arylation ${}^{2}NH_2{}^{\alpha–A}/{}^{2}CO_2H^{\beta}$ and a corresponding decarboxylative variation ${}^{2}NH_2{}^{\alpha–A}/{}^{2}CO_2H^{\beta–B}$ were achieved via Gooßen's conditions[51] giving **15** or **16** in 64% and 68% isolated yield, respectively.

The mapping of amine–acid coupling space provides opportunities to devise reaction methods, and we discovered one reaction within this system. We reasoned that fruitful combinations of reagents, catalysts, ligands, activating groups and directing groups could be identified to realize hypothetical reactions. Towards this objective, we applied high-throughput experimentation techniques to interrogate the coupling of amine–acid derivatives using transition metal complexes, ligands and additives (13-Extended Data Fig. 8). We discovered a reaction based on the ${}^{2}NH_2^{\alpha-A}/{}^{2}CO_2H^{B[O]}$ transformation, which generated ester **17** from acid **11** and the diazonium salt of **10** under influence of copper(i) iodide, silver nitrate and pyridine. This reaction transforms a C–N bond into the C–O bond of the ester. The product, **17**, is a matched molecular pair to the corresponding amide **12**, but bears one less hydrogen bond donor. Thus, starting from **10** and **11** and simply by varying reaction conditions, we could produce the traditional amide (**12**), as well as closely related analogues that are basic (**13**), acidic (**15**), neutral and lipophilic (**14**, **16** and **17**), or neutral and hydrophilic (**12**). The products we obtained experimentally span a substantial portion of the full range of molecular weight, log*P* and polar surface area values achievable from all coupling transformations of **10** and **11** (4-Fig. 3b), showcasing the utility of our approach for fine-tuning molecular properties.

Many complex molecules contain an amine or an acid functional group, so we anticipated that the application of diverse amine–acid coupling transformations to late-stage diversification would enable access to congeners with diverse property profiles. We used chemoinformatics to evaluate late-stage diversification in the amine–acid coupling system on a series of complex substrates (14-Extended Data Fig. 9). Examination of the properties of the products reveals that the choice of transformation can determine whether an analogue will pass or fail the Lipinski rule of five[52], leading to a range in desirability score (quantitative estimate of drug-likeness)[45] of 0.31 to 0.70 for the couplings of the acid-containing antibiotic levofloxacin with 3,5-dichloroaniline, and 0.29 to 0.61 for the pairing of yohimbine and α-methylbenzylamine (14-Extended Data Fig. 9).

To experimentally demonstrate the value of the late-stage diversification concept (5-Fig. 4), enones derived from yohimbine (**18**), sulfadoxin (**22**) and lithocholic acid benzyl ether (**27**) were used as substrates. In the first instance, **18** was esterified and then converted to amide **19** in 78% yield by heating with **10** in the presence of lithium hexamethyldisilazide. Concurrently, **18** was esterified and then β-arylated to produce **20** upon palladium-catalysed Heck–Matsuda arylation using the diazonium salt of *p*-toluidine (**10**). We determined that **18** could be converted to **19**, then treated with magnesium in a one-pot operation to introduce an additional stereocentre, as in **21**. Likewise, **22** and cyclohexane carboxylic acid (**23**) coupled to form amide **24**, ester **25** by our copper(i)-promoted C–N to C–O $^{2}\mathrm{NH}_{2}^{\alpha\text{–}\mathrm{A}}/^{3}\mathrm{CO}_{2}\mathrm{H}^{\mathrm{B[O]}}$ reaction, or amine **26** under decarboxylative conditions[53]. Finally, **27** served as a framework to produce amide **29**, amine **30** or aminoester **31** via a one-pot $^{3}\mathrm{NH}_{2}^{\mathrm{A}}/^{3}\mathrm{CO}_{2}\mathrm{H}^{\alpha}$ α-amination sequence using piperidine (**28**). In this work we focused on amines and acids, but transformations of any pair of functional groups can be enumerated to serve as inspiration for the development of novel reaction methods and as a strategy for chemical-space exploration. All of the transformations mapped in the amine–acid coupling system could exist, but most are not yet linked to viable reaction conditions, making this transformation space a fertile proving ground for manual or automated reaction discovery.



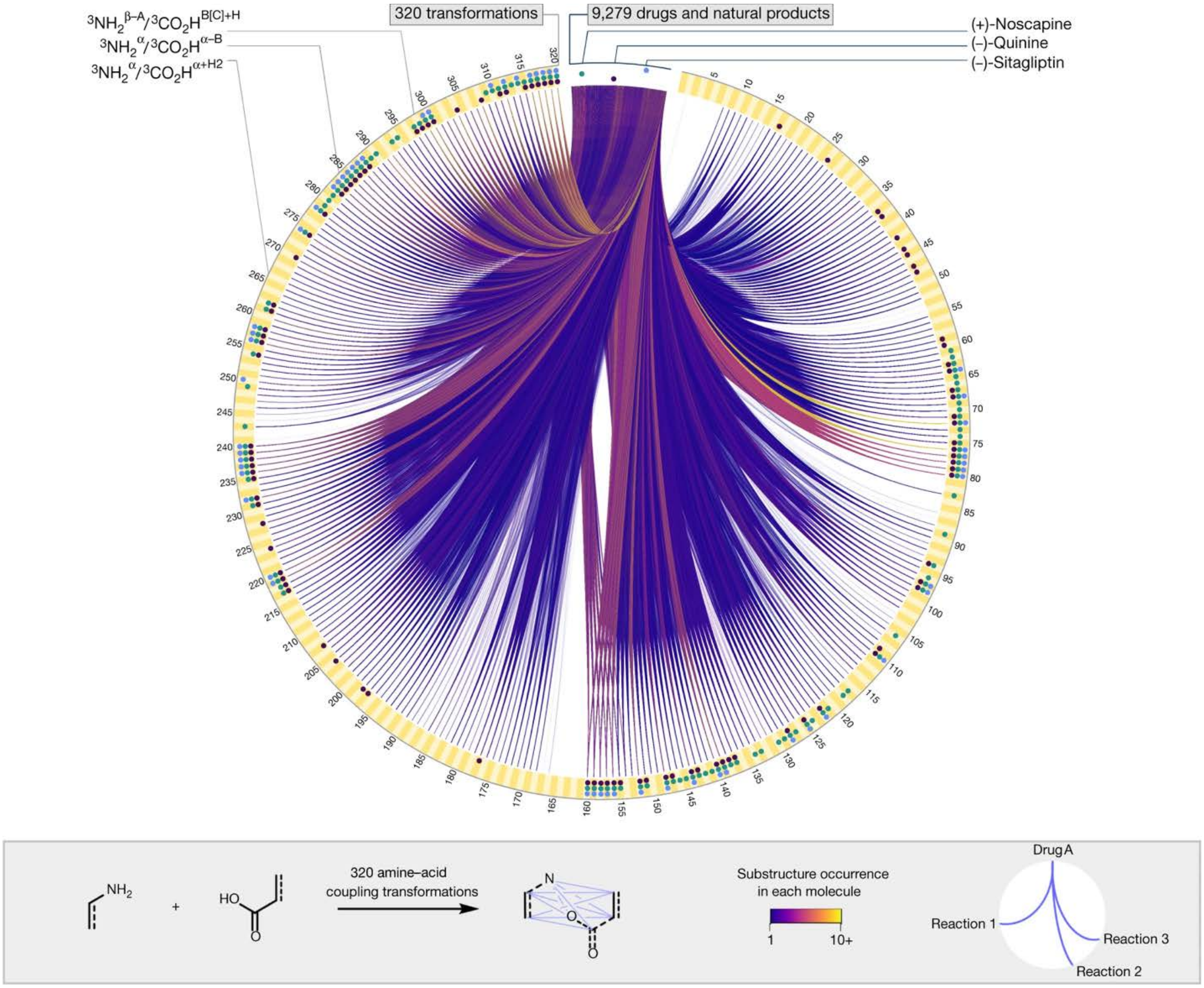
a The amide coupling

b Conceivable amine–acid coupling transformations

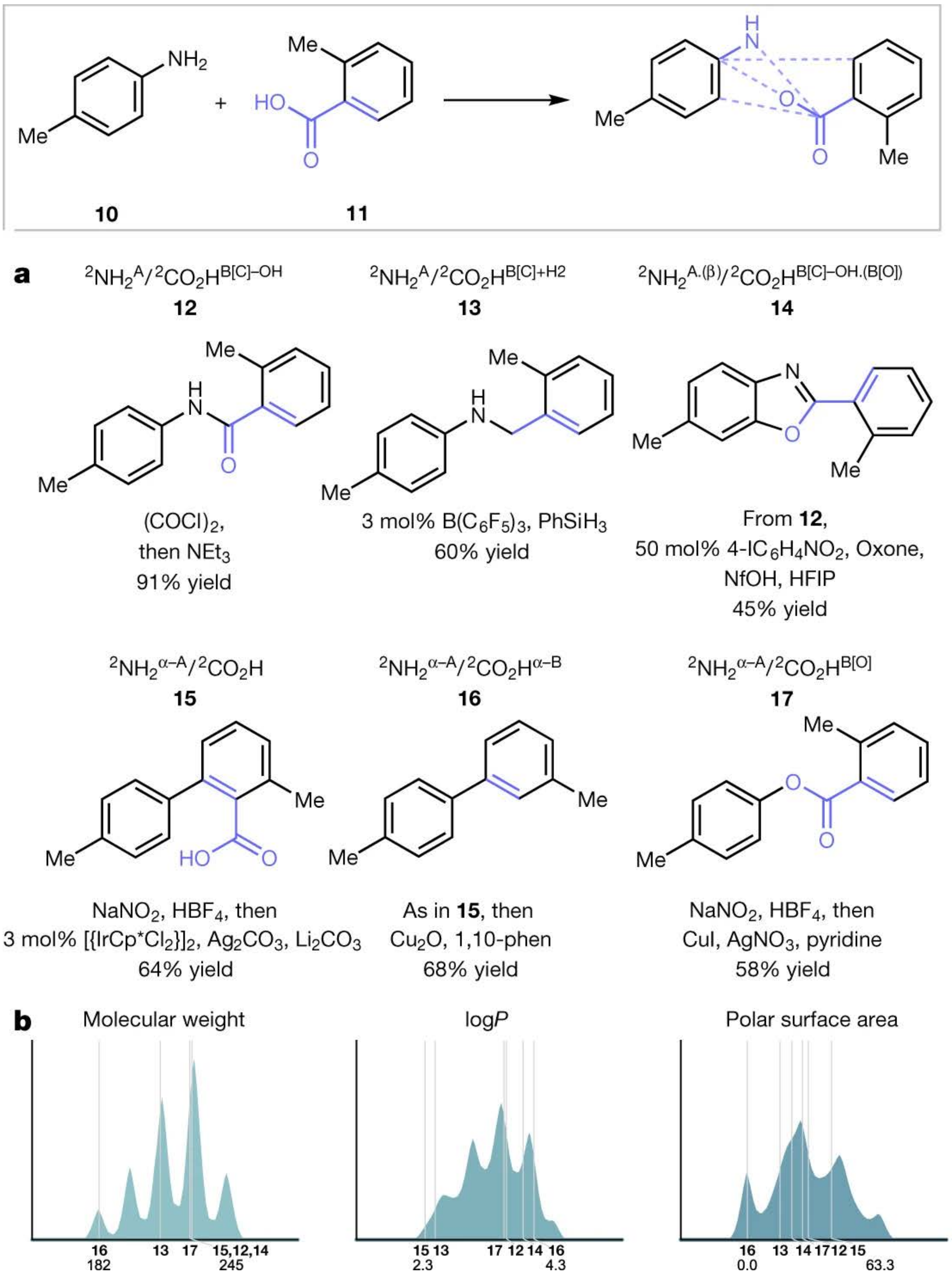
c Transformation notation

NH2
HO
Me
O
1
2
3
HN
4
5
6
7
8
9
$^3NH_2{}^A/^3CO_2H^{\alpha-B}$
$^3NH_2{}^{\alpha-A}/^3CO_2H^{B[O]}$
$^3NH_2{}^{\alpha-A}/^3CO_2H^{\alpha-B}$
$^3NH_2{}^{\alpha}/^3CO_2H^{\alpha}$
$^3NH_2{}^{\alpha-A}/^3CO_2H^{B[C]-OH}$
$^3NH_2{}^{\alpha}/^3CO_2H^{B[C]+H}$
OH
A
B
h = hybridization (sp^2 or sp^3)
X = reacting atom (A, B[C], B[O], etc.)
Y = modification (–A, –B, –OH, +H, +H2)
$^hA^{XY}/^hB^{XY}$
$^3A^A/^3B^{\alpha-B}$
$^3A^A/^2B^{\alpha-B}$
$^2A^{\alpha-A}/^2B^{\alpha-B}$
$^3A^{\alpha}/^3B^{\alpha}$
$^3A^{\alpha-A}/^3B^{\alpha}$

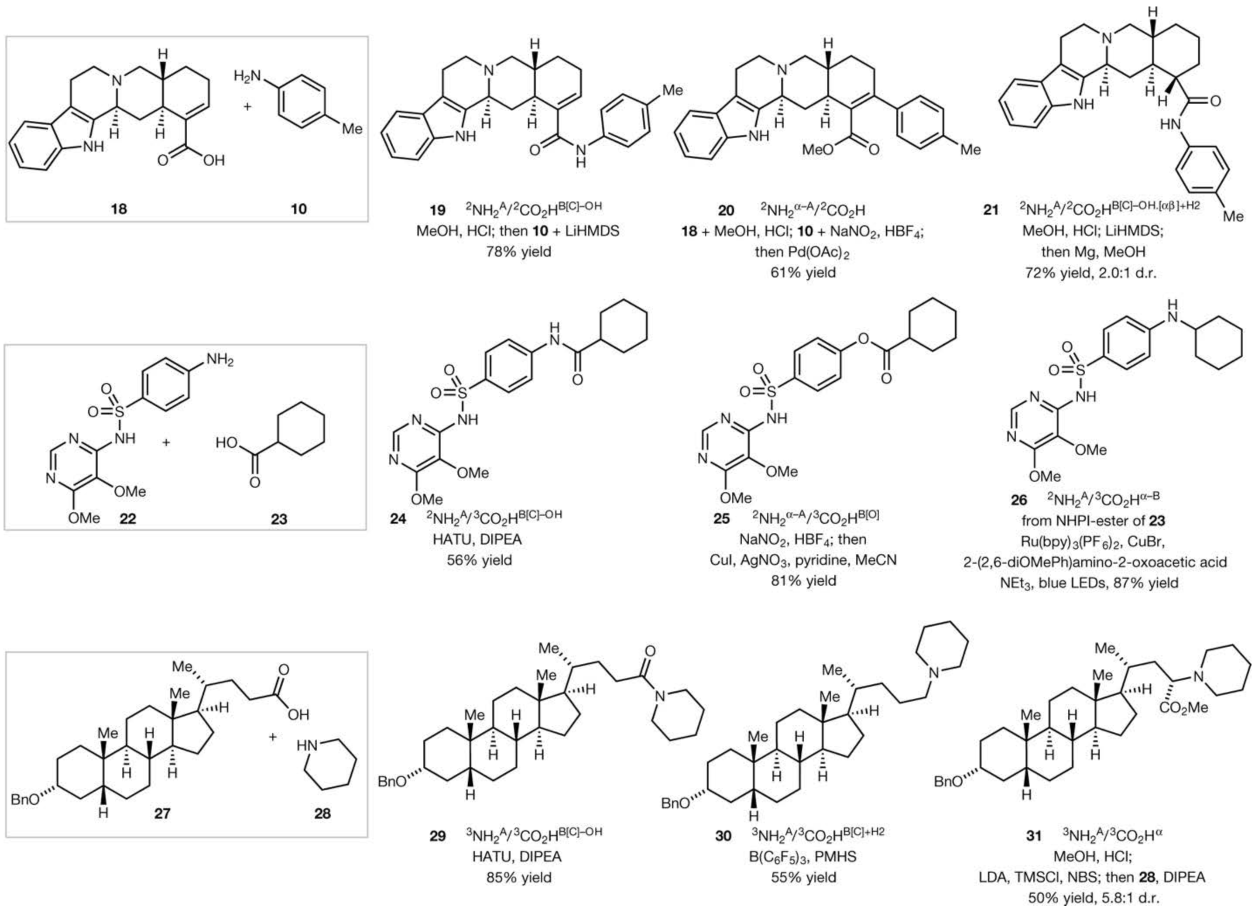
2-Fig. 1 | Transformation enumeration strategy and notation. a, Ethylamine (**1**) and propanoic acid (**2**) can couple to form amide **3**, but can also couple to form 79 other products, including **4**–**9**. **b**, Enumerating all combinations of sp^2 or sp^3 hybridization for the 80 coupling patterns yields 320 product substructures. **c**, A notation system for classifying transformations; see also 6-Extended Data Fig. 1.



3-Fig. 2 | Substructure search of 320 amine–acid coupling transformations within 9,279 complex molecules from DrugBank. Each line represents the appearance of a product substructure of a transformation in a complex molecule, and the colour of the line represents the frequency of occurrence of that substructure in that molecule. The dots around the periphery denote which specific transformations appear in complex molecules (+)-noscapine (green dots), (−)-quinine (purple dots) and (−)-sitagliptin (blue dots), which connect to 112, 96 and 55 transformations, respectively. The numbers around the periphery can be matched to a full list of transformation notation labels found in Extended Data Table 1.

NH$_2$
Me
+
HO
O
Me
H
N
O
O
Me
Me
10
11
a
^{2}NH$_2$^{A}/^{2}CO$_2$H$^{B[C]-OH}$
12
N
H
O
Me
Me
$(COCl)_2$,
then NEt_3
91% yield
^{2}NH$_2$^{A}/^{2}CO$_2$H$^{B[C]+H2}$
13
N
H
Me
Me
3 mol% $B(C_6F_5)_3$, $PhSiH_3$
60% yield
^{2}NH$_2$$^{A.(\beta)}$/^{2}CO$_2H^{B[C]-OH.(B[O])}$
14
N
O
Me
Me
From **12**,
50 mol% 4-$IC_6H_4NO_2$, Oxone,
NfOH, HFIP
45% yield
^{2}NH$_2$$^{\alpha-A}$/^{2}CO$_2$H
15
Me
HO
O
Me
NaNO$_2$, HBF$_4$, then
3 mol% $[\{IrCp^*Cl_2\}]_2$, Ag_2CO_3, Li_2CO_3
64% yield
^{2}NH$_2$$^{\alpha-A}$/^{2}CO$_2H^{\alpha-B}$
16
Me
Me
As in **15**, then
Cu_2O, 1,10-phen
68% yield
^{2}NH$_2$$^{\alpha-A}$/^{2}CO$_2H^{B[O]}$
17
Me
O
O
Me
NaNO$_2$, HBF$_4$, then
CuI, AgNO$_3$, pyridine
58% yield
b
Molecular weight
16
13
17
15,12,14
182
245
log*P*
15 13
17 12 14 16
2.3
4.3
Polar surface area
16
13 14 17 12 15
0.0
63.3

4-Fig. 3 | Experimental exploration of the *sp*^{2}–*sp*^{2} amine–acid coupling transformation space. **a**, Products with different property profiles (for example, basic, acidic, neutral, lipophilic) can be produced from the same two starting materials, **10** and **11** (top), by varying the transformation and reaction conditions. NfOH, nonafluorobutanesulfonic acid; HFIP, hexafluoroisopropanol; Cp*, pentamethylcyclopentadiene; 1,10-phen, 1,10-phenanthroline. **b**. Kernel density estimation plots show the range of accessible molecular weight, partition coefficient (log*P*) and polar surface area by coupling **10** to **11** via various amine–acid coupling transformations. Grey lines denote the molecular weight, log*P* and polar surface area of **12**–**17**. The calculations use **13** and **15** in their charged protonation state.



18 **10**

19 $^{2}NH_2{}^{A}/^{2}CO_2H^{B[C]-OH}$
MeOH, HCl; then **10** + LiHMDS
78% yield

20 $^{2}NH_2{}^{\alpha-A}/^{2}CO_2H$
18 + MeOH, HCl; **10** + NaNO$_2$, HBF$_4$;
then Pd(OAc)$_2$
61% yield

21 $^{2}NH_2{}^{A}/^{2}CO_2H^{B[C]-OH.[\alpha\beta]+H2}$
MeOH, HCl; LiHMDS;
then Mg, MeOH
72% yield, 2.0:1 d.r.

22 **23**

24 $^{2}NH_2{}^{A}/^{3}CO_2H^{B[C]-OH}$
HATU, DIPEA
56% yield

25 $^{2}NH_2{}^{\alpha-A}/^{3}CO_2H^{B[O]}$
NaNO$_2$, HBF$_4$; then
CuI, AgNO$_3$, pyridine, MeCN
81% yield

26 $^{2}NH_2{}^{A}/^{3}CO_2H^{\alpha-B}$
from NHPI-ester of **23**
Ru(bpy)$_3$(PF$_6$)$_2$, CuBr,
2-(2,6-diOMePh)amino-2-oxoacetic acid
NEt$_3$, blue LEDs, 87% yield

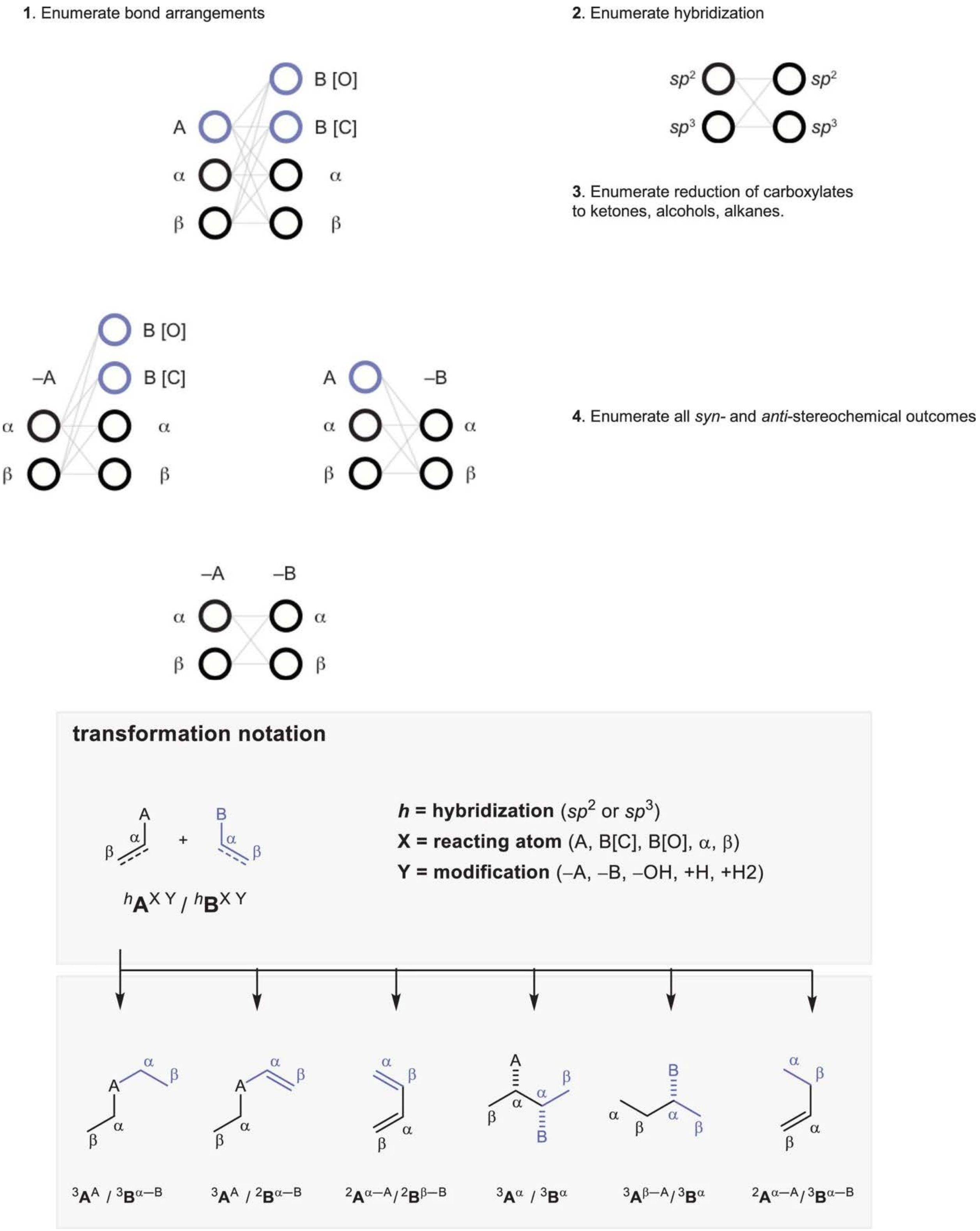
27 **28**

29 $^{3}NH_2{}^{A}/^{3}CO_2H^{B[C]-OH}$
HATU, DIPEA
85% yield

30 $^{3}NH_2{}^{A}/^{3}CO_2H^{B[C]+H2}$
B(C$_6$F$_5$)$_3$, PMHS
55% yield

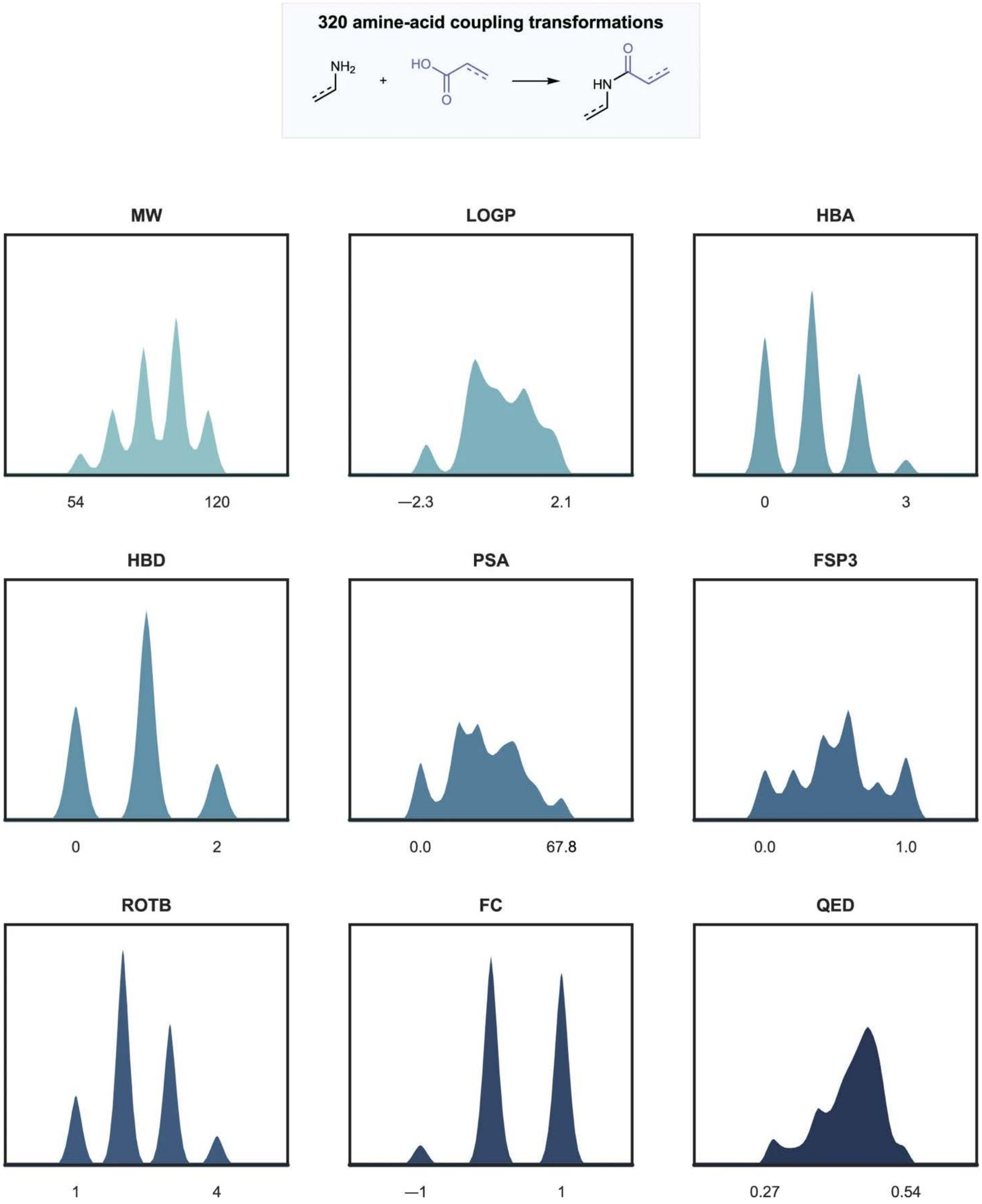
31 $^{3}NH_2{}^{A}/^{3}CO_2H^{\alpha}$
MeOH, HCl;
LDA, TMSCl, NBS; then **28**, DIPEA
50% yield, 5.8:1 d.r.

5-Fig. 4 | Late-stage diversification. Various transformations enable the diversification of the complex molecules **18**, **22** and **27**. We performed a virtual enumeration of other complex molecules—shown in 14-Extended Data Fig. 9—wherein the full transformation set was enumerated for four complex molecule pairings to demonstrate that a wide range of properties can be accessed, depending on which transformation is selected. LiHMDS, lithium bis(trimethylsilyl)amide; HATU, 1-[bis(dimethylamino)methylene]-1*H*-1,2,3-triazolo[4,5-*b*]pyridinium 3-oxide hexafluorophosphate; DIPEA, *N*,*N*-diisopropylethylamine; NHPI, *N*-hydroxyphthalimide; PMHS, polymethylhydrosiloxane; LDA, lithium diisopropylamide; TMSCl, trimethylsilyl chloride; NBS, *N*-bromosuccinimide.

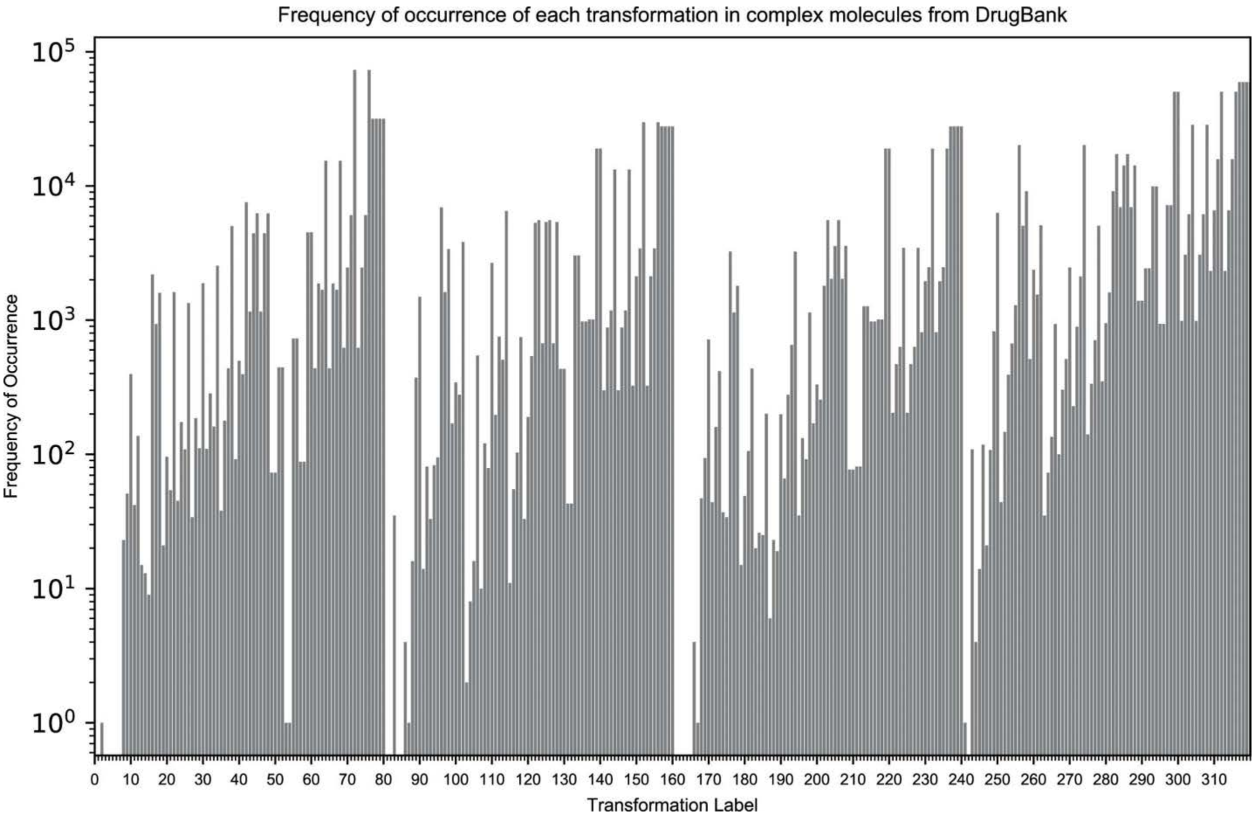

1. Enumerate bond arrangements
A
α
β
B [O]
B [C]
α
β
–A
B [O]
B [C]
α
β
α
β
A
–B
α
β
α
β
–A
–B
α
β
α
β
2. Enumerate hybridization
sp^2
sp^3
sp^2
sp^3
3. Enumerate reduction of carboxylates to ketones, alcohols, alkanes.
4. Enumerate all syn- and anti-stereochemical outcomes
transformation notation
A
α
β
+
B
α
β
$^h\mathbf{A}^{X\,Y}$ / $^h\mathbf{B}^{X\,Y}$
h = hybridization (sp^2 or sp^3)
X = reacting atom (A, B[C], B[O], α, β)
Y = modification (–A, –B, –OH, +H, +H2)
$^3\mathbf{A}^{A}$ / $^3\mathbf{B}^{\alpha-B}$
$^3\mathbf{A}^{A}$ / $^2\mathbf{B}^{\alpha-B}$
$^2\mathbf{A}^{\alpha-A}$ / $^2\mathbf{B}^{\beta-B}$
$^3\mathbf{A}^{\alpha}$ / $^3\mathbf{B}^{\alpha}$
$^3\mathbf{A}^{\beta-A}$ / $^3\mathbf{B}^{\alpha}$
$^2\mathbf{A}^{\alpha-A}$ / $^3\mathbf{B}^{\alpha-B}$

6-Extended Data Fig. 1 | Workflow for enumeration of amine–acid transformations.

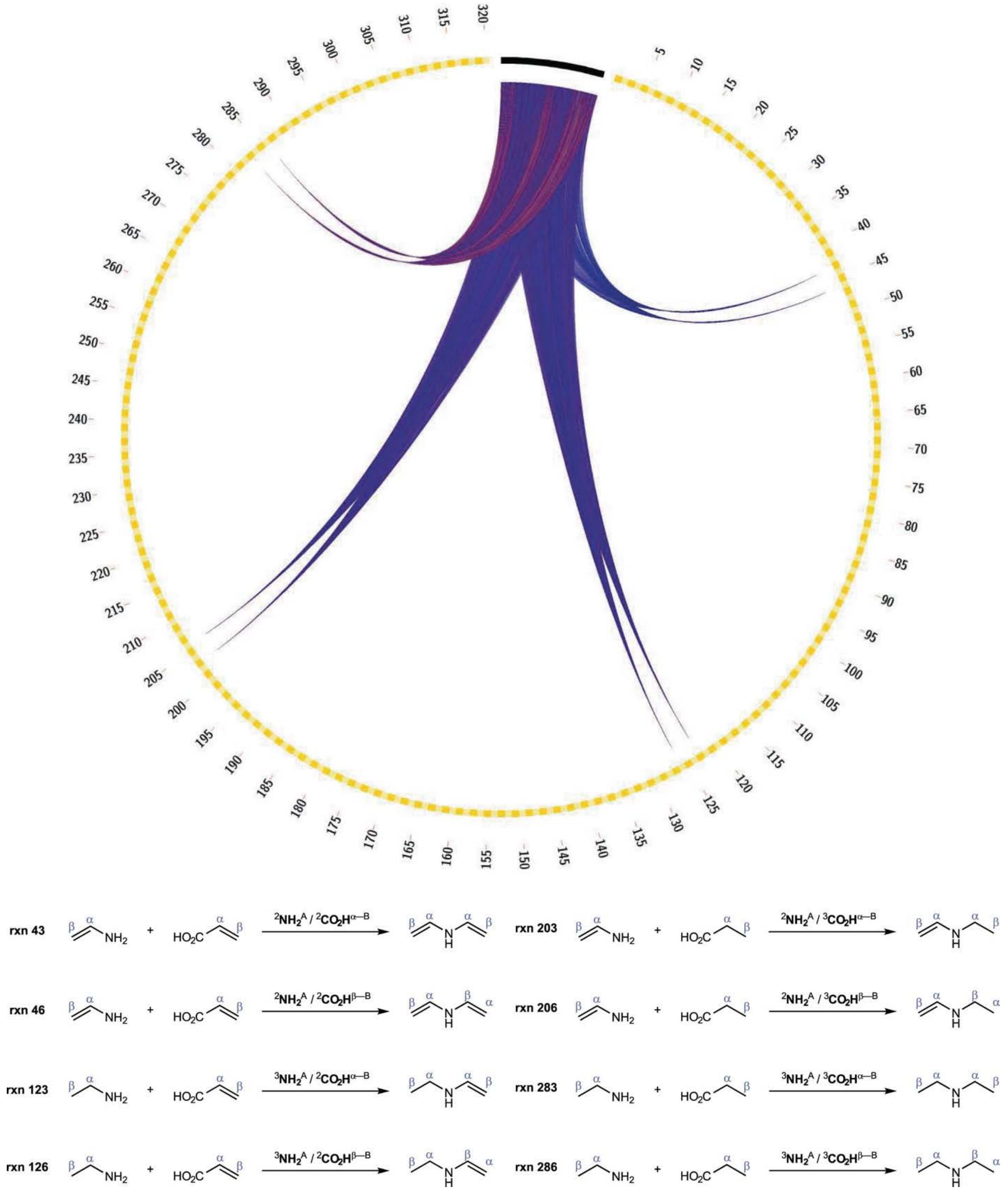
For a pair of coupling partners, we consider a reaction at the functional group A (amine) and B (carboxylic acid oxygen, B[O] or carbon, B[C]). Deamination reactions are noted as −A and decarboxylation reactions are noted as −B. Enumeration following steps 1–3 produces 320 transformations. For the enumeration of all *syn*- and *anti*-diastereomers (step 4), consult also 11-Extended Data Fig. 6.



7-Extended Data Fig. 2 | Kernel density estimate plots for 320 conceivable amine–acid coupling transformations. Distribution of common physical properties from the achiral amine–acid coupling of ethylamine, ethenamine, propanoic acid and acrylic acid. MW, molecular weight; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; PSA, polar surface area; FSP3, fraction *sp*^{3}; ROTB, rotatable bonds; FC, formal charge; QED, quantitative estimate of drug-likeness.

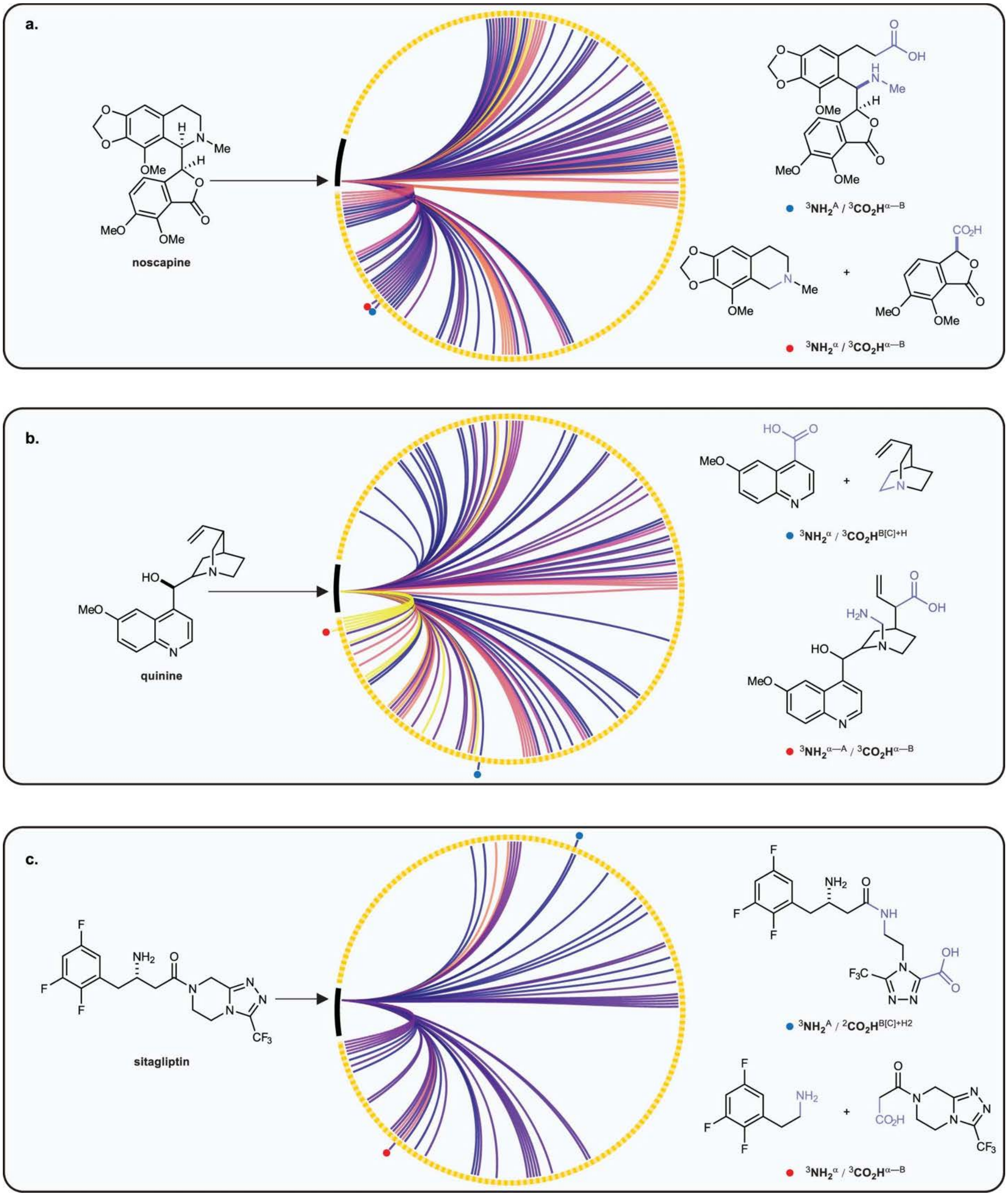


8-Extended Data Fig. 3 | Number of DrugBank hits per transformation. This bar chart shows how many times a transformation is found in the DrugBank database. Each number on the abscissa maps to a transformation listed in Extended Data Table 1.

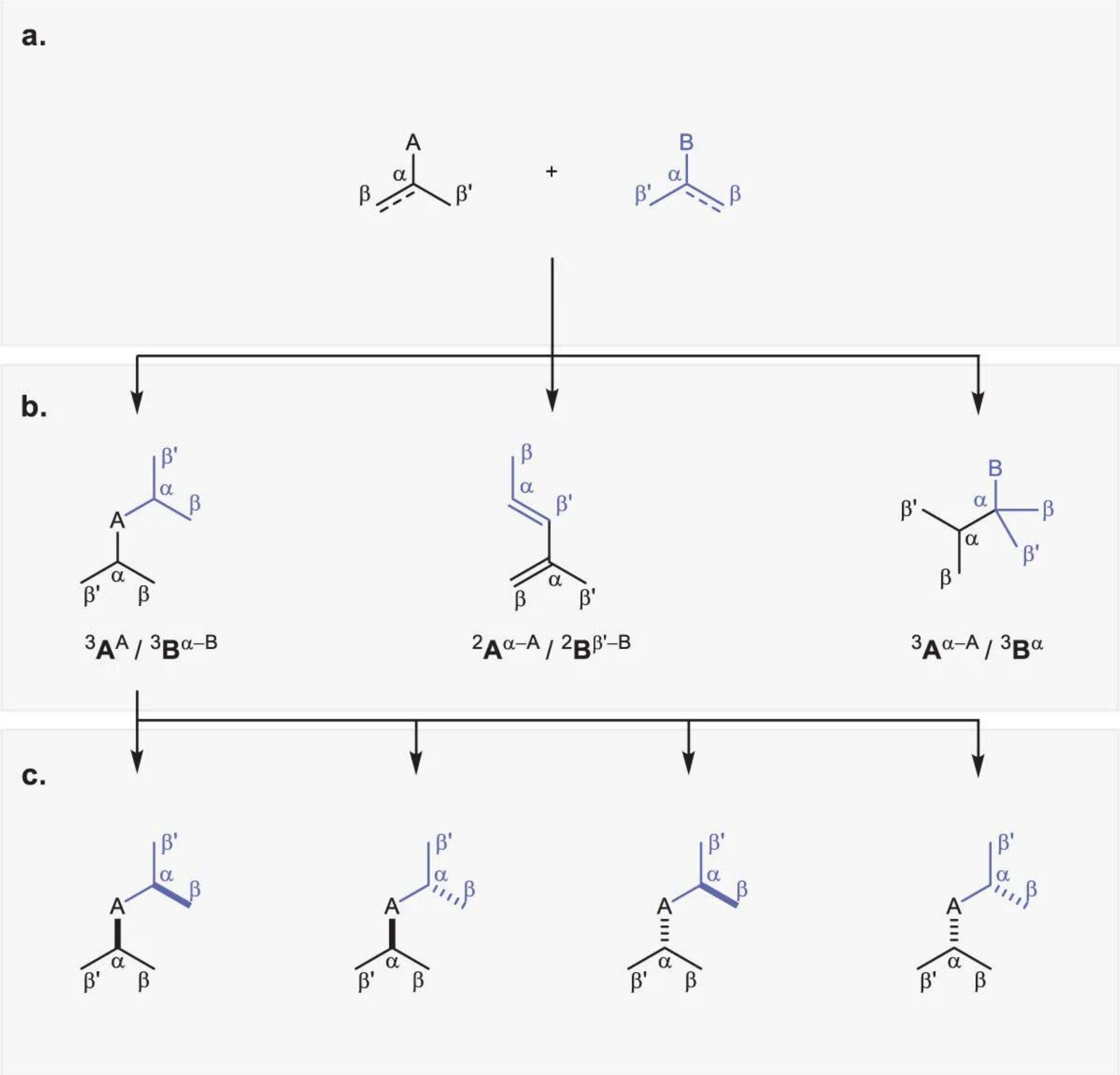


5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
rxn 43
$^2NH_2^A / ^2CO_2H^{\alpha—B}$
rxn 46
$^2NH_2^A / ^2CO_2H^{\beta—B}$
rxn 123
$^3NH_2^A / ^2CO_2H^{\alpha—B}$
rxn 126
$^3NH_2^A / ^2CO_2H^{\beta—B}$
rxn 203
$^2NH_2^A / ^3CO_2H^{\alpha—B}$
rxn 206
$^2NH_2^A / ^3CO_2H^{\beta—B}$
rxn 283
$^3NH_2^A / ^3CO_2H^{\alpha—B}$
rxn 286
$^3NH_2^A / ^3CO_2H^{\beta—B}$

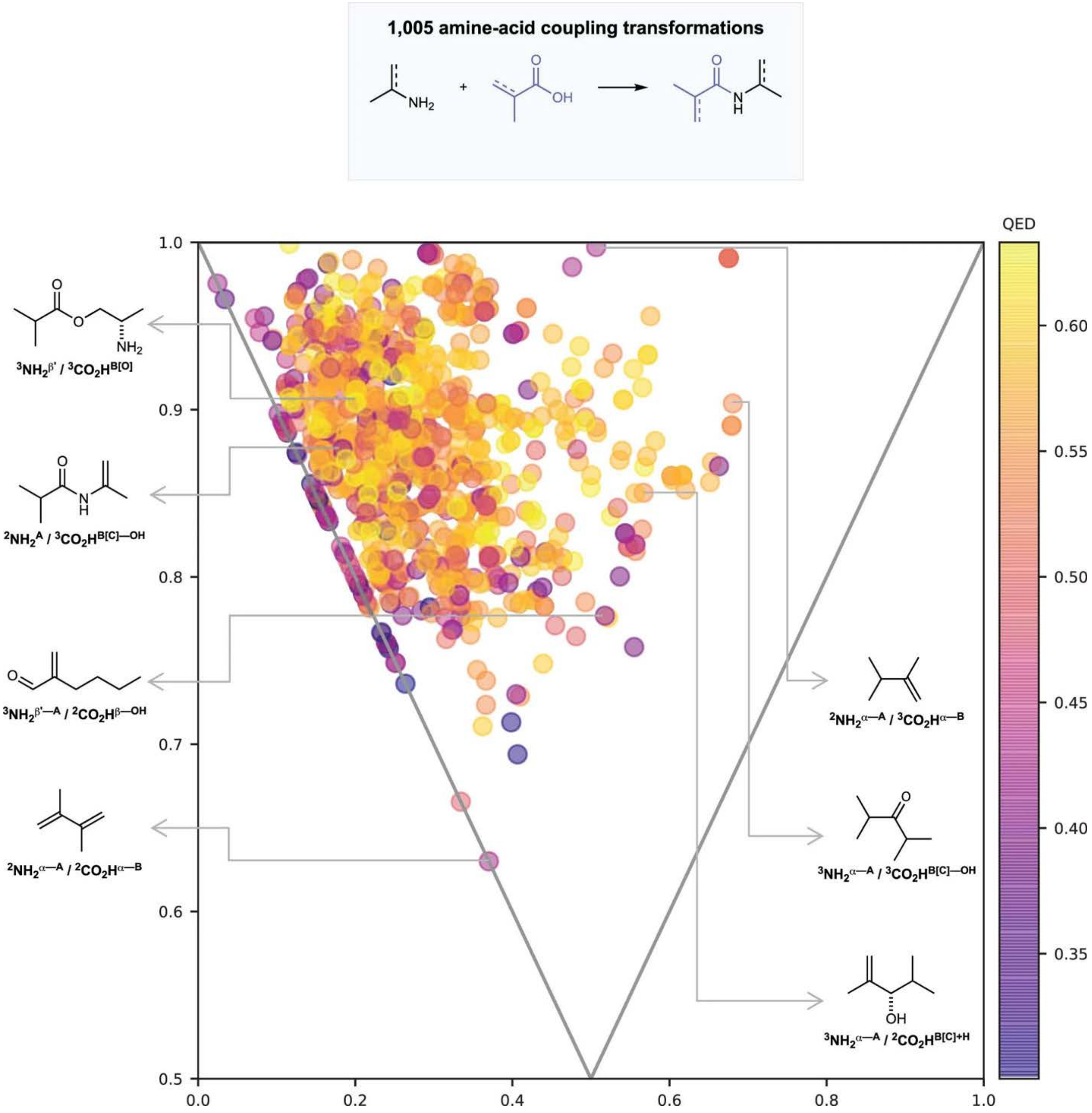
9-Extended Data Fig. 4 | Decarboxylative transformations from the enumeration scheme. Decarboxylative reactions that produce an amine bound to an sp^3 or sp^2 carbon chain appear in high frequency. These reactions can be used to synthesize a large number of drugs contained in DrugBank. Each transformation can be found by its corresponding number in Extended Data Table 1. The colour scale is the same as in 3-Fig. 2. rxn, reaction.


a.
noscapine
● $^3NH_2^A$ / $^3CO_2H^{\alpha—B}$
● $^3NH_2^\alpha$ / $^3CO_2H^{\alpha—B}$
b.
quinine
● $^3NH_2^\alpha$ / $^3CO_2H^{B[C]+H}$
● $^3NH_2^{\alpha—A}$ / $^3CO_2H^{\alpha—B}$
c.
sitagliptin
● $^3NH_2^A$ / $^2CO_2H^{B[C]+H2}$
● $^3NH_2^\alpha$ / $^3CO_2H^{\alpha—B}$

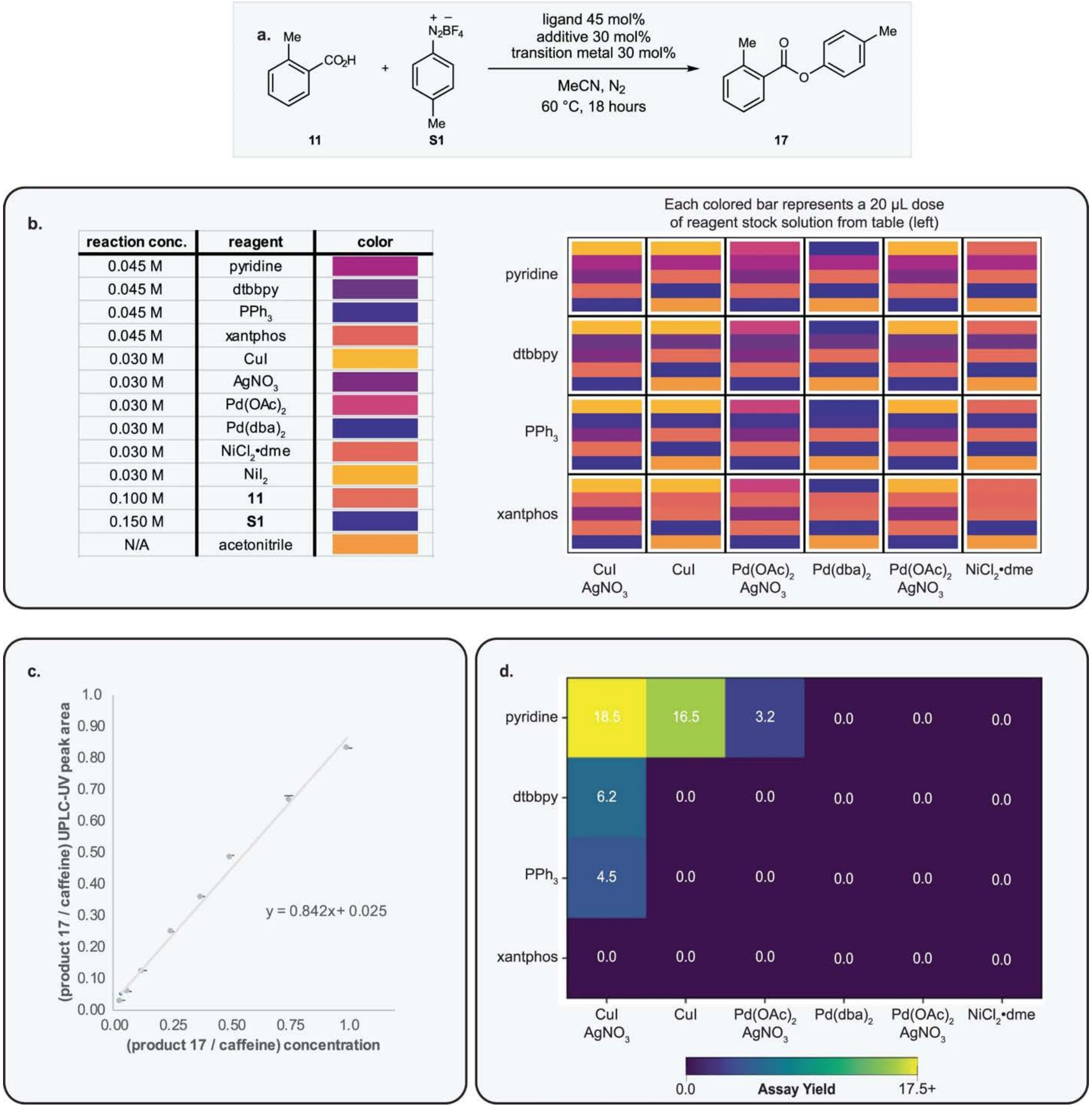
10-Extended Data Fig. 5 | Transformations from the enumeration scheme found in specific drugs. The chord diagrams show connectivity of transformation substructures as retrosynthetic disconnections in target molecules, with red and blue dots highlighting the transformations shown at left in each panel. **a**, Noscapine connects to 112 of the transformations. **b**, Quinine connects to 96 transformations. **c**, Sitagliptin connects to 55 transformations. The colour scale is the same as in 3-Fig. 2.


a.
A
α
β
β'
+
B
α
β'
β
b.
$^3\mathbf{A}^{A} / {}^3\mathbf{B}^{\alpha\text{–}B}$
$^2\mathbf{A}^{\alpha\text{–}A} / {}^2\mathbf{B}^{\beta'\text{–}B}$
$^3\mathbf{A}^{\alpha\text{–}A} / {}^3\mathbf{B}^{\alpha}$
c.

11-Extended Data Fig. 6 | Enumeration of regioisomers and diastereomers. a, The transformation substructures enumerated in 4-Fig. 3 are from the 320 achiral bond arrangements available from coupling **1**, **2** and their *sp*^{2} variants ethenamine and acrylic acid. **b**, To sample three-dimensional and regiochemical space, a β′ substituent was added as a differentiating substituent. The β′ substituent may be any substituent, but is enumerated as being distinct from the β substituent. Considering this regiochemical enumeration increases the 320 achiral coupling transformations to 588. **c**, Subsequent enumeration of all possible diastereomers leads to 1,005 chiral coupling transformations. These 1,005 three-dimensional substructures were used as inputs in the principal moment of inertia plot in 12-Extended Data Fig. 7.



12-Extended Data Fig. 7 | Principal moment of inertia plot of 1,005 amine–acid coupling transformations incorporating stereochemistry and regiochemistry. In this expanded three-dimensional space, regiochemistry and stereochemistry of the transformations were considered. A total of 1,005 ways to connect an amine to an acid were found. The products presented a diversity of properties and three-dimensional shapes. Each molecule is coloured by its quantitative estimate of drug-likeness.

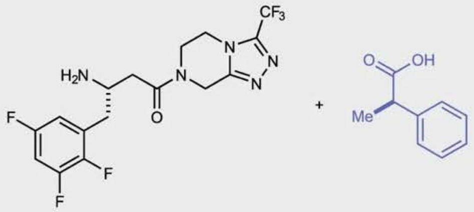


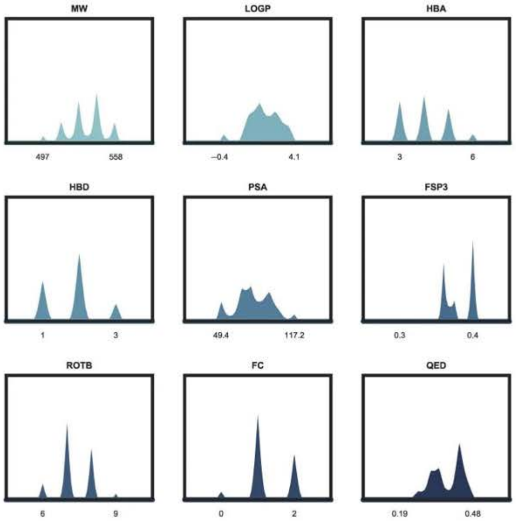
reaction conc.	reagent	color
0.045 M	pyridine	🟪
0.045 M	dtbbpy	🟪
0.045 M	PPh_3	🟦
0.045 M	xantphos	🟥
0.030 M	CuI	🟨
0.030 M	$AgNO_3$	🟪
0.030 M	$Pd(OAc)_2$	🟥
0.030 M	$Pd(dba)_2$	🟦
0.030 M	$NiCl_2$•dme	🟥
0.030 M	NiI_2	🟨
0.100 M	**11**	🟥
0.150 M	**S1**	🟦
N/A	acetonitrile	🟧

	CuI AgNO$_3$	CuI	Pd(OAc)$_2$ AgNO$_3$	Pd(dba)$_2$	Pd(OAc)$_2$ AgNO$_3$	NiCl$_2$•dme
pyridine	18.5	16.5	3.2	0.0	0.0	0.0
dtbbpy	6.2	0.0	0.0	0.0	0.0	0.0
PPh_3	4.5	0.0	0.0	0.0	0.0	0.0
xantphos	0.0	0.0	0.0	0.0	0.0	0.0

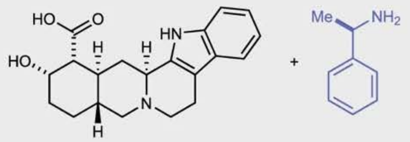
13-Extended Data Fig. 8 | High-throughput experimentation for the discovery of a copper-promoted esterification reaction. a, An esterification reaction discovered through reaction screening of transition metals with ligands and additives. **b**, Recipe and well mapping. **c**, Calibration curve, for product **17** versus caffeine internal standard, used to convert the ultraperformance liquid chromatography with ultraviolet–visible spectrometry peak area to concentration, and thus to assay yield. Error bars show deviation among triplicate injections. **d**, Heat map depicting assay yield screening results. CuI with $AgNO_3$ and pyridine showed the most promising results, achieving 18.5% assay yield using 30 mol% CuI with $AgNO_3$.

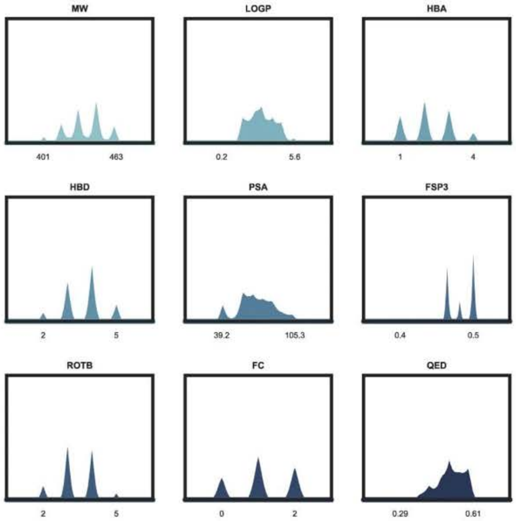
112 amine-acid late stage coupling transformations



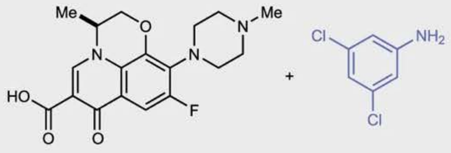


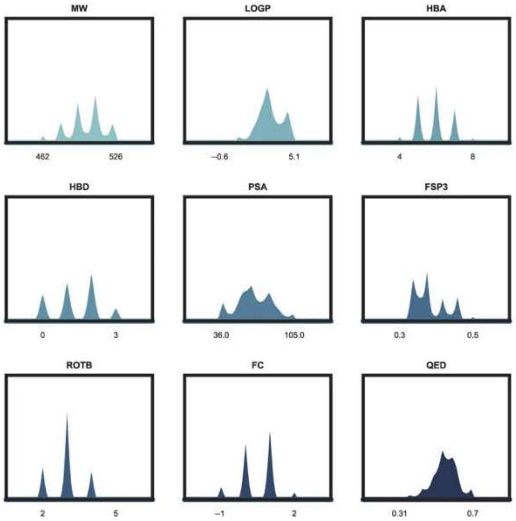
105 amine-acid late stage coupling transformations



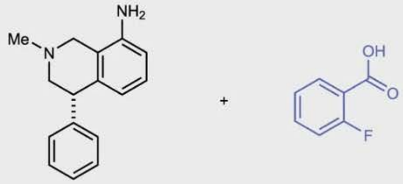


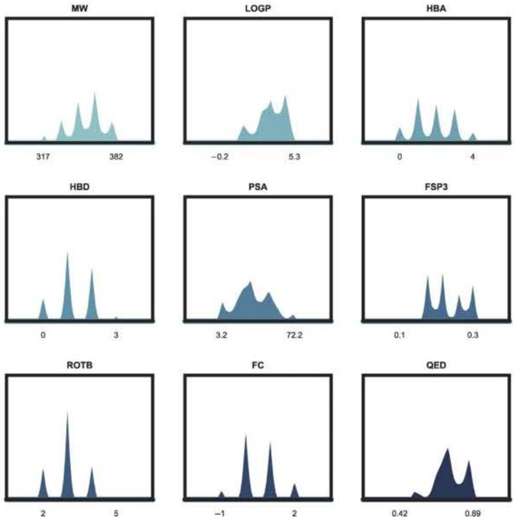
77 amine-acid late stage coupling transformations



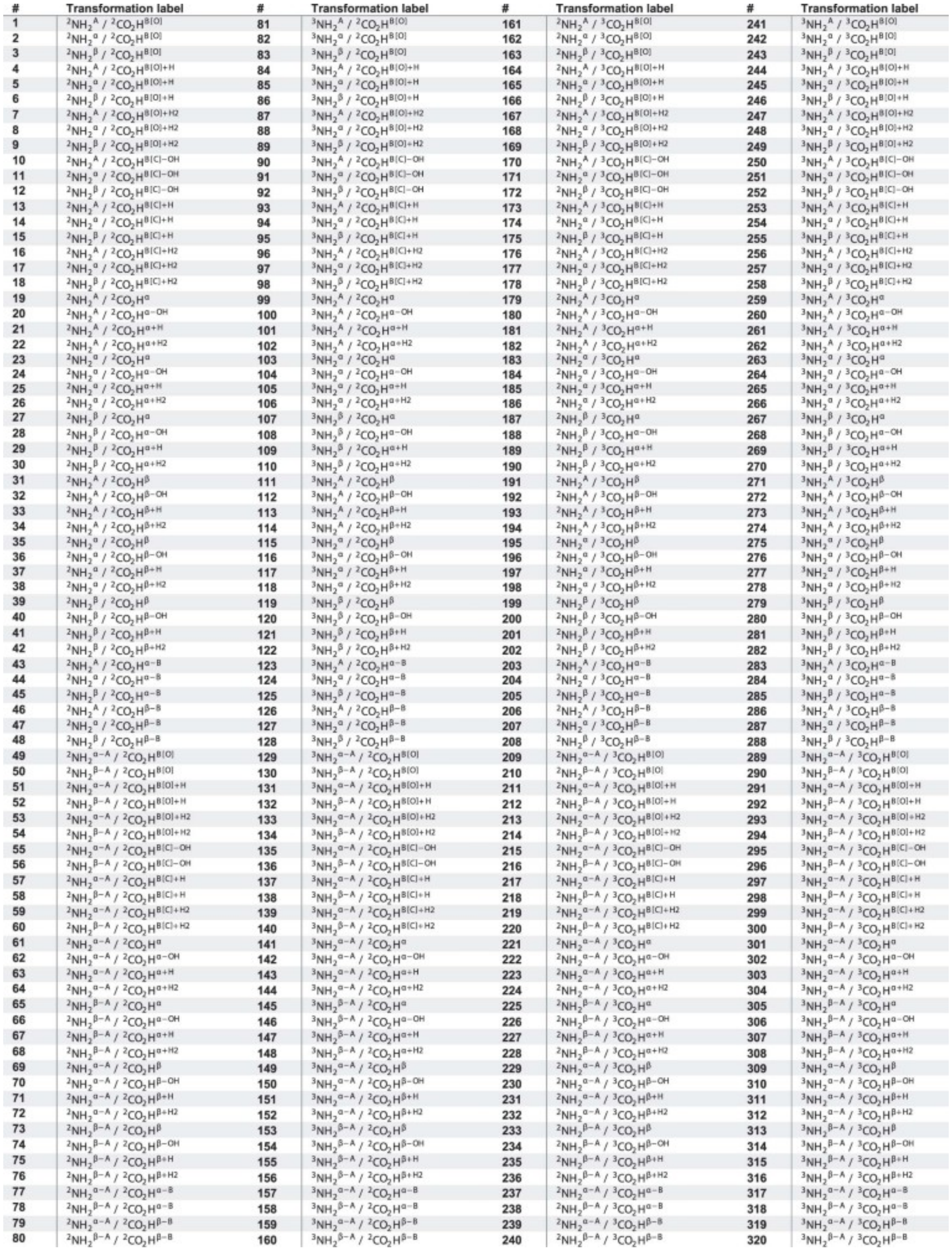


80 amine-acid late stage coupling transformations





14-Extended Data Fig. 9 | Kernel density estimate plots of a series of complex molecules as substrates in the amine–acid coupling system. The amine–acid pair depicted was used as an input to combinatorial enumeration, and the number of valid products identified is noted for each pairing. Distributions of common physical properties are shown for each coupling set. Abbreviations are as in 7-Extended Data Fig. 2.

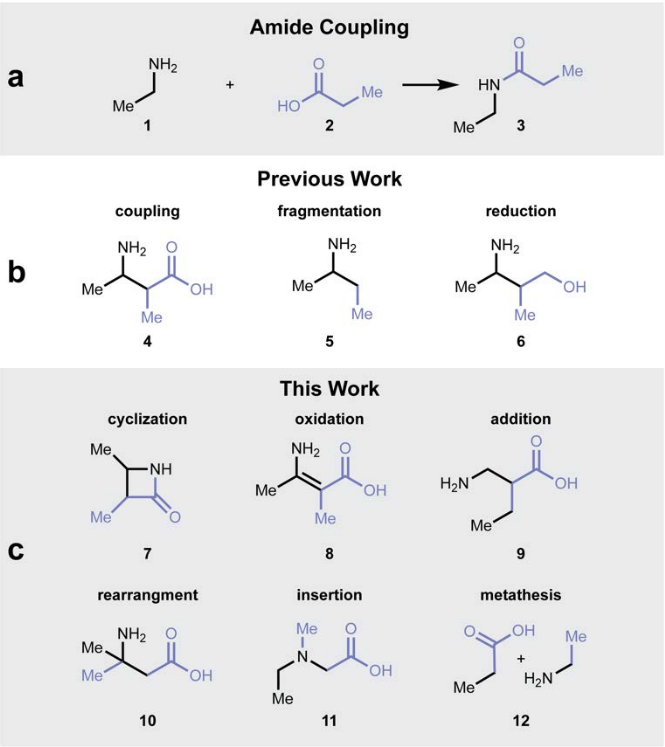


#	Transformation label	#	Transformation label	#	Transformation label	#	Transformation label
1	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[O]}$	81	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[O]}$	161	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[O]}$	241	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[O]}$
2	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]}$	82	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]}$	162	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]}$	242	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]}$
3	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]}$	83	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]}$	163	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]}$	243	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]}$
4	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[O]+H}$	84	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[O]+H}$	164	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[O]+H}$	244	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[O]+H}$
5	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]+H}$	85	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]+H}$	165	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]+H}$	245	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]+H}$
6	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]+H}$	86	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]+H}$	166	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]+H}$	246	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]+H}$
7	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[O]+H2}$	87	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[O]+H2}$	167	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[O]+H2}$	247	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[O]+H2}$
8	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]+H2}$	88	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[O]+H2}$	168	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]+H2}$	248	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[O]+H2}$
9	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]+H2}$	89	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[O]+H2}$	169	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]+H2}$	249	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[O]+H2}$
10	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[C]-OH}$	90	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[C]-OH}$	170	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[C]-OH}$	250	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[C]-OH}$
11	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]-OH}$	91	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]-OH}$	171	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]-OH}$	251	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]-OH}$
12	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]-OH}$	92	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]-OH}$	172	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]-OH}$	252	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]-OH}$
13	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[C]+H}$	93	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[C]+H}$	173	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[C]+H}$	253	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[C]+H}$
14	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]+H}$	94	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]+H}$	174	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]+H}$	254	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]+H}$
15	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]+H}$	95	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]+H}$	175	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]+H}$	255	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]+H}$
16	$^{2}NH_2^{A}$ / $^{2}CO_2H^{B[C]+H2}$	96	$^{3}NH_2^{A}$ / $^{2}CO_2H^{B[C]+H2}$	176	$^{2}NH_2^{A}$ / $^{3}CO_2H^{B[C]+H2}$	256	$^{3}NH_2^{A}$ / $^{3}CO_2H^{B[C]+H2}$
17	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]+H2}$	97	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{B[C]+H2}$	177	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]+H2}$	257	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{B[C]+H2}$
18	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]+H2}$	98	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{B[C]+H2}$	178	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]+H2}$	258	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{B[C]+H2}$
19	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\alpha}$	99	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\alpha}$	179	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\alpha}$	259	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\alpha}$
20	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\alpha-OH}$	100	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\alpha-OH}$	180	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\alpha-OH}$	260	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\alpha-OH}$
21	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\alpha+H}$	101	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\alpha+H}$	181	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\alpha+H}$	261	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\alpha+H}$
22	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\alpha+H2}$	102	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\alpha+H2}$	182	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\alpha+H2}$	262	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\alpha+H2}$
23	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha}$	103	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha}$	183	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha}$	263	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha}$
24	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha-OH}$	104	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha-OH}$	184	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha-OH}$	264	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha-OH}$
25	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha+H}$	105	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha+H}$	185	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha+H}$	265	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha+H}$
26	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha+H2}$	106	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha+H2}$	186	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha+H2}$	266	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha+H2}$
27	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha}$	107	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha}$	187	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha}$	267	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha}$
28	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha-OH}$	108	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha-OH}$	188	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha-OH}$	268	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha-OH}$
29	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha+H}$	109	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha+H}$	189	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha+H}$	269	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha+H}$
30	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha+H2}$	110	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha+H2}$	190	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha+H2}$	270	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha+H2}$
31	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\beta}$	111	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\beta}$	191	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\beta}$	271	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\beta}$
32	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\beta-OH}$	112	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\beta-OH}$	192	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\beta-OH}$	272	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\beta-OH}$
33	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\beta+H}$	113	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\beta+H}$	193	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\beta+H}$	273	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\beta+H}$
34	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\beta+H2}$	114	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\beta+H2}$	194	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\beta+H2}$	274	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\beta+H2}$
35	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta}$	115	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta}$	195	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta}$	275	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta}$
36	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta-OH}$	116	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta-OH}$	196	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta-OH}$	276	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta-OH}$
37	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta+H}$	117	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta+H}$	197	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta+H}$	277	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta+H}$
38	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta+H2}$	118	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta+H2}$	198	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta+H2}$	278	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta+H2}$
39	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\beta}$	119	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\beta}$	199	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\beta}$	279	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\beta}$
40	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\beta-OH}$	120	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\beta-OH}$	200	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\beta-OH}$	280	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\beta-OH}$
41	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\beta+H}$	121	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\beta+H}$	201	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\beta+H}$	281	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\beta+H}$
42	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\beta+H2}$	122	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\beta+H2}$	202	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\beta+H2}$	282	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\beta+H2}$
43	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\alpha-B}$	123	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\alpha-B}$	203	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\alpha-B}$	283	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\alpha-B}$
44	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha-B}$	124	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\alpha-B}$	204	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha-B}$	284	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\alpha-B}$
45	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha-B}$	125	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\alpha-B}$	205	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha-B}$	285	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\alpha-B}$
46	$^{2}NH_2^{A}$ / $^{2}CO_2H^{\beta-B}$	126	$^{3}NH_2^{A}$ / $^{2}CO_2H^{\beta-B}$	206	$^{2}NH_2^{A}$ / $^{3}CO_2H^{\beta-B}$	286	$^{3}NH_2^{A}$ / $^{3}CO_2H^{\beta-B}$
47	$^{2}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta-B}$	127	$^{3}NH_2^{\alpha}$ / $^{2}CO_2H^{\beta-B}$	207	$^{2}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta-B}$	287	$^{3}NH_2^{\alpha}$ / $^{3}CO_2H^{\beta-B}$
48	$^{2}NH_2^{\beta}$ / $^{2}CO_2H^{\beta-B}$	128	$^{3}NH_2^{\beta}$ / $^{2}CO_2H^{\beta-B}$	208	$^{2}NH_2^{\beta}$ / $^{3}CO_2H^{\beta-B}$	288	$^{3}NH_2^{\beta}$ / $^{3}CO_2H^{\beta-B}$
49	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]}$	129	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]}$	209	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]}$	289	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]}$
50	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]}$	130	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]}$	210	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]}$	290	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]}$
51	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]+H}$	131	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]+H}$	211	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]+H}$	291	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]+H}$
52	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]+H}$	132	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]+H}$	212	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]+H}$	292	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]+H}$
53	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]+H2}$	133	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[O]+H2}$	213	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]+H2}$	293	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[O]+H2}$
54	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]+H2}$	134	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[O]+H2}$	214	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]+H2}$	294	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[O]+H2}$
55	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]-OH}$	135	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]-OH}$	215	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]-OH}$	295	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]-OH}$
56	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]-OH}$	136	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]-OH}$	216	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]-OH}$	296	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]-OH}$
57	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]+H}$	137	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]+H}$	217	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]+H}$	297	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]+H}$
58	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]+H}$	138	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]+H}$	218	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]+H}$	298	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]+H}$
59	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]+H2}$	139	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{B[C]+H2}$	219	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]+H2}$	299	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{B[C]+H2}$
60	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]+H2}$	140	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{B[C]+H2}$	220	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]+H2}$	300	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{B[C]+H2}$
61	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha}$	141	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha}$	221	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha}$	301	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha}$
62	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha-OH}$	142	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha-OH}$	222	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha-OH}$	302	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha-OH}$
63	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha+H}$	143	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha+H}$	223	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha+H}$	303	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha+H}$
64	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha+H2}$	144	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha+H2}$	224	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha+H2}$	304	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha+H2}$
65	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha}$	145	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha}$	225	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha}$	305	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha}$
66	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha-OH}$	146	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha-OH}$	226	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha-OH}$	306	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha-OH}$
67	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha+H}$	147	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha+H}$	227	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha+H}$	307	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha+H}$
68	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha+H2}$	148	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha+H2}$	228	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha+H2}$	308	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha+H2}$
69	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta}$	149	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta}$	229	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta}$	309	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta}$
70	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta-OH}$	150	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta-OH}$	230	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta-OH}$	310	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta-OH}$
71	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta+H}$	151	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta+H}$	231	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta+H}$	311	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta+H}$
72	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta+H2}$	152	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta+H2}$	232	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta+H2}$	312	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta+H2}$
73	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta}$	153	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta}$	233	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta}$	313	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta}$
74	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta-OH}$	154	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta-OH}$	234	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta-OH}$	314	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta-OH}$
75	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta+H}$	155	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta+H}$	235	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta+H}$	315	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta+H}$
76	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta+H2}$	156	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta+H2}$	236	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta+H2}$	316	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta+H2}$
77	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha-B}$	157	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\alpha-B}$	237	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha-B}$	317	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\alpha-B}$
78	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha-B}$	158	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\alpha-B}$	238	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha-B}$	318	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\alpha-B}$
79	$^{2}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta-B}$	159	$^{3}NH_2^{\alpha-A}$ / $^{2}CO_2H^{\beta-B}$	239	$^{2}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta-B}$	319	$^{3}NH_2^{\alpha-A}$ / $^{3}CO_2H^{\beta-B}$
80	$^{2}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta-B}$	160	$^{3}NH_2^{\beta-A}$ / $^{2}CO_2H^{\beta-B}$	240	$^{2}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta-B}$	320	$^{3}NH_2^{\beta-A}$ / $^{3}CO_2H^{\beta-B}$

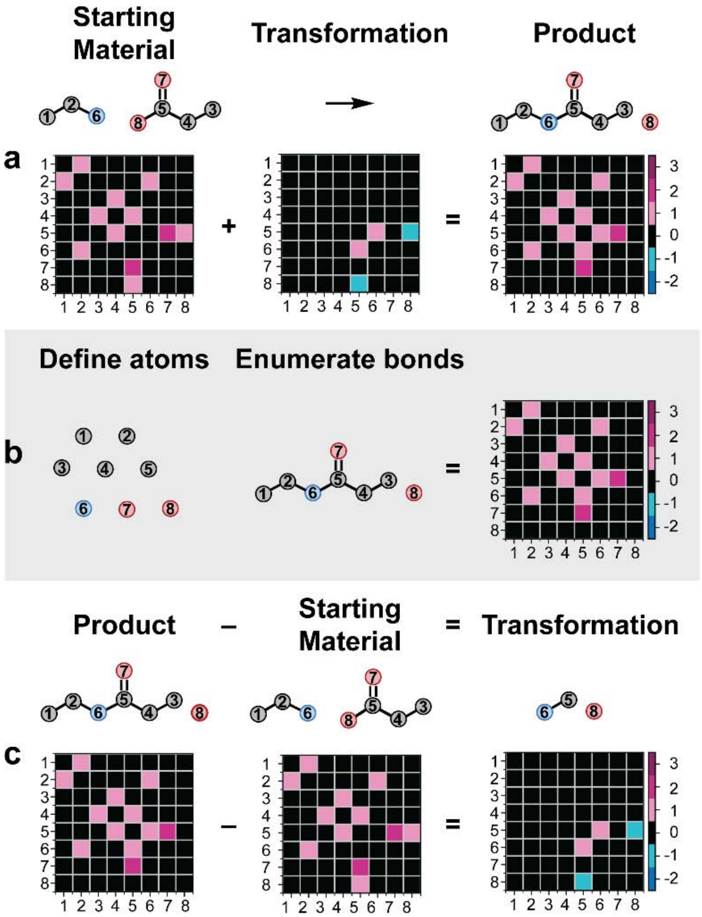
15-Extended Data Table 1 | Transformation labels. This table maps each transformation number from the periphery of the chord diagram in 3-Fig. 2 to a transformation label.

Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing

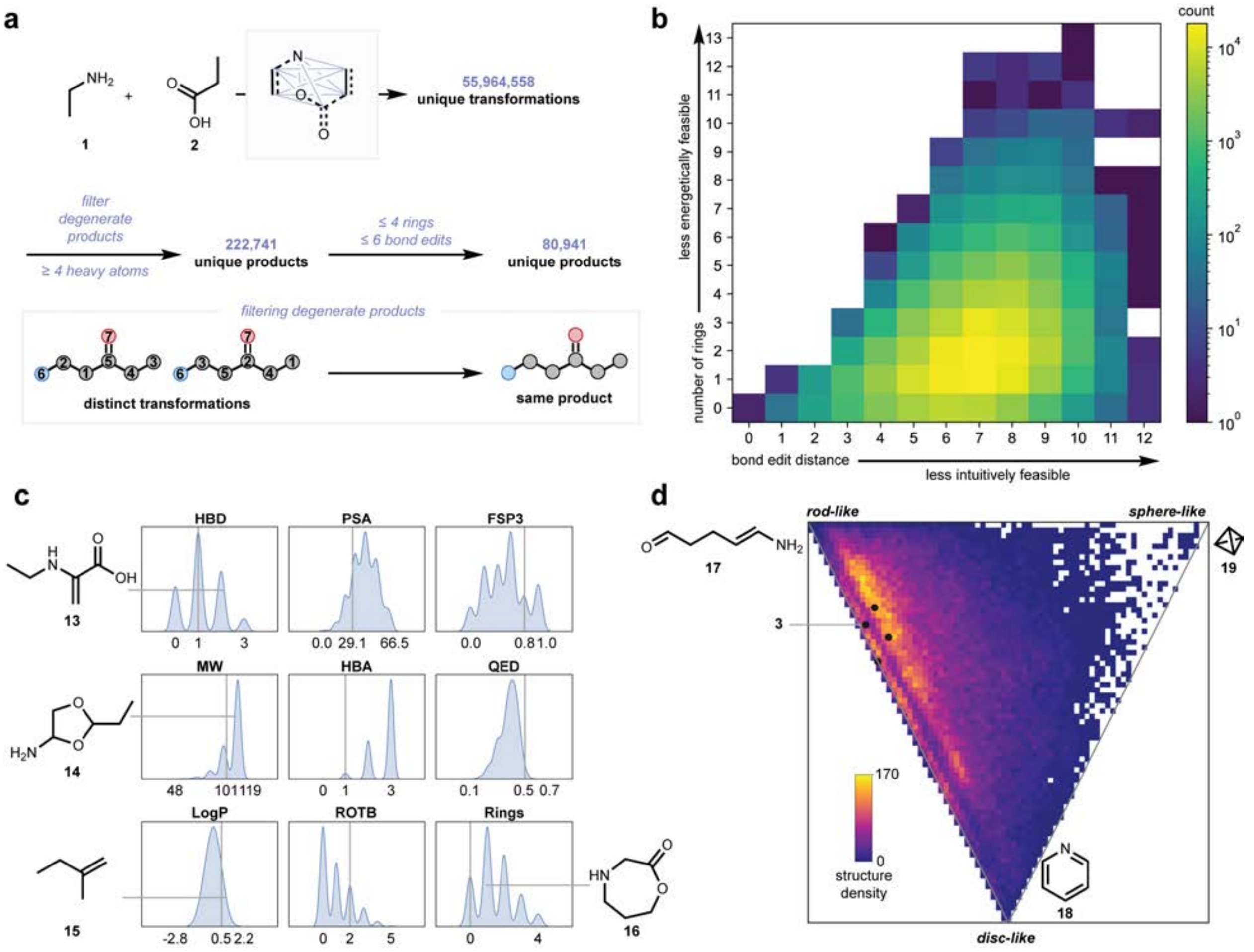
In this project, led by Sam Zhang, the enumerative combinatorics concept introduced in the previous paper was expanded. Instead of only looking at a certain subset of more obviously feasible reactions, all and any hypothetical transformation is considered. In this case, any and multiple bonds can be formed or broken between atoms. This expansion led to an explosion of potential transformations. While many may be currently synthetically infeasible, the extent of potential properties emergent from the transformations reveals the power having control how two substrates merge can have in designing a compound with specific profiles.



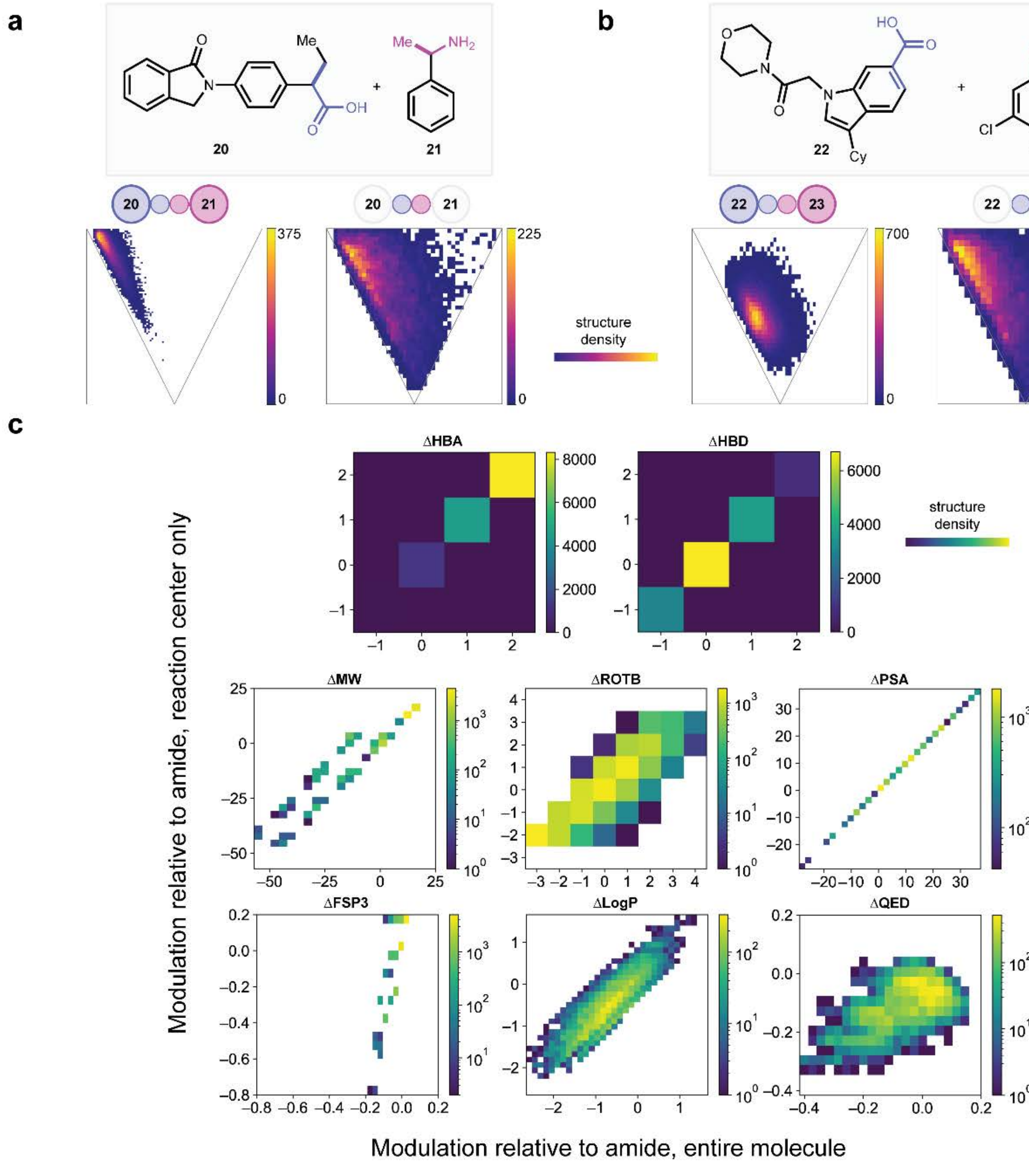
16-Fig. 1 | a. Given an amine **1** and carboxylic acid **2,** the most popular transformation to unite this pair of building blocks produces the amide **3**. **b.** Coupling products arising from a curated subset of chemical transformations charted by our prior work. **c.** Examples of transformations excluded from our prior work.



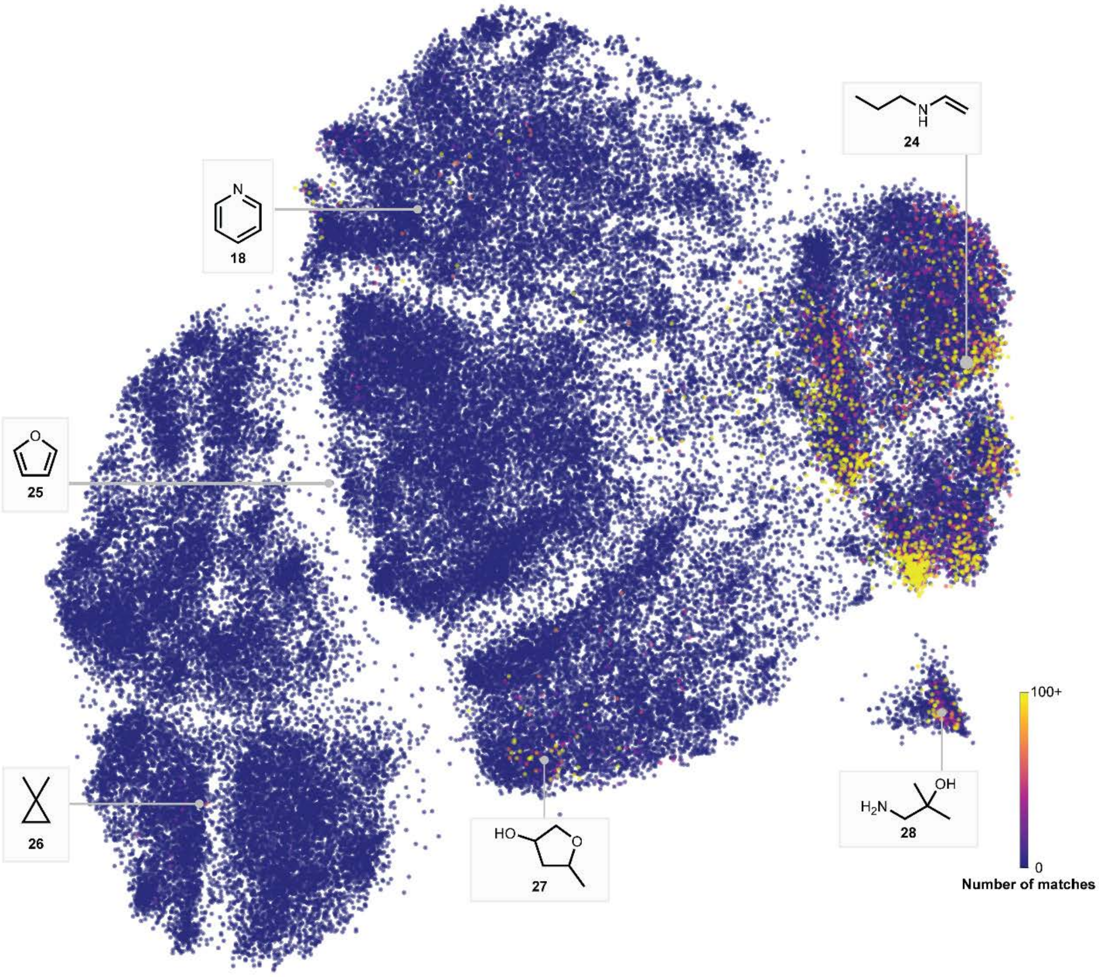
17-Fig. 2 | a. Demonstration of matrix-encoding of reaction from panel **1a**. In the molecular diagrams, white circles represent carbon, blue circles nitrogen, and red circles oxygen. The numbers in the adjacency matrix correspond to the atom indices in the cartoon atoms, while the color represents bond order. **b.** Workflow for exhaustive generation of amine–acid coupling products. **c.** The difference between the product matrix and starting material matrix is the transformation matrix.



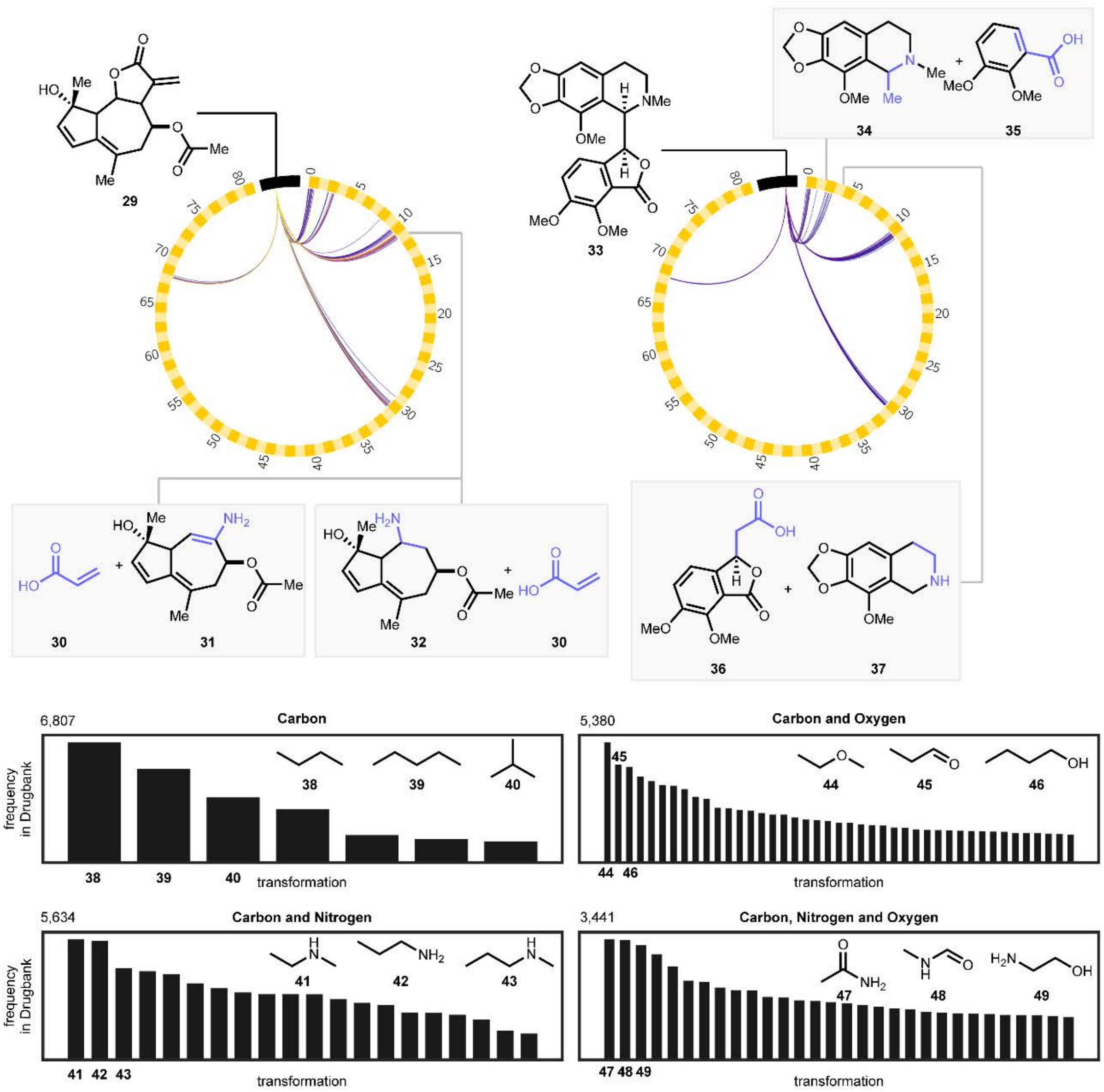
18-Fig. 3 | a. Schematic of enumeration from amine **1** and acid **2** to yield 56 million unique transformation matrices, which are filtered first to 222,740 unique products assuming carbon and oxygen atoms are degenerate, and further to 80,941 unique products after eliminating highly improbable structures. **b.** 2-dimensional histogram showing distribution of ring count and bond edit distance of the initial 222,740 products. **c.** Kernel density estimate (KDE) plots of various physiochemical properties of the expanded amine-acid coupling system, along with selected products. The respective property of the classic amide is shown by the vertical grey line. HBD = hydrogen bond donor, PSA = polar surface area, FSP3 = fraction sp^3-atoms, MW = molecular weight, HBA = hydrogen bond acceptors, QED = quantitative estimate of drug-likedness, LogP = partition coefficient, ROTB = number of rotatable bonds, Rings = number of rings. **d.** Principal Moment of Inertia (PMI) ratio distributions of all products from the expanded enumeration.



19-Fig. 4 | PMI ratio plots showing shape space distribution of reaction enumerated late-stage diversification of two drug-like molecules, **a.** showing an *sp*^{3}–*sp*^{3} coupling, and **b.** showing an *sp*^{2}–*sp*^{2} coupling. For each pair of plots, the left shows the distribution of full coupling products, while the right shows the distribution of only the atoms at the reaction center (*cf*. 18-Fig. 3d). **c.** Joint distribution plots of physicochemical property modulation within the system depicted in **a** (**20** + **21**). See Fig. S2 in Supporting Information for corresponding plots for **22** + **23**. The x-axes show modulation of the entire molecule relative to the amide, while the y-axes show modulation of only the atoms at the reaction center. HBD = hydrogen bond donor, PSA = polar surface area, FSP3 = fraction sp^3-atoms, MW = molecular weight, HBA = hydrogen bond acceptors, QED = quantitative estimate of drug-likedness, LogP = partition coefficient, ROTB = number of rotatable bonds, Rings = number of rings.



20-Fig. 5 | UMAP projection of reaction fingerprints from simple amine–acid starting materials onto 80,941 enumerated products onto DrugBank. Dots are colored by the number of product substructure matches in drugs with the majority of product substructures appearing as unexplored chemical space.



21-Fig. 6 | Chord diagram showcasing connectivity of two drugs to enumerated amine–acid coupling products. Drug targets lie on the black band, and the exhaustive matrix enumeration products lie on the checkered yellow band. The color of the chord represents the substructure's frequency of occurrence, ranging from one to greater than ten (see Figure S7 for further discussion). Two disconnections are shown for athamontanolide (**29**), both into acrylic acid (**30**) and an amine (**31, 32**) with new disconnections generated by the current enumeration workflow. Similarly, noscapine (**33**) can be disconnected to amine **34** and acid **35**, or to alkyl acid **36** and secondary amine **37**. The frequencies of the top 100 most abundant substructure matches in DrugBank are plotted as a histogram, categorized by their elemental makeup. The top three most abundant structures for each group are shown (**38**–**49**).

Reinforcing the Supply Chain of Umifenovir and Other Antiviral Drugs with Retrosynthetic Software

This project was an ad hoc project performed in response to SARS-COV-2/COVID-19 pandemic that emerged in 2020. As the world rushed to develop therapeutics to halt the spread of the pandemic, we posited that any winning drug would experience a strained supply line. In this project, I designed a data structure to encode synthetic routes to over ten potential therapeutics that were being studied for efficacy against the virus. With the help of other lab members, we collected every and all published synthetic route to each of these molecules. Simultaneously, we utilized the retrosynthetic software Synthia to devise routes to these drugs that excluded the same starting materials used in the published routes. In this fashion, new routes would avoid any supply chain constraints imposed by induced demand of the compound. Yingfu Lin and Rose Zhang experimentally validated several of these routes in lab, while I analyzed and developed a software to evaluate the proposed routes against the collected published ones. This report is featured in *Nature Communications*.[54]

Abstract

The global disruption caused by the 2020 coronavirus pandemic stressed the supply chain of many products, including pharmaceuticals. Multiple drug repurposing studies for COVID-19 are now underway. If a winning therapeutic emerges, it is unlikely that the existing inventory of the medicine, or even the chemical raw materials needed to synthesize it, will be available in the quantities required. Here, we utilize retrosynthetic software to arrive at alternate chemical supply chains for the antiviral drug umifenovir, as well as eleven other antiviral and anti-inflammatory drugs. We have experimentally validated four routes to umifenovir and one route to bromhexine. In one route to umifenovir the software invokes conversion of six C–H bonds into C–C bonds or functional groups. The strategy we apply of excluding known starting materials from search results can be used to identify distinct starting materials, for instance to relieve stress on existing supply chains.

Introduction

In 2020, the scourge of coronavirus highlighted the fragility of diverse supply chains, affecting the world's pipeline of hand sanitizer[55], toilet paper[56], and pharmaceutical starting materials[57]. A diverse array of antiviral and anti-inflammatory drugs was investigated in the hopes that an existing medicine could be repurposed for use against COVID-19[58, 59]. The scale of the pandemic[60] would easily stress the pharmaceutical supply chain[61]. For instance, at the time of our study, the only drug with an emergency use authorization for treating COVID-19 was remdesivir (**1**). While its producer, Gilead Sciences, Inc., ramped up production of **1** significantly, there were only 5000 doses of **1** available when the outbreak began[62].

We realized that the availability of alternative starting materials to promising synthetic therapeutics could alleviate pressure on supply chains. We reasoned that modern retrosynthetic software could be used to analyze the diverse chemical agents that were being studied clinically and pre-clinically when we initiated our study[58]. For the sake of preparedness, we chose to consider multiple therapeutic synthetic targets simultaneously, even though many would eventually prove to be irrelevant in the war on COVID-19. We focused our experimental attention on umifenovir (**2**), which had been used against SARS-CoV1, and is effective against SARS-CoV2 in vitro[63]. Automated retrosynthesis has already been used to design contingency plans for the investigational COVID-19 therapeutics hydroxychloroquine and remdesivir (**1**)[64]. This reaction-centric study identified novel sequences computationally, but the routes initiate with known starting materials, or from starting materials that added significantly to route length or reagent cost.

We present herein a starting material-centric retrosynthetic analysis of 12 investigational COVID-19 drugs. We identify distinct raw materials that are of comparable cost, and which feature in routes of comparable length to known routes to the 12 diverse targets. From our perspective, it was not detrimental to intercept established synthetic routes to the selected drugs, as long as the overall route had a competitive step-count and initiated with distinct starting materials of a comparable or better price. The simultaneous design of multistep preparative routes to diverse targets, which circumvent the use of established raw materials, presents a considerable data handling challenge that would be a good test of modern computer-assisted synthesis planning software[26-28, 39, 65-70]. We enlisted the SYNTHIA^{TM} retrosynthesis platform[26-28] to facilitate navigation of requisite parameters including availability, pricing, and novelty of starting materials, route brevity, issues of chemo-, regio-, and stereoselectivity, as well as route visualization, documentation, sharing, and storage. We present herein predicted retrosynthetic routes to twelve diverse COVID-19 therapeutic candidates that initiate, as frequently as possible, with starting materials that are distinct from those used in published or patented syntheses. Few studies of modern retrosynthetic software validate routes experimentally, so it was important for us to realize some of our calculated routes.

Results and Discussion

Crowd-sourced data collection and automated retrosynthesis

Our study commenced with a crowd-sourcing approach wherein each member of our lab collected all published and patented synthetic routes for one of the drugs in 22-Fig. 1. The routes were then encoded via their simplified molecular-input line-entry system (SMILES) strings. From this dataset we built an interactive route visualizer, available for free at http://covidroutes.cernaklab.com,[71] to facilitate review of existing routes. The concatenated list of starting material SMILES from each target was used as an exclusion criterion in each retrosynthetic search. This approach allowed us to rapidly navigate to novel starting materials. Each search result contained 50 route proposals, and the user-defined search heuristic was generally set to minimize starting material cost. A single search heuristic worked for most targets, but occasionally, the preference for minimized cost reagents would result in proposed routes with more reaction steps than desired. In these cases, the search heuristic was modified by relaxing the preference for reagent cost and increasing the software's beam search width. Predicted routes were manually reviewed for step count, synthetic feasibility, and ease of execution of proposed reactions on the multikilogram scale, for instance by biasing towards routes that minimized the use of cryogenic cooling or pyrophoric reagents. The final heuristic used for each target is shown in the Supplementary Information.

Among small molecules being considered (23-Fig. 2), we focused on remdesivir (**1**), umifenovir (**2**), bromhexine (**3**), galidesivir (**4**), ritonavir (**5**), cobicistat (**6**), ribavirin (**7**), camostat (**8**), darunavir (**9**), nelfinavir (**10**), favipiravir (**11**), and baricitinib (**12**). In most cases, the proposed route has the same number, or fewer, steps than the established routes, and initiates from distinct starting materials. Our analysis yielded alternate starting material proposals for **1**–**12**, which can relieve pressure on the fine chemical supply chain. Using galidesivir (**4**) as an example (24-Fig. 3), the software proposed a sequence hinging on a *trans*-hydroiodination of alkyne **13**, an Evans alkylation to form **18**, an Ullman coupling to form **19**, and an enantioselective Heck-coupling to give **22**. The software proposed that the latter reaction mixture could be subjected in situ to hydrochloric acid to remove the Boc-protecting groups in a one-pot operation. Dihydroxylation of **22** would complete the synthesis of **4**. Exemplary starting materials **23**–**25** were excluded from the search based on the appearance of their SMILES strings in published routes to **4**. The algorithm successfully navigated around five established pyrrolopyrimidine starting materials to arrive at **21**, which is cost-competitive with the established nucleobase sources; for instance, the 7-*des*-bromo analog of **21**, 4-chloro-5H-pyrrolo[3,2-d]pyrimidine, is listed at $9.90/g while a 4–OH analog of **21**, 7-bromo-5H-pyrrolo[3,2-d]pyrimidin-4-ol, which is used in the reported synthesis of **4**, is listed at $280/g. On a production scale, all starting material pricing would likely be customized based on competitive bidding, but in any event, the high list price of **21** is comparable to starting material analogs currently described in patents. The proposed use of an Evans auxiliary to produce **18** highlights the software's desire to select robust chemistry, but this step could likely be replaced with a catalytic protocol to avoid auxiliary use for large-scale production if needed. Indeed, the overall route proposes a variety of catalytic operations. For the production of **1**–**12** on large scale, routes could be found that minimized use of cryogenic conditions, pyrophoric reagents, or expensive catalysts, which were the main biases imposed in our manual review of answer sets beyond route length and starting material cost. For **1**, novel starting materials were identified but the route bore high similarity to known routes, mirroring the automated retrosynthesis findings of the Grzybowski lab for this target[64]. While the software was challenged by esoteric functionalities like a chiral phosphorus atom, the predicted route to **4** discussed above, and those to **2** and **3** discussed below, represent typical outputs.

Retrosynthesis and experimental syntheses of Umifenovir

Umifenovir (**2**), is an antiviral drug developed to combat influenza infections whose use against SARS-CoV1 made it an attractive synthetic target for this study. It is believed that **2** inhibits the entry of viruses into human cells, and the antiviral has been used in many clinical trials as an investigational COVID-19 therapy[63]. Although **2** shows promising in vitro activity against the novel coronavirus, recent clinical results suggest limited efficacy for **2** against COVID-19 in humans[72]. Using the search criteria described above, we arrived at a series of routes to **2** based on the oxidative cyclization of aniline with a β-ketoester[73]. Since it is a commodity chemical, ethyl acetoacetate **27** was not included in the exclusion criteria of our heuristic and appears as a starting material here, although it has been used previously in the synthesis of **2**. Starting materials **28**–**30**, among several others (see Supplementary Information), were excluded. A key theme that separated the predicted routes from the established indole-forming routes, and enabled the use of distinct starting materials, was the incorporation of a Baeyer–Villiger oxidation to utilize an acetyl group as a surrogate to the requisite hydroxyl group at C1. We found this proposal of a Baeyer–Villiger oxidation to be a surprising solution. Among other proposals that were non-obvious to us was the suggested C–H oxidation of an ethyl group where the use of C–H functionalization logic[74-77] reduces the cost of the starting materials. The software proposed an inefficient three-step sequence to *N*-methylate the indole, whereas published syntheses of **2** reported *N*-methylation directly from the indole with methyl iodide and sodium hydride. We opted to use this one-step precedent instead of the software's three-step proposal. In another search, a proposed sequence to **2** was initiated with a pre-installed halogen coupling handle, instead of a C–H bond, to enable a related indole formation, but instead invoked a Bamberger rearrangement to functionalize the C–H bond. As described below, these four routes were reduced to experimental practice with only minor modifications to reaction conditions and sequences proposed by the software.

To experimentally validate routes to **2**, we first investigated the proposed indole formation from 1-(4-aminophenyl)ethan-1-one (**26**, \$1.15/g) and ethyl acetoacetate (**27**, \$0.03/g) using oxidative reaction conditions (25-Fig. 4 route A). Pretreatment of **26** and **27** with 1 mol% indium(III) bromide, to form **31**, was followed by oxidative cyclization to form **32**. While the published reaction conditions for the suggested reaction[73] did provide the desired indole **32**, the yield was only 20%. Using magnesium sulfate to promote the formation of **31** improved the yield of **32** to 47%. As described below, other implementations of this reaction gave much higher yields. *N*-Methylation occurred smoothly to produce **33** in 99% yield. An issue was encountered in the experimental realization of the Baeyer–Villiger oxidation using *m*CPBA in that a mixture of oxidation products was obtained. Unstable products we believe to be from oxidation of the indole's double bond accounted for the bulk of the reacted material, and only traces of **34** were isolated. While the formation of **34** was accurately predicted, the subtle interplay of electronics that govern the preference for the desired Baeyer-Villiger oxidation over the undesired Prilezhaev oxidation could not be teased out by the software, and the best modification of reaction conditions we found yielded small amounts of **34** as a mixture with undesired oxidation products. A literature search on related indoles revealed that the α-chloroketone **35** should be a viable substrate for the Baeyer–Villiger[78], with the chlorogroup acting as a directing group to favor oxidation of the ketone. We thus modified the route and, indeed, chlorination of **33** led to **35**, which underwent selective Baeyer–Villiger oxidation to produce **36**. Subsequent bromination produced **37**, which underwent thioetherification with **38** and in situ saponification to produce **39**. Here, the route intercepts known syntheses of **2** via alkylation with **40**[79]. All intermediates predicted by the software were observed, but a modification to incorporate a chlorine directing group was necessary to achieve usable levels of selectivity in the formation of **36**. This change led us to demonstrate the bromination of **36** to produce **37**, instead of brominating **34**, yet the selective bromination of **34** en route to **39** is a known reaction[80].

The output of a SYNTHIA™ search is a ranked list of route proposals. Several other computed routes to **2** were also experimentally vetted. One route, based on a variation of the same indole-formation and Baeyer–Villiger sequence described above, proposed a benzylic C–H oxidation of indole **43** (25-Fig. 4, route B). The indole-synthesis was more productive with **41** than with **26**, yielding **42** in 79% yield. Methylation gave **43** in 92% yield. SYNTHIA™ predicts reaction sequences. Corresponding reaction conditions are recommended based on what was reported in the source literature. While these recommendations work well for a majority of substrates, exact recipes for specific substrates may require user direction. Software-recommended conditions of Oxone® with potassium bromide[81] for the C–H oxidation of **43** were unsuccessful in our hands. An experimental survey of oxidants revealed the recently disclosed Baran-Roček oxidation[82, 83] could selectively oxidize C14 in 62% yield, thus intercepting the previous route to **2**.

While the generation of chromium waste is only viable on small scale, this result validates the proposed C–H functionalization hypothesis.

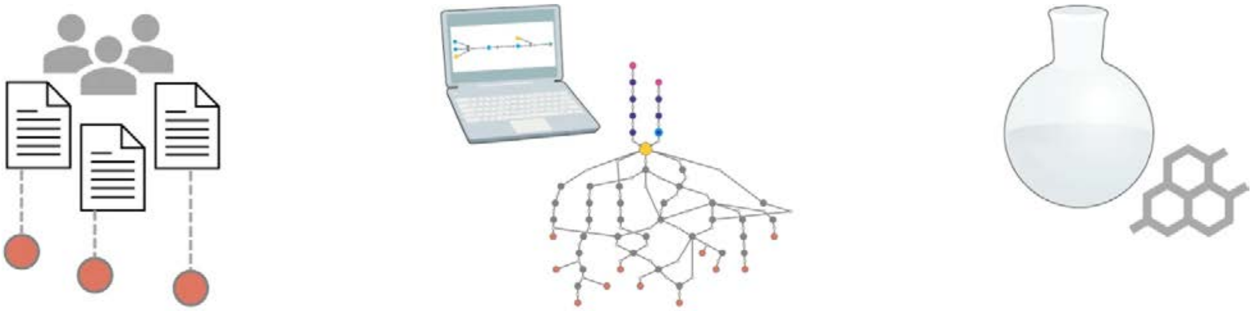
The direct installation of the chloromethyl ketone via a Friedel–Crafts acylation would provide a concise and alternative route to **2**. Indeed, the software had proposed a route that used a Friedel–Crafts acylation (25-Fig. 4, route C). This route was intriguing in that it initiated the synthesis from **45**, an exceptionally cheap starting material. While the software proposed a Friedel–Crafts acylation with acetyl chloride, we modified the route to use instead chloroacetyl chloride (**48**, \$0.13/g), thus installing the chloride directing group in a single step. Experimentally, oxidative indole-coupling to form **46**, followed by methylation to form **47**, occurred smoothly. Friedel–Crafts acylation of **47** with **48** under influence of aluminum(III) chloride gave **35** and intercepted the other routes. The 2:1 regioselectivity of the acylation would require optimization for production on production scale. Aside from this reaction, the regioselectivity for desired isomers was excellent for all other C–H functionalization events. We expect the frequent suggestion by the software to convert C–H bonds into other functionalities is the result of the preference for low-cost starting materials in our heuristic, with C–H bonds in many instances being cheaper than other functionalities. The Friedel–Crafts acylation route described replaces six C–H bonds with new functionalities over seven reactions to convert **27** and **45** into **2**.

We next employed a different tactic. Most routes to **2** hinge on a Nenitzescu indole coupling[79] between 1,4-benzoquinone and a β-aminocrotonic ester[71]. Indeed, the Nenitzescu reaction using known starting materials featured as a proposal in our query results when default search criteria were used, so the keyword “Nenitzescu” was used as an exclusion criterion. This heuristic did not employ a SMILES exclusion criterion, so starting material **52** was employed even though this chemical has been used in a prior synthesis of **2**. The results of this search led to yet another proposal to use a C–H bond as a feedstock, via a Bamberger rearrangement to install the C1 hydroxyl (25-Fig. 4, route D). SYNTHIA^{TM} proposed the use of 5-bromo-2-chloro-1-nitrobenzene as a starting material. In our hands, the requisite indole coupling on the chloride gave only traces of **39**, and we ultimately modified the starting material to use 2,5-dibromo-1-nitrobenzene (**49**) instead. This modification allowed the indole coupling to proceed, as discussed below, with the added benefit that **49** is cheaper than the corresponding chloride. In practice, **49** was reduced to the hydroxylamine, and treated with aqueous trifluoroacetic acid to affect the Bamberger rearrangement yielding **50**, which was methylated to arrive at **51**. Copper-catalyzed coupling to **53**, itself obtained through the union of **52** and **38**, produced **39** in 66% yield when **54** was used as a ligand. These conditions were the result of a rapid optimization campaign using high-throughput experimentation (see Supplementary Information). Subsequent alkylation of **39** with **40** produced **2**. With this latter route, convergency is maximized, so the longest-linear sequence is just five steps.

One-step synthesis of bromhexine

Finally, we looked at **3** (26-Fig. 5), a transmembrane protease, serine 2 (TMPRSS2) inhibitor that was being investigated in five clinical trials for COVID-19. A SYNTHIA^{TM} search provided new reaction sequences of comparable length to known routes, identifying **55** as a novel starting material[71] by navigating around known starting materials **56**–**59** and others. The predicted route invoked a C–H oxidation of the benzylic methyl group, presumably to arrive at cheaper starting materials, which readied **60** for reductive amination with **61**. The proposed route completed the synthesis of **3** by *N*-methylation of **62** with **63**. We considered instead that **3** could be synthesized from 2,4,6-tribromoaniline (**64**, \$0.51/g), which is used in the textile industry and readily available, with *N*,*N*-dimethylcyclohexylamine (**65**, \$0.10/g), a commodity chemical used in oil refining, via the direct C–H functionalization recently reported by Shirakawa[84]. While this manually designed route does not serve as a test of the software's capability per se, our motivation here was largely to do what we could as synthetic chemists to support the production of a potentially beneficial medicine during a pandemic. The key reaction was added to the SYNTHIA^{TM} database so it would appear as a general solution to subsequent searches, and indeed this route came up as a top hit in a subsequent search for **3**. To experimentally realize this one-step route, we found that **64** could be heated in excess **65** in the presence of *tert*-butylperoxide[84] to produce **3** in 41% yield. Further optimization of reaction conditions—to improve yield, ease of purification of **3**, and address the hazard of using peroxide on large scale—would be needed for commercial production. Nonetheless, this strategic disconnection reduces **3**, in a single step, to starting materials that are considerably cheaper than those in commercial use.

In this work, we show disconnection of drugs into affordable reagents for 12 drugs through the merger of crowd-sourcing and retrosynthetic software. Four predicted routes to **2** and one route to **3**, manually designed but added to the software's database for future use, were experimentally validated. Navigating the combinatorial explosion of routes towards twelve distinct synthetic targets to arrive at distinct and affordable starting materials was a data handling challenge that could only be navigated with automated retrosynthesis. Our work was performed over nine weeks in Spring 2020 against the backdrop of a developing pandemic. While full process development would require a longer timeline—for instance, reagents such as peroxides would likely be replaced for production on commercial scale—our results show that automated retrosynthetic predictions can rapidly highlight alternative starting material supply chains to pharmaceuticals.

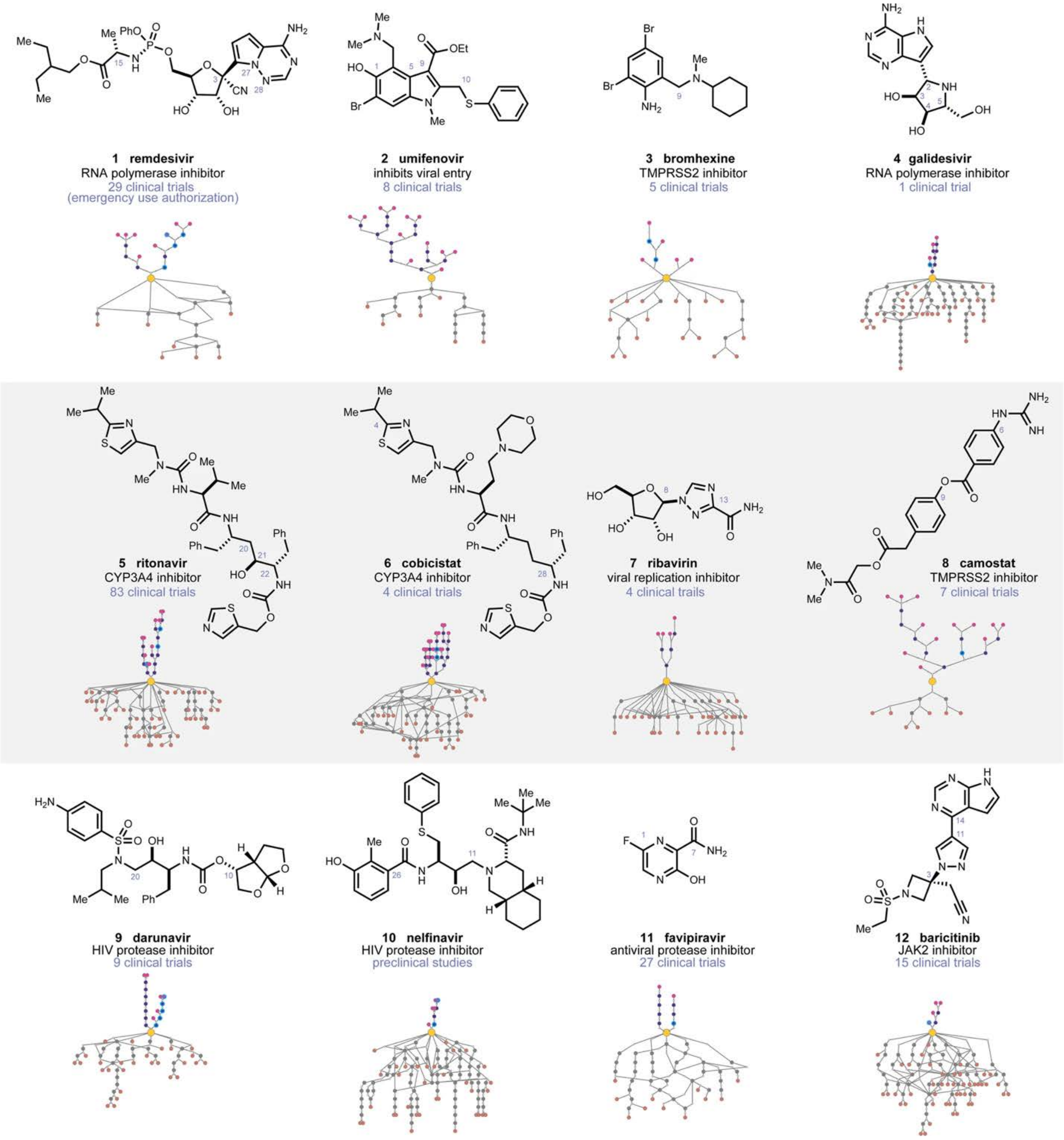


step 1.
Crowd-sourced literature review identifies known starting materials

step 2.
Retrosynthetic search specifically excludes known starting materials

step 3.
Experimental validation of predicted synthetic route

22-Fig. 1 | Workflow to identify alternate starting material supply chains. 1. Published starting materials are encoded as SMILES strings, which are 2. excluded from automated retrosynthetic searching. 3. Select routes are validated in the wet lab.



1 remdesivir
RNA polymerase inhibitor
29 clinical trials
(emergency use authorization)

2 umifenovir
inhibits viral entry
8 clinical trials

3 bromhexine
TMPRSS2 inhibitor
5 clinical trials

4 galidesivir
RNA polymerase inhibitor
1 clinical trial

5 ritonavir
CYP3A4 inhibitor
83 clinical trials

6 cobicistat
CYP3A4 inhibitor
4 clinical trials

7 ribavirin
viral replication inhibitor
4 clinical trails

8 camostat
TMPRSS2 inhibitor
7 clinical trials

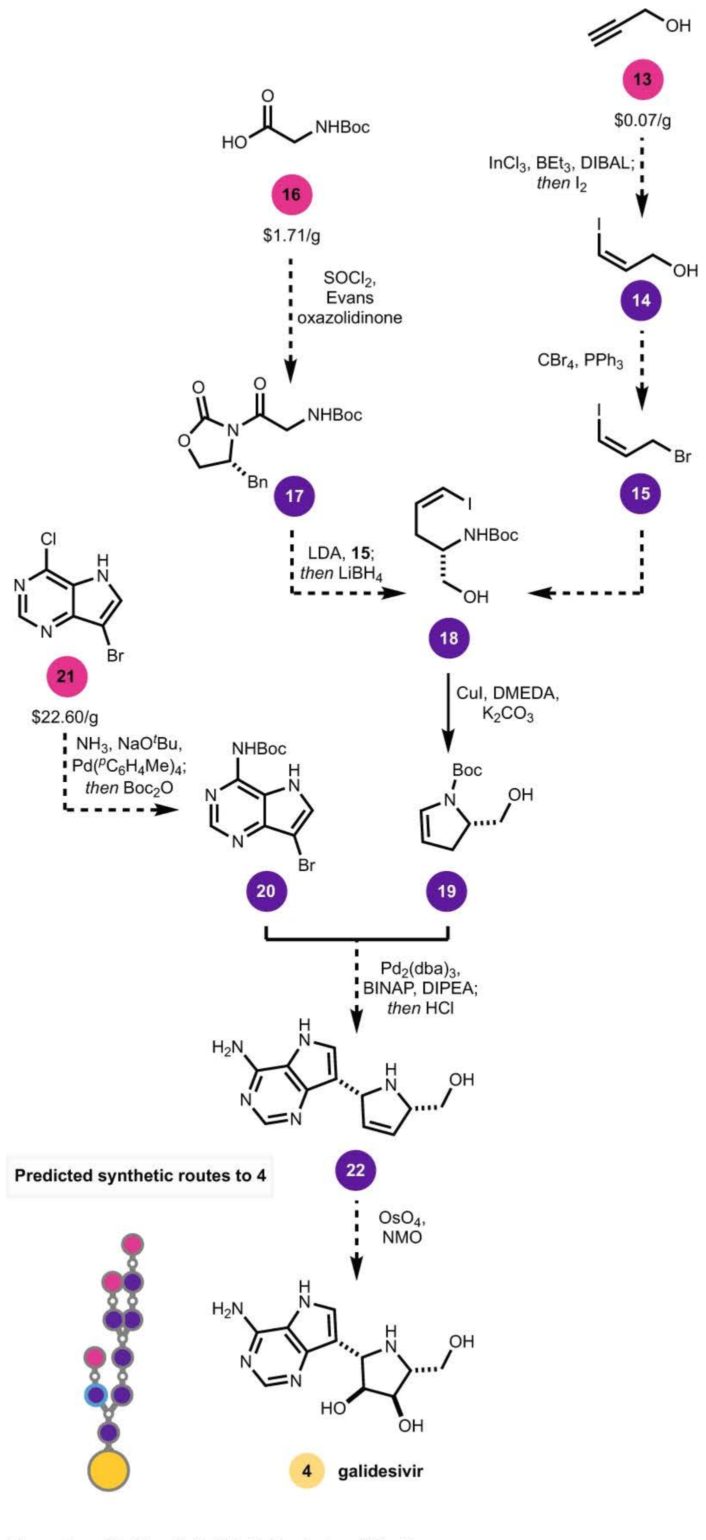
9 darunavir
HIV protease inhibitor
9 clinical trials

10 nelfinavir
HIV protease inhibitor
preclinical studies

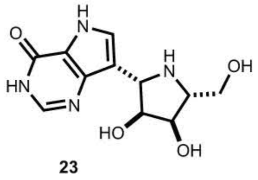
11 favipiravir
antiviral protease inhibitor
27 clinical trials

12 baricitinib
JAK2 inhibitor
15 clinical trials

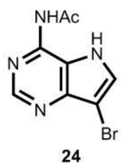
23-Fig. 2 | Therapeutics being evaluated for COVID-19 and their retrosynthetic analysis networks. The number of clinical trials is based on search results for all listed trials, completed, active, or planned, found on www.clinicaltrials.gov (accessed July 7th, 2020). For the route networks, the yellow dot at the center is the target molecule. Routes in grey and orange below the target are published, and the routes in purple and pink above the target are routes predicted. Intersection nodes in literature routes reflect common intermediates or starting materials. A version of this diagram is available as an interactive route visualizer.


13
$0.07/g
$InCl_3$, BEt_3, DIBAL; *then* I_2
14
CBr_4, PPh_3
15
16
$1.71/g
$SOCl_2$, Evans oxazolidinone
17
LDA, **15**; *then* $LiBH_4$
18
CuI, DMEDA, K_2CO_3
19
21
$22.60/g
NH_3, NaO^tBu, $Pd(^pC_6H_4Me)_4$; *then* Boc_2O
20
$Pd_2(dba)_3$, BINAP, DIPEA; *then* HCl
22
OsO_4, NMO
Predicted synthetic routes to 4
4 **galidesivir**

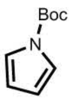
Examplary Starting Materials in Exclusion Criteria



23

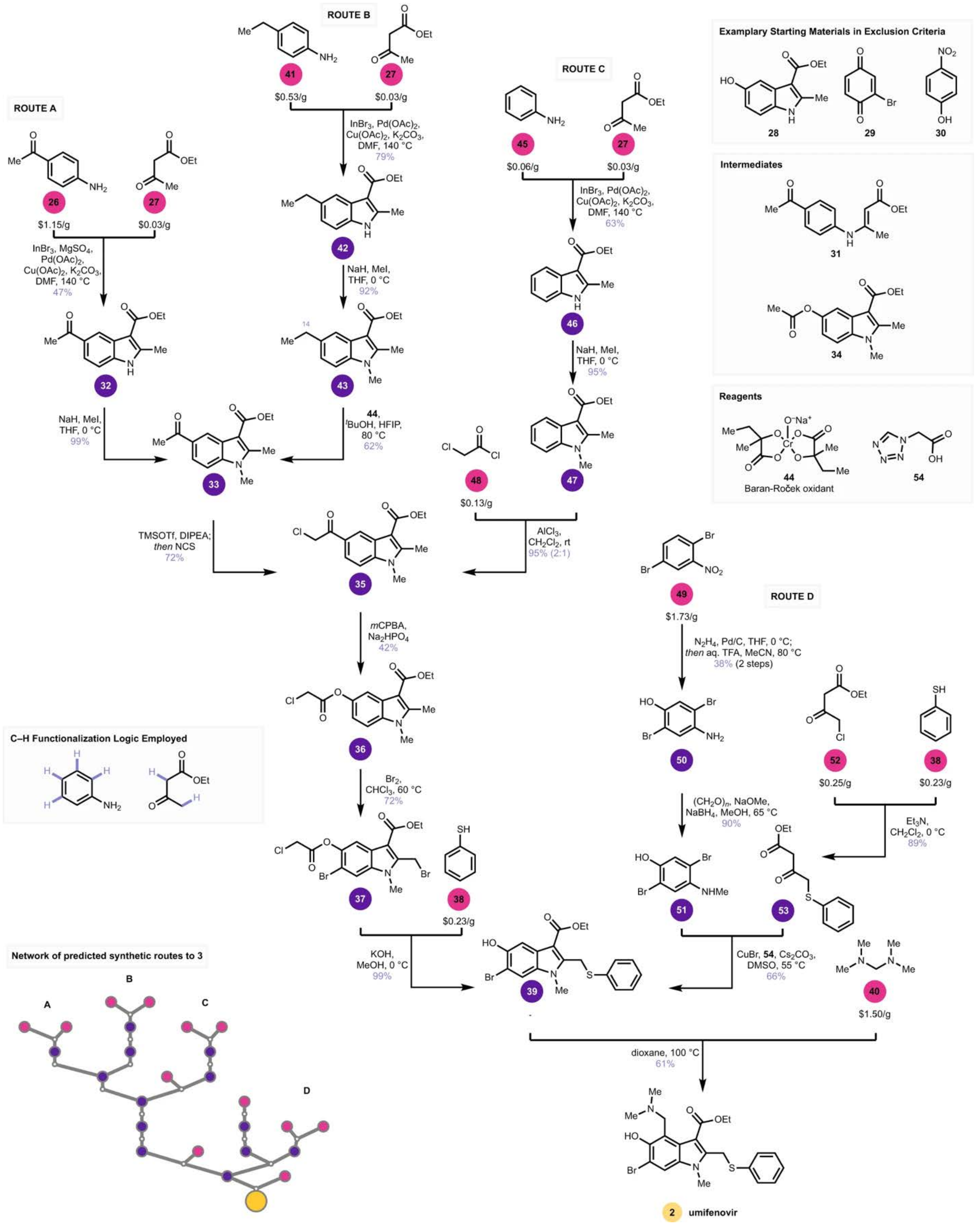


24

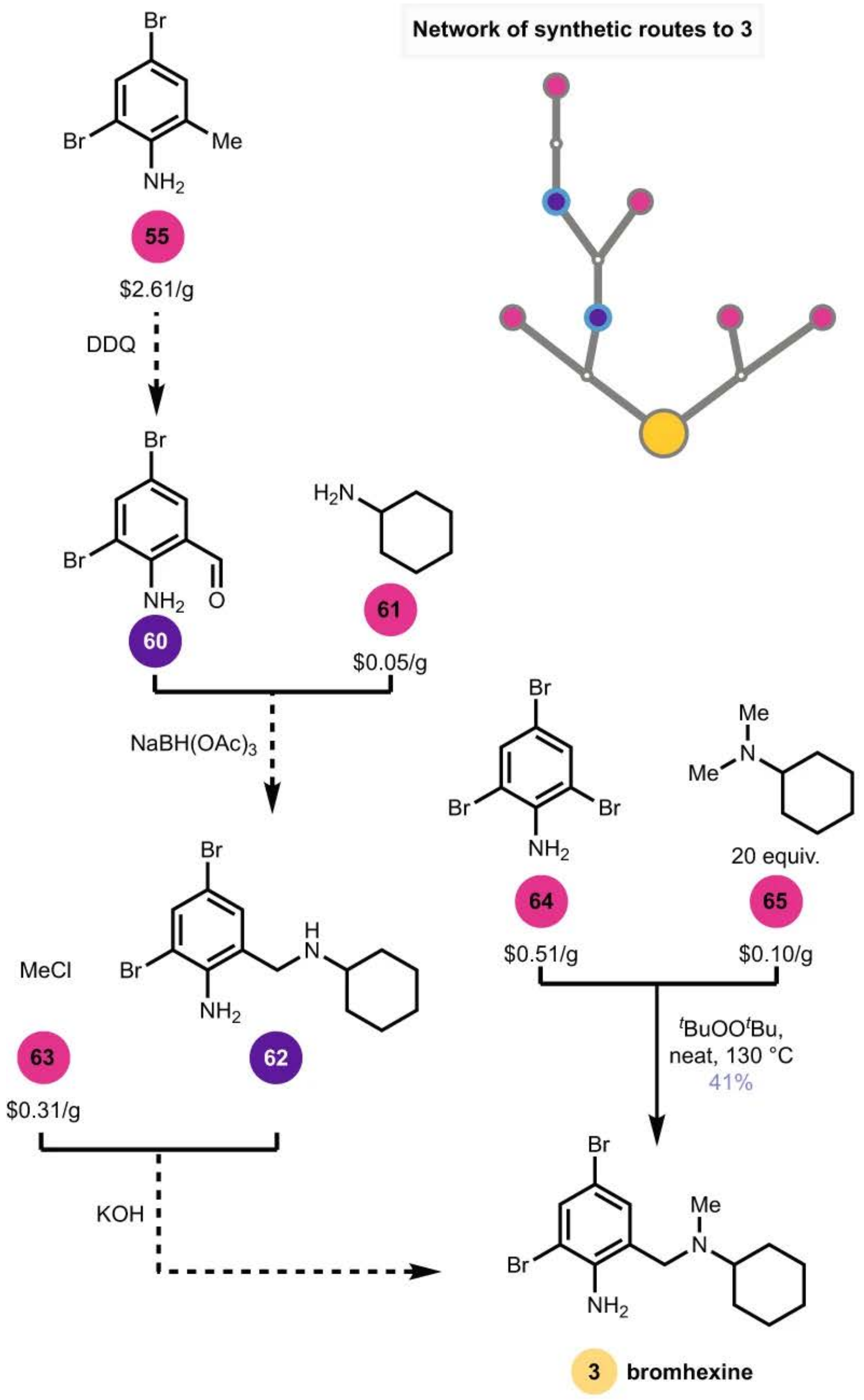


25

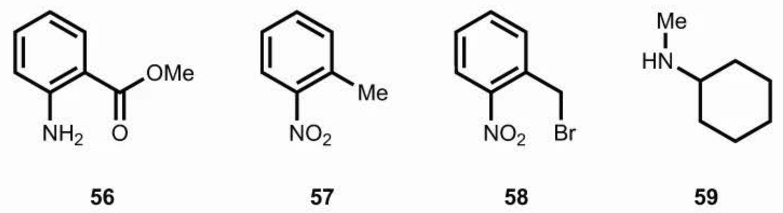
24-Fig. 3 | Predicted synthetic sequence to 4. Reaction conditions are proposals from the SYNTHIA™ software.



25-Fig. 4 | Experimental demonstration of four routes to 2. Each route terminates in inexpensive starting materials. **26**, **41**, **45**, **48**, and **49** have not been used in a published synthesis of **2**. The conversion of **33** into **35**, and the use of starting material **49**, instead of 5-bromo-2-chloro-1-nitrobenzene, were manual modifications to the software's prediction. See main text for details.

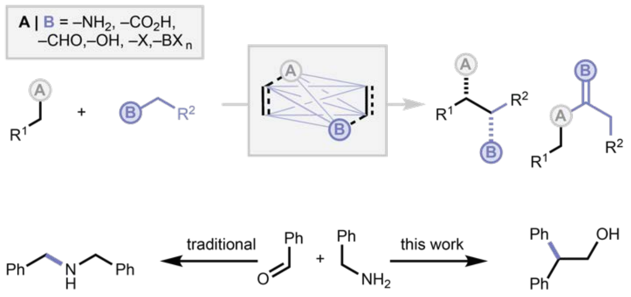


Examplary Starting Materials in Exclusion Criteria



26-Fig. 5 | Predicted and realized routes to 3. Dashed arrows are predicted, while the solid arrow is manually designed and experimentally validated, with the subsequent addition of the key reaction to the SYNTHIA™ database.

Analyzing Conceivable Reactions from Commercially Available Chemical Building Blocks



27-Graphical Abstract | Enumerative combinatorics is taken beyond amines and acids to all common building block types.

This project contains information regarding distributions and analyses of common building block classes in the Sigma-Aldrich catalog. It is shown that enumerated transformations are plentiful between various building block types that are perhaps less often used compared to common building blocks such as boronates and halides. Using the extent of transformations found in drug compounds and the abundance of building block classes in commercial catalogs, a metric is developed to gauge the most valuable transformations. Several of these unconventional reactions between building blocks are experimentally shown.

Abstract

The navigation of chemical space is highly dependent on the availability of chemical building blocks. A common drug discovery tactic is to attach commercially available building blocks to a core moiety using robust chemical transformations. This systematic matching of commonly available building blocks such as those bearing amines, acids, alcohols, aldehydes, halides and boronates to popular reactions such as the amide coupling, Buchwald-Hartwig coupling, Fisher esterification, or Suzuki coupling has formed the basis for modern medicinal chemistry. Here, we explore the interplay of commercially available building blocks with existing and hypothetical reaction methods to evaluate purchasable substrates classes through chemoinformatic analyses in Python. Several case studies are presented to showcase examples of high value transformation space.

Main

Modern pharmaceutical synthesis favors a subset of robust reactions including amide coupling, reductive amination, Suzuki coupling, Buchwald-Hartwig coupling and related aryl-heteroatom couplings (28-Fig. 1A).[30, 85-88] These well-known reactions perform robustly across many substrate classes and also enjoy a diversity of commercially available amine, acid, halide, boronate and aldehyde feedstocks. Often likened to the plastic blocks of the popular Lego® toy, the availability of chemical building blocks frequently dictates the physicochemical property space that can be reached in a drug discovery program.[89-92] Commercially available building blocks provide access to a broad swath of physicochemical property space, and are typically united with pharmacophoric cores using popular transformations such as the amide coupling.[93] The merger of building blocks with predictable reactions that unite them forms the basis of modern Make-On-Demand virtual libraries, which have been demonstrated to offer high quality drug leads.[94, 95] Building on these trends, large virtual libraries developed from building block libraries via robust chemical coupling reactions have become an important source of lead discovery through ultralarge-scale virtual screening.[96-99]

We have recently shown that chemical reactions themselves play a major role in impacting physicochemical properties,[1, 100] and have developed a series of reaction methods based on diverse amine–acid couplings.[101-104] Here, we expand the analysis to include other popular building blocks such as alcohols, halides, boronates, and aldehydes (28-Fig 1B) and specifically link the reactions to commercially available building blocks. Several examples of high impact transformations between abundant building blocks as identified in this report have been published recently.[105-107]

Rational modulation of physicochemical properties is critical to the macroscopic phenotype of a small molecule, such as its solubility, metabolic stability, or affinity to a target protein. Traditionally, such property modulation has relied on the availability of diverse building blocks. In a classic medicinal chemistry optimization campaign, it may be necessary to evaluate several hundred building blocks, united with a single pharmacophoric core using a single reaction (Fig. S1). Our lab has been exploring a complementary approach, wherein we consider coupling a single building block to a single pharmacophoric core using hundreds of conceivable reactions, even if the requisite reaction method has not yet been developed.[1, 100] The majority of commercially available building blocks fall around 12 heavy atoms, whereas most reactions change 2–6 heavy atoms at a time (28-Fig. 1C). Thus, in the status quo, the tuning of property diversity by interchanging building blocks requires a coarse-tuning modulation of around 12 heavy atoms at a time. Contrarily, our reaction-centric approach to property modulation changes only a few atoms at a time but achieves comparable physiochemical property diversity, thus offering a fine-grain modulation of properties.[17, 18]

While the coupling reactions we propose may not yet exist, we envision acceleration of reaction method development through the advent of high-throughput experimentation,[36, 108-113] machine learning,[114] and novel chemical reactivity. Emerging chemical technologies[115, 116] such as C–H functionalization,[117-119] photoredox operations,[120] transition metal catalysis, and electrochemistry[121] to name a few, will further expand the breadth of imaginable coupling reactions.

While we propose that the coupling of available building blocks should be accomplished via many diverse chemical transformations, only a handful of these chemical transformations exist. It currently requires significant resources to invent recipes for new chemical reaction methods, but the process is becoming increasingly automated.[114] The analysis described in 29-Fig. 2 provides one possible formula to triage specific high-impact coupling reactions for invention. Since the objective is to identify novel reaction methods of high impact, our analysis is completely agnostic of reaction mechanism or feasibility. As per our earlier analysis, we only require that products observe the octet rule.

One key parameter we consider is the prevalence of reaction product substructures in drugs. This is shown in chord diagrams in 29-Fig. 2A and 29-Fig. 2B where 15 antivirals under trial for COVID-19 treatment including bromhexine, remdesivir, and umifenovir are linked to conceivable reactions in boronate–bromide coupling space and amine–acid coupling space.

Each line in 29-Figs 2A-2C represents the observation of the reaction product substructure in a pharmaceutical, and the lines are colored by the number of times the substructure appears in each pharmaceutical. It is easy to see that the amine–acid space is more densely populated than the boronate–bromide space, which is rationalized since most of the “white space” in the boronate–bromide chord diagram represents product substructures containing a boronate or a bromide, which occur rarely in drugs. In contrast, products where both the boronate and bromide do not appear, as in the products of Suzuki coupling and related C–C coupling methods, have high prevalence in drugs. Expanding the analysis to other common functional groups prevalent as reactive handles in building blocks – acids, aldehydes, alcohols, amines, halides, and boronates – shows that the amine–acid coupling space yields the most product substructures (3,050) that appear in the targeted pharmaceuticals (29-Fig. 2C and 29-Fig. 2D). Meanwhile, within the boronate–boronate couplings, there are only 71 substructure matches across the 15 pharmaceuticals, demonstrating that there are fewer reactions of high impact that preserve one or both boronate groups, and reaction discovery should be focused towards coupling reactions that form C–C bonds while losing both boronates. The richness of opportunity to impact drug discovery makes the amine–acid coupling space an ideal venue for reaction method development.

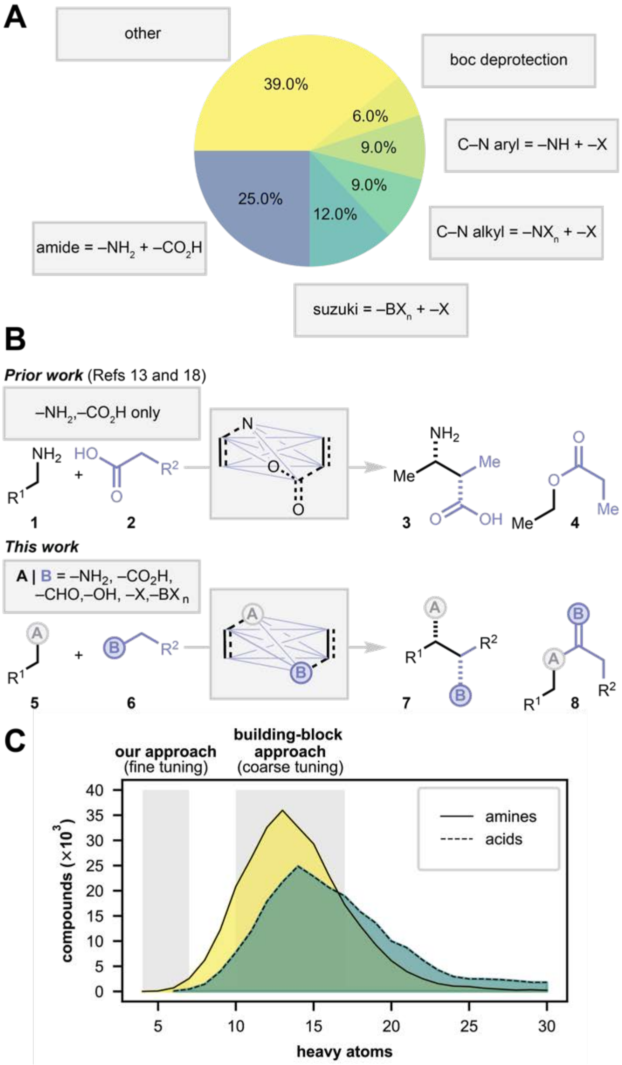
By considering both the availability of commercial building blocks and the prevalence of the transformations born from their respective enumeration, we can gauge potential reaction classes for pareto optimality (30-Fig. 3). Through our analysis, it is revealed that reactions utilizing carboxylic acids, amines, and alcohols push the pareto frontier, and are thus highly valuable targets for research. In our studies, we have discovered several non-traditional couplings between these Pareto-optimal building blocks. For instance, we have discovered an unusual rearrangement between aldehyde **9** and activated benzylamine **10** to form **12**, as opposed to the traditional reductive amination to form **11**.[122] In studies on the carbon–nitrogen coupling between bromide **13** and piperidine **14** to form Buchwald-Hartwig product **15**, we realized a significant impact of base leading to rearranged product **16**, similar to that recently reported by Bandar et al.[123] Finally, as an example of a non-expected C–H functionalization to form a carbon-nitrogen bond, we showcase our previously reported one-pot synthesis of bromhexine (**19**) when treating compound **18** with excess tribromoaniline **17**.

Two examples that are under exploration in our lab are the amine–acid deaminative–decarboxylative C–C coupling of alkyl amines and alkyl and aryl acids,[104] and the amine–acid esterification of aryl amines and aryl or alkyl carboxylic acids.[102] For each transformation, we conducted a head-to-head comparison with the boronate-bromide Suzuki coupling. Building blocks that had the relevant functional groups were extracted from the Sigma-Aldrich catalog, and their chemical space overlap plotted as Venn diagrams (31-Fig. 4). In 31-Fig. 4A, the transformation involved forming a *sp*^{3}–*sp*^{3} C–C bond at the α position of the building block while losing both amine and carboxylic acid functional groups, for instance creating **22** from **20** and **21**, whereas 31-Fig. 4B considers a *sp*^{3}–*sp*^{2} C–C bond formation with an aryl carboxylic acid, such as forming **25** from activated amine **23** and acid **24**. In each instance, the building blocks were coupled with our developed C–C coupling methods, in tandem with the Suzuki coupling. Morgan fingerprints were subsequently computed for all products, and they were clustered via t-distributed stochastic neighbor embedding (t-SNE).

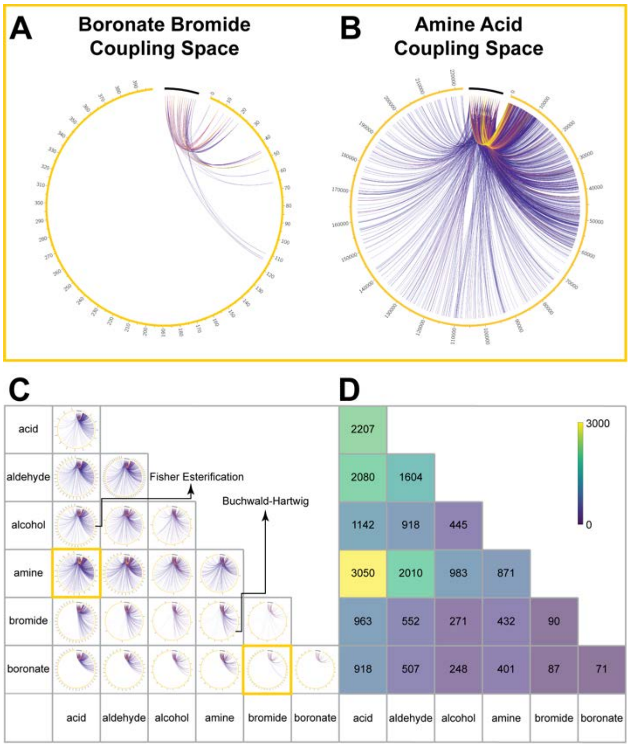
Similarly, amine–acid esterification chemistry was analyzed, where diazonium salts like **26** undergo deaminative esterification with acids such as **27** to form compounds like **28**. (Fig. S2A).[101] Aryl amines that are commercially available building blocks but non-existent as alcohols provide an alternative to the Fisher esterification. (Fig. S2B) Both *para-* and *ortho*-substituted anilines provide access to chemical space unreachable by their phenolic counterparts. (Fig. S2C) An inspection of the physicochemical property profile of anilines and phenols showed that purchasable anilines are more druglike (Fig. S2D), as evidenced by aniline distributions scoring higher on druglikedness scores compared to the phenol distributions. A similar analysis was performed on another amine–acid esterification between alkyl amines and alkyl acids,[103] and it was found that alkyl amines provide access to more druglike molecules compared to alcohols. (Fig. S3) Finally, head-to-head analyses of the physicochemical properties of the building blocks within the classes used in the esterification chemistries reveal the substrate versatility reactivities need to achieve desired property distributions in a product molecule. (Fig. S4)

Conclusion

The common drug discovery tactic of unifying diverse building blocks with select robust reactions has resulted in large commercial libraries containing popular functional groups. We have demonstrated that by leveraging novel reactivities between common functional groups, the accessible chemical space can be vastly expanded. Many reactions fit the criteria of high prevalence of product substructure in pharmaceuticals and broad commercial availability of building blocks. Mining transformation and building block space will continue to guide the efforts of reaction discovery. In conjunction with reaction informatics, the most valuable transformations can be identified and invented to access relevant chemical space with higher efficiency. Furthermore, modern efforts in physical organic chemistry are elucidating the mechanistic understanding of cross-coupling reactions. Mapping the extent of cross-coupling reactivity is an invaluable step in applying current mechanistic understanding towards the discovery of high value transformations.

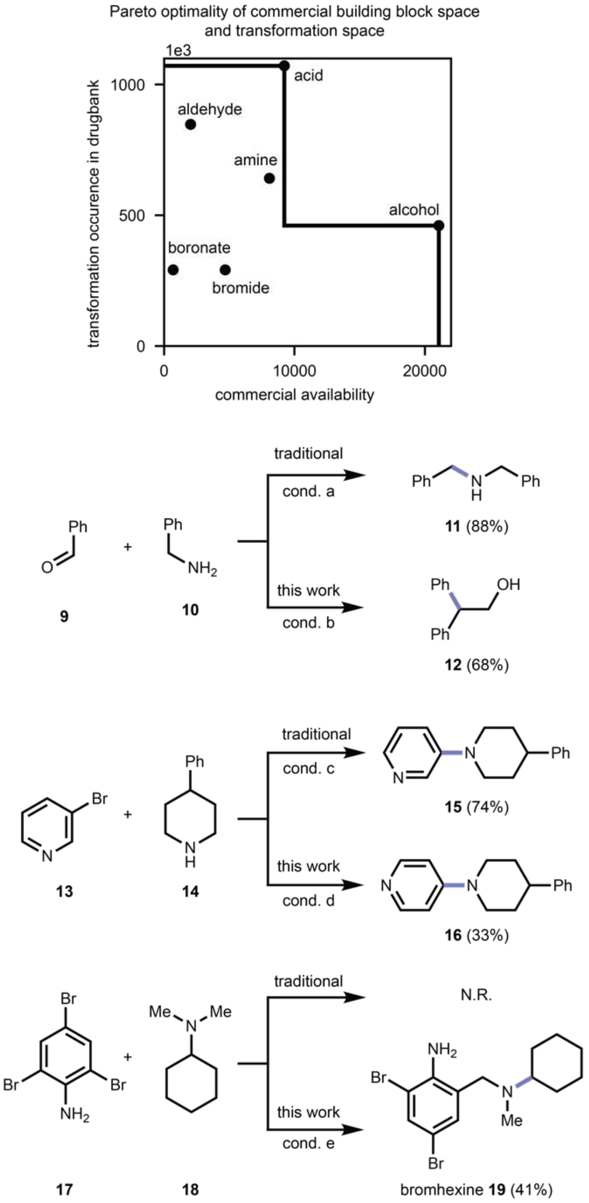


28-Fig. 1 | Analysis of building blocks from the MilliporeSigma catalog. (A) Popular reactions in medicinal chemistry and popular building block classes. (B) Comparison of scope between our prior work and current work. (-X = halides, -BX_n = boronates) (C) Distribution of heavy atoms in commercially available amines and acids.

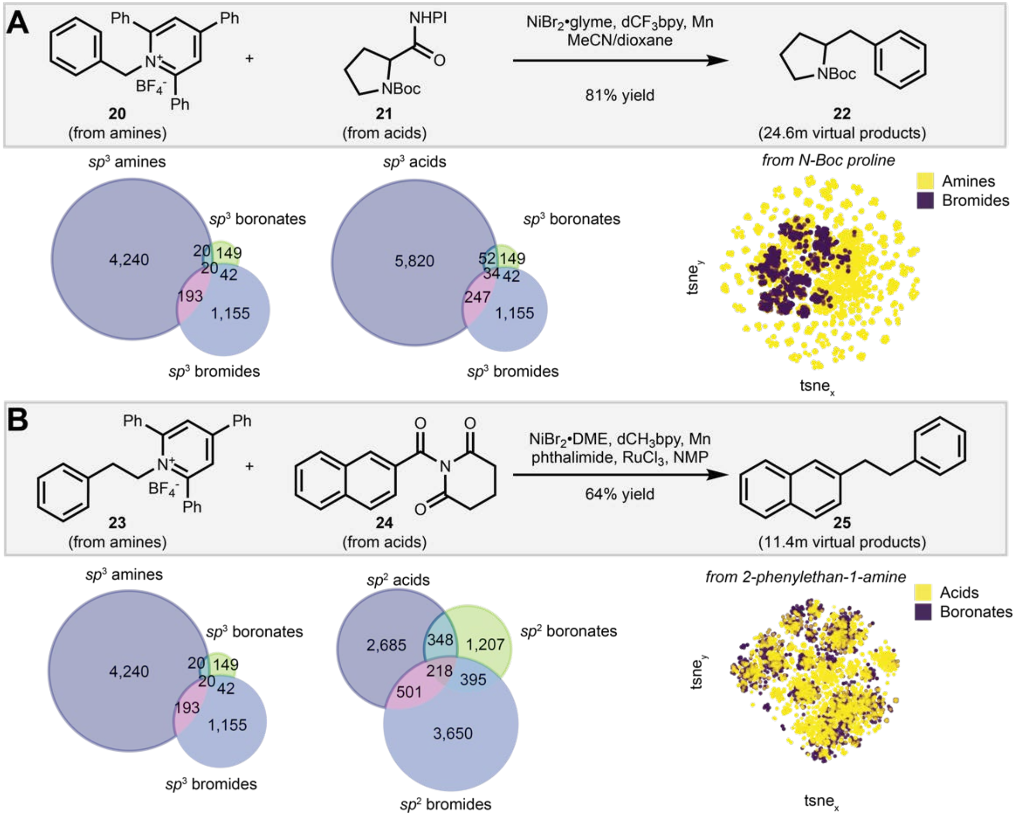


	acid	aldehyde	alcohol	amine	bromide	boronate
acid	2207					
aldehyde	2080	1604				
alcohol	1142	918	445			
amine	3050	2010	983	871		
bromide	963	552	271	432	90	
boronate	918	507	248	401	87	71

29-Fig. 2 | (A) Chord diagram linking 399 conceivable boronate–bromide transformations to 15 antiviral drugs. A transformation (yellow arc) to drug (black arc) linking lines indicate the corresponding substructure produced by the transformation exists in the respective drug. (B) All conceivable reactions (222,740 transformations) between an amine and acid provides considerable access into drug-like chemical space. (C) Distribution of transformations in examined drugs from all conceivable transformations between building blocks. (D) Number of transformations between building blocks that exist in examined drugs.



30-Fig. 3 | Scatter plot of the commercial availability of building blocks against the transformation frequency of substructures formed found when enumerating transformations for the respective building block. High value transformations identified on the Pareto front are exemplified. Isolated yields are shown. Conditions: (A) $NaBH_4$ (1 equiv.), *p*-toluenesulfonic acid monohydrate (1 equiv.), 25 ºC, 0.5 h. (B) 1) triphenylpyrylium (1.0 equiv.), ethanol (1.0 M), 80 ºC 2) benzaldehyde (1.0 equiv.), Cs_2CO_3 (1.5 equiv.), DMSO (0.1 M), 60 ºC 3) $InBr_3$ (1.0 equiv.), chloroform (0.1 M), 60 ºC. 4) $NaBH_4$ (1.0 equiv.) (C) $Pd(OAc)_2$ (0.1 equiv.), XantPhos (0.1 equiv.), KO^{t}Bu (2.0 equiv.), toluene (0.2 M), 80 ºC, 18 h. (D) KO^{t}Bu (2.0 equiv.), DMSO (0.2 M), r.t., 18 h. (E) ^{t}BuOO^{t}Bu, 130 ºC, 40 h.



31-Fig. 4 | Recently reported amine–acid chemistries analyzed for chemical space access. On the left are Venn analyses of commercially available *sp*^{3} and *sp*^{2} amines, bromines, boronates, and carboxylic acids indicating novel chemical space reachable by recently discovered amine–acid chemistry. The t-SNEs to the right compares the structural access provide when using acids or amines instead of bromides or boronates. (A) The analysis performed for an *sp*^{3}–*sp*^{3} deaminative–decarboxylative C–C cross coupling. (B) The analysis performed for an *sp*^{3}–*sp*^{2} deaminative–decarboxylative C–C cross coupling.

One Step Retrosynthesis of Drugs by Cross-Coupling from Building Block Catalogs

In this work reaction targeting was used to focus on retrosynthetic routes that formed target molecules via one transformation from one or two substrates that are purchasable. Specifically, DrugBank was used as a target catalog and the SigmaAldrich commercial catalog was used as the building block dataset. Each bond of each drug was decomposed, and the resultant synthons were enumerated with building block handles then cross-referenced against the commercial catalog. If both synthons (or synthon, in the case of an intramolecular disconnection) could be purchased, then the reaction was saved and organized by reaction type (bond formed, local atomic features). A stunning number of one-step cross coupling routes were identified.

Abstract

Enumerative combinatorics can be utilized to decompose target molecules into components. Cross-referencing these synthons to commercial datasets can reveal valuable reactivities. In this report, the compounds listed in DrugBank were mapped to Sigma-Aldrich's commercial catalog through reaction agnostic one step retrosynthetic decomposition. Decomposed synthons were linked to target molecules by transformations that form one bond after the elimination of each synthon's respective functional handle, as indicated by their building block class. The analysis is performed at the highest level, considering only the bond formed and the minimal substructure template of each building block class. Specific reactivities were analyzed after post-hoc refinement and clustering of commercial synthons. Maps between boronates, bromides, iodides, amines, acids, chlorides, alcohols, and various C–H motifs to form alkyl–alkyl, alkyl–aryl, and aryl–aryl carbon–carbon, carbon–nitrogen, and carbon–oxygen bonds are reported herein, with specific examples for each provided.

Introduction

Computer aided synthesis is a valuable tool in designing and evaluating experiments to create chemical matter efficiently and optimally.[54, 124-127] Modern retrosynthetic algorithms rely on reactivities reported in the literature to propose synthetic routes to targets,[128-133] and the exploration of chemical space is largely biased by known reactivities as indicated by the large majority of reactions run in the pharmaceutical industry being one of four or five common types (amide, Suzuki, Buchwald-Hartwig, Boc (de-)protection, etc.)[30, 134] In contrast, enumerative combinatorics can generate hypothetical transformations that have yet to be discovered between substrate classes.[1, 135] In reaction targeting, enumerated transformations are evaluated for utility and impact through a variety chemoinformatic means such as chemical space access analysis, retrosynthetic value, and property or topological distribution analysis. When extending the enumerative combinatorics algorithm to a variety of common building block classes such as amines, acids, alcohols, halides, boronates, etc., a wealth of hypothetical yet realistic and feasible reactivities are revealed.

The most valuable reactivities are those that utilize highly available commercial building blocks and provide access to many drug or druglike structures (32-Fig. 1A). In this study, we focus on a specific transformation identified through enumerative combinatorics – the general class of cross-coupling reactions that form a single bond between carbon, nitrogen, or oxygen atoms after the elimination of a building block handle, for instance by coupling **1** and **2** to form structure **3**. (32-Fig. 1B). A distributable algorithm was used to decompose compounds listed in DrugBank into synthons, which were enumerated as building blocks and cross-referenced against Sigma Aldrich's commercial catalog. In the case where two commercial compounds could cross-couple into a drug molecule, an entry was recorded. Retrosynthetic entries were visualized in bulk via chord diagrams, where one synthon (synthon a) is arrayed along the bottom-left arc, the other synthon (synthon b) is arrayed along the bottom-right arc, and target molecules are arrayed along the top arc. A chord between a synthon and the target molecule indicates that the synthon can be used to form the target in one step when merged with a compound found in the other synthon arc. (32-Fig. 1C)

In the analysis reported in this paper, building blocks are linked to synthetic targets in DrugBank through enumerated hypothetical cross-couplings. We consider the formation of a single bond between alkyl and aryl carbon atoms, nitrogen atoms, and oxygen atoms due to their high prevalence in druglike structures (32-Fig. 1D). Furthermore, the analysis focuses on common building blocks found in commercial catalogs to maximize the utility of available feedstock (32-Fig. 1E). Post-hoc refinement of the dataset enables analysis of more precise building blocks. For instance, alcohol building blocks hits can be grouped into primary, secondary, tertiary, or aryl subclasses, and C–H building blocks can be split into benzylic and non-benzylic motifs. The results reveal a wealth of viable reactivities that can form druglike molecules in one step via cross-coupling between two commercial substrates, some of which have been previously reported as methodologies. We posit that many of these reactivities can be developed through literature mining and high throughput experimentation.[108, 136, 137]

In 33-Fig. 2A, the utility of bonds formed at C–H motifs are valuable in expanding chemical space and in synthesizing drug compounds in one step from commercial compounds. Particularly, benzyl activations such as the formation of **6** from **4** and **5** or the formation of **9** from **7** and **8** are high value reactions due to their commercial prevalence. Similarly, forming a bond at a C–H motif alpha to a heteroatom is a valuable reactivity to form drugs directly, such as the formation of bromhexine **12** from **10** and **11**, which was experimentally proven. (33-Fig. 2B)

The analysis was initially performed on carbon-nitrogen and carbon-oxygen bonds. (34-Fig. 3) Examples of these disconnections that were experimentally validated on drug-like molecules are shown. We now focus on one step reactions where one synthon is an activated C–H motif. (35-Fig. 4) As an example, we show how drug **15** could theoretically be formed in one step from commercial compounds **13** and **14** using a benzylic activation. The analysis is expanded to other common building block classes. 36-Fig. 5 and 37-Fig. 6 showcase the chord diagrams and example reactions (**16** – **33**) for one-step syntheses of drugs from alcohols, boronates, amines, acids, and halides, organized by bond formed.

Methods

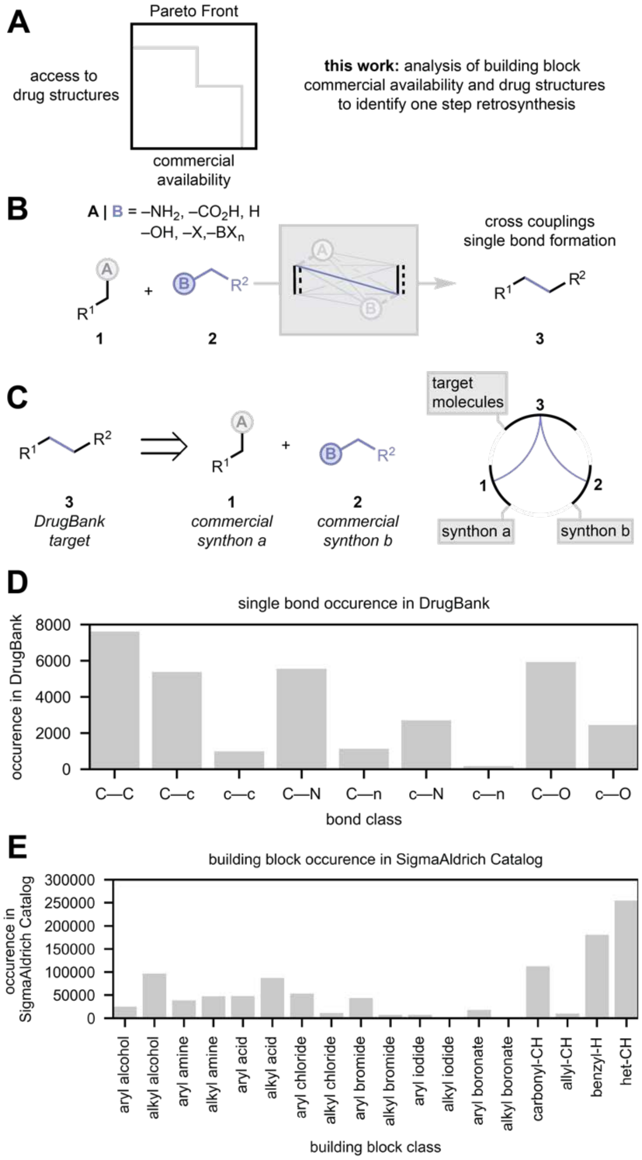
In the analysis reported herein, the DrugBank dataset and commercial catalogs were filtered to compounds with molar mass less than 500 and 200, respectively, and desalted. A command line script was written to cross reference targets with commercial datasets. Given a commercial data set, target molecule, and building block parameters, the script will produce the one step disconnection map between the catalog and the target molecule. This script is massively parallelizable and can be utilized trivially on high performance clusters.

The algorithm is briefly described:

1. Identify target bond in target compound dataset. Store all compounds that contain the target bond.
2. Decompose the target bond for all hits, forming two synthons
3. Enumerate the synthons at the disconnection point to generate building block analogs
4. Cross-reference the enumerated building blocks against the commercial catalog. Store the transformation if both synthons can be found as commercial analogs.

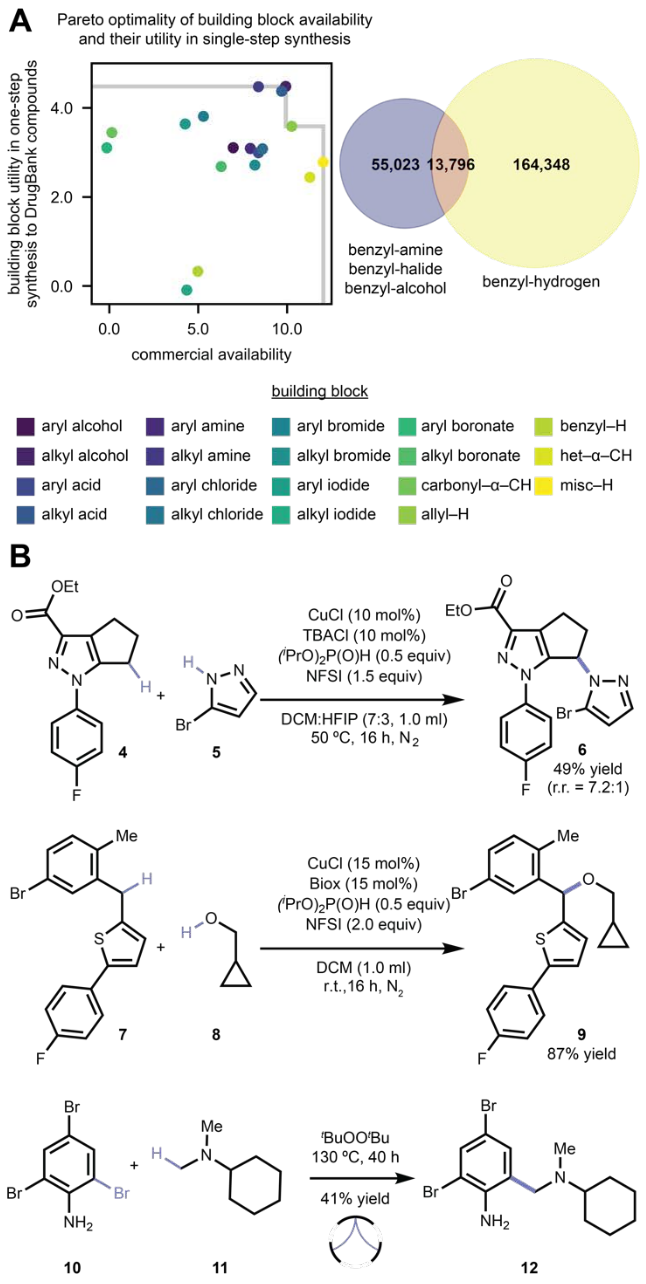
Conclusion

A method to rapidly form single-bond cross coupling maps between target molecules and commercial catalogs is reported and showcased via an analysis of compounds in DrugBank and the commercial catalog provided by Sigma Aldrich. Coupling maps are grouped and filtered to reveal various perspectives of reaction targeting. Several reactivities that have been previously reported were showcased as transformations found in the analysis, and examples of prospective chemistries are shared. Mining other maps of coupling modes will continue to reveal the most valuable reactions to target in synthetic campaigns.

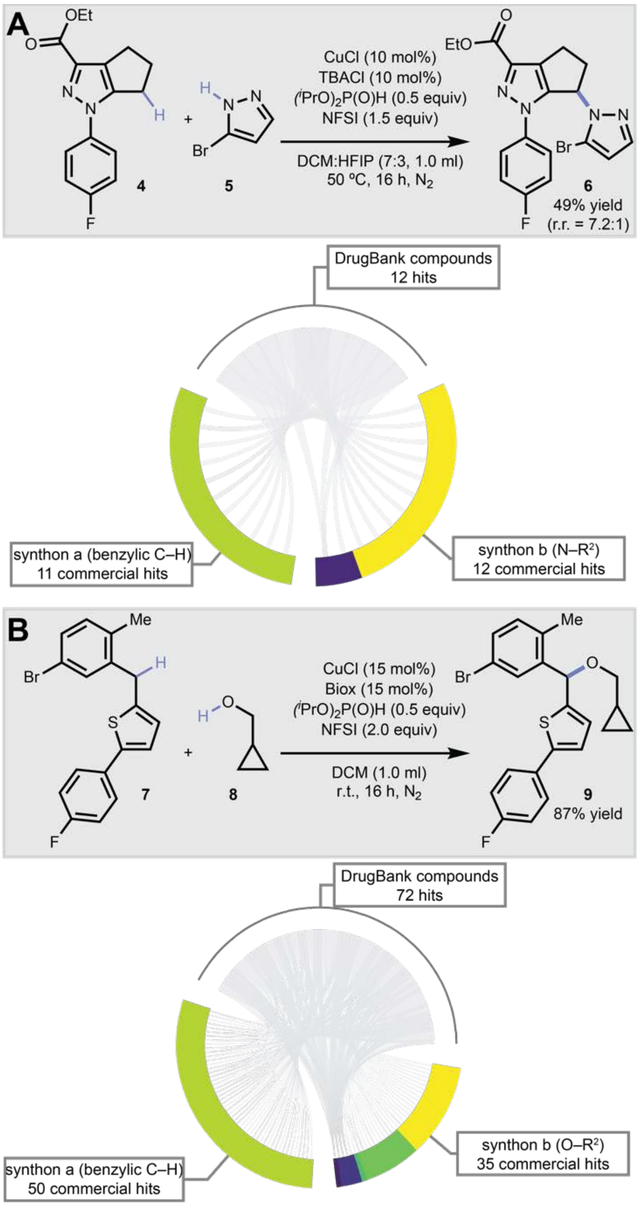


32-Fig. 1 | An analysis of commercial building blocks and reported drug structures. **a)** A Pareto front cartoon exemplifying the value of certain building blocks. Building blocks that are highly available and can be used to form drugs are of high value. **b)** Enumerative combinatorics can identify target bonds between building blocks. In this study we focus on cross-couplings that form one bond after the elimination of both building block handles. **c)** Retrosynthesis of target and its representation as a chord diagram. **d)** Prevalence of bond types in DrugBank compounds. The alkyl-alkyl C–C bond is the most common bond type found, followed by the alkyl-alkyl C–N and C–O bonds and the alkyl-aryl C–C bond.

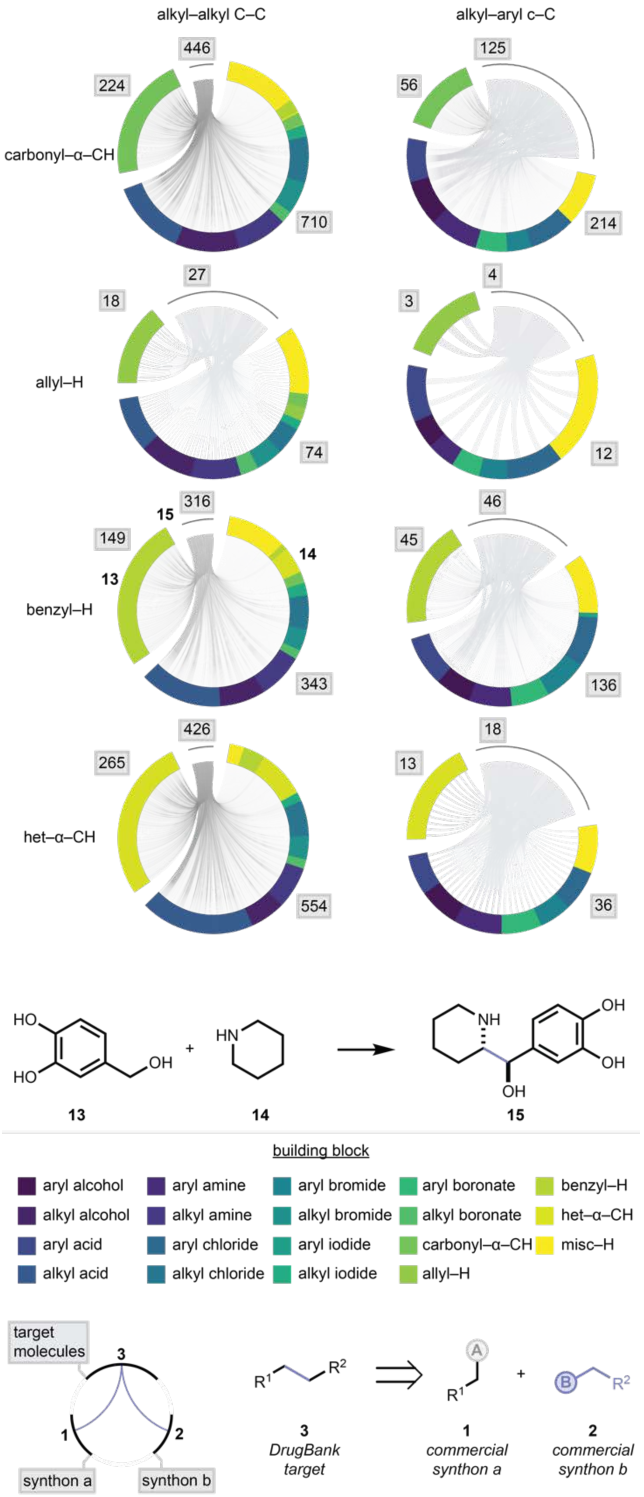
e) Prevalence of building block classes in the Sigma Aldrich commercial catalog.



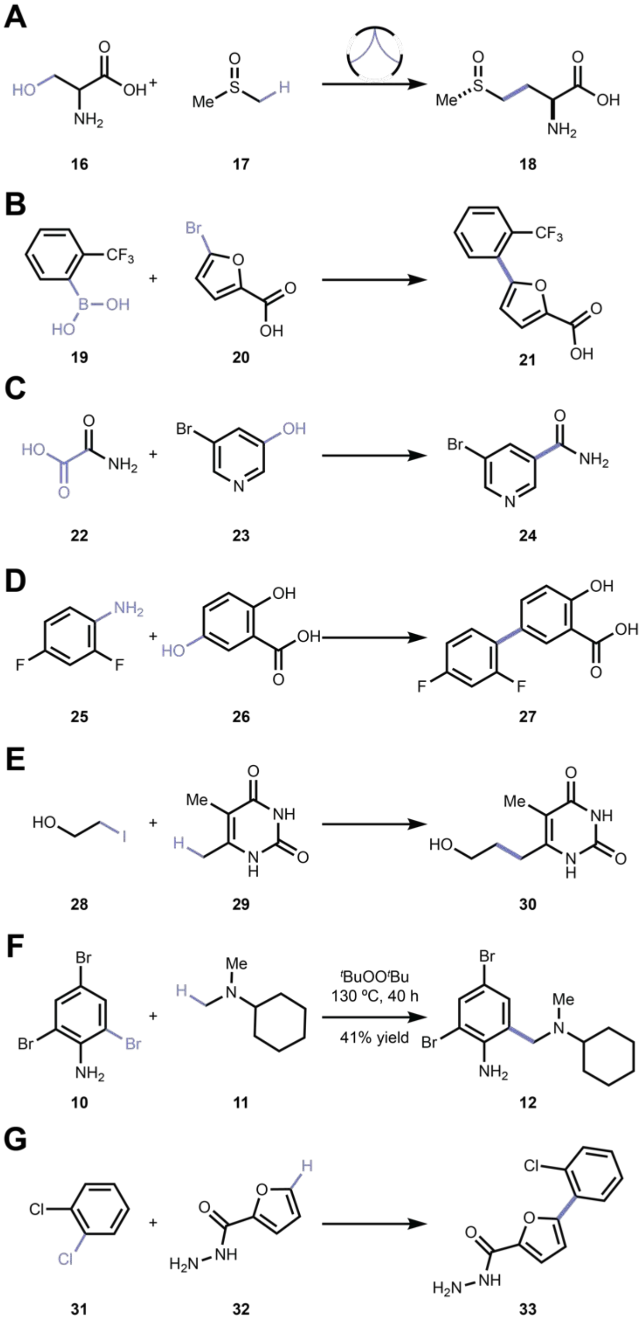
33-Fig. 2 | C–H motifs can be viewed as valuable and versatile building blocks. a) Pareto front and building block analysis of the commercial dataset. Classes of building blocks that have many commercial substrates and can be found often in one step retrosyntheses of drug compounds lie on the Pareto front. Core structures of benzyl–H and benzyl–amine, benzyl-halide, and benzyl-alcohol compounds in the commercial dataset are compared in the Venn diagram. Many benzylic compounds do not exist as their amine/halide/alcohol functionalized counterpart. **b)** Examples of couplings using unconventional building blocks.



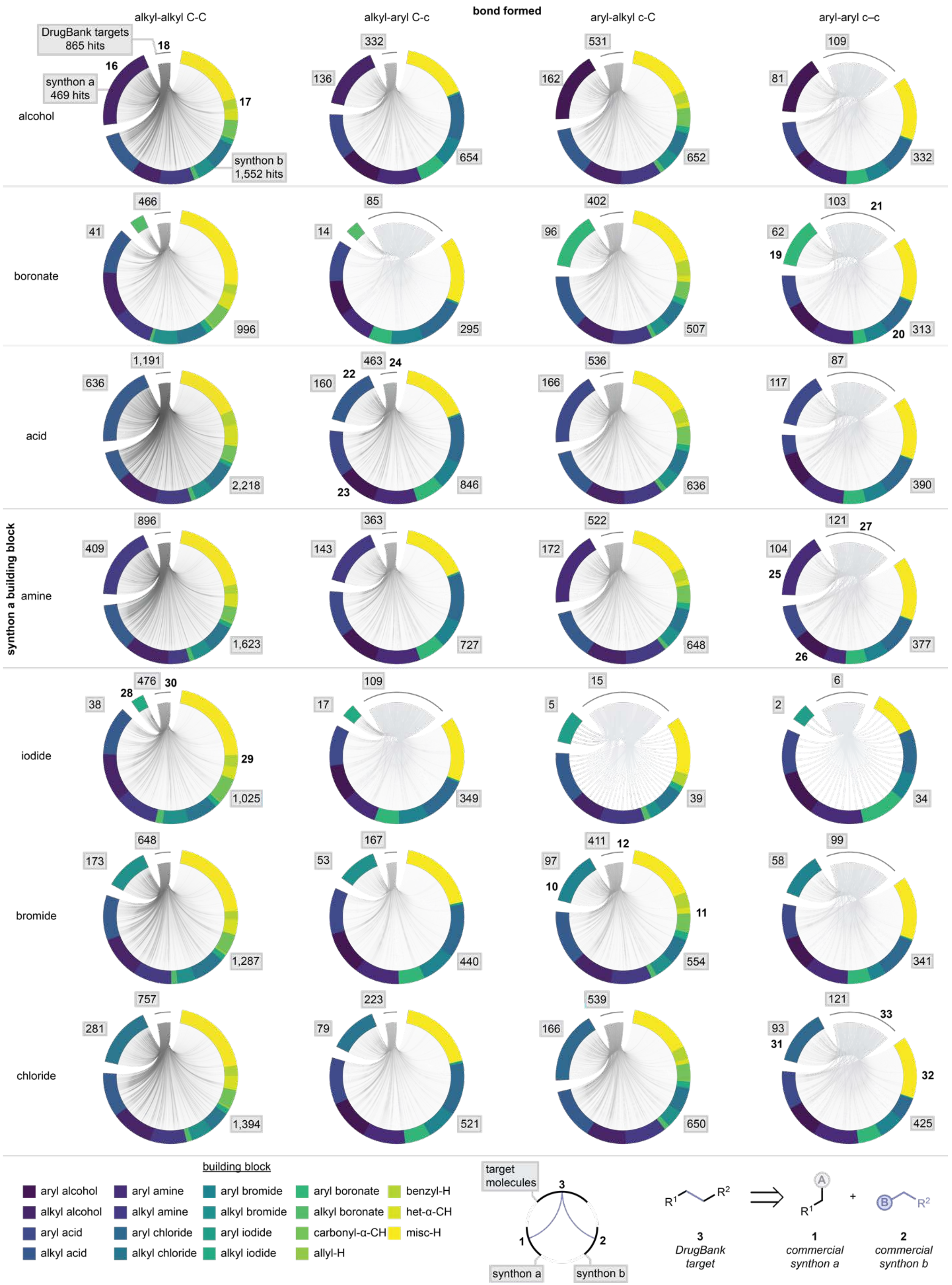
34-Fig. 3 | a) All hypothetical single bond C–H cross couplings between benzylic carbon atoms and nitrogen atoms to form DrugBank compounds from commercial substrates. A benzylic functionalization previously reported by the Stahl group is exemplified as a reaction identified in the analysis.[138] **b)** All hypothetical single bond C–H cross couplings between benzylic carbon atoms and oxygen atoms to form DrugBank compounds from commercial substrates. Another example from the Stahl group is showcased here as a reaction found by the targeting algorithm.[139].



35-Fig. 4. | One step cross coupling maps for various activated C–H motifs. In the first row and column, synthon a is grouped by commercial compounds that react at a carbon alpha to a carbonyl group with another alkyl building block. In the third row and first column, synthon a is grouped by compounds that couple at a benzylic hydrogen with another building block, such as benzyl alcohol **13** reacting with **14** to form drug **15**.



36-Fig. 5 | a-g) Examples of one step syntheses found in the analysis. **f)** The synthesis of bromhexine identified in this analysis was previously validated.


bond formed
alkyl-alkyl C-C
alkyl-aryl C-c
aryl-alkyl c-C
aryl-aryl c–c
synthon a building block
DrugBank targets
865 hits
synthon a
469 hits
synthon b
1,552 hits
alcohol
18
16
17
136
332
654
162
531
652
81
109
332
boronate
466
41
996
85
14
295
402
96
507
103
21
62
19
20
313
acid
1,191
636
2,218
463
24
22
160
23
846
536
166
636
87
117
390
amine
896
409
1,623
363
143
727
522
172
648
121
27
104
25
26
377
iodide
476
30
28
38
29
1,025
109
17
349
15
5
39
6
2
34
bromide
648
173
1,287
167
53
440
411
12
97
10
11
554
99
58
341
chloride
757
281
1,394
223
79
521
539
166
650
121
33
93
31
32
425
building block
aryl alcohol
alkyl alcohol
aryl acid
alkyl acid
aryl amine
alkyl amine
aryl chloride
alkyl chloride
aryl bromide
alkyl bromide
aryl iodide
alkyl iodide
aryl boronate
alkyl boronate
carbonyl-α-CH
allyl-H
benzyl-H
het-α-CH
misc-H
target molecules
3
1
2
synthon a
synthon b
R1
R2
A
B
3
DrugBank target
1
commercial synthon a
2
commercial synthon b

37-Fig. 6 | Single step cross coupling maps between DrugBank and purchasable compounds in Sigma-Aldrich’s catalog, trellised by synthon A building block and bond formed. In the first column, both synthons link at an alkyl carbon. In the second column, synthon A, which is the building block group for the row, is alkyl and the remaining synthon b is aryl. In the third column, the building block arc contains aryl synthons and synthon b contains alkyl synthons. In the final column, both synthons are aryl. Compound numbers refer to retrosynthetic reactions shown in 36-Fig. 5. An example of a one-step synthesis of bromhexine experimentally validated is shown above and indicated on the bromide aryl-alkyl C-c bond chord plot.[54]

Chapter III. High Throughput Experimentation

The dawn of miniaturized chemistry has ushered in an era that closely mirrors the evolution of high throughput screening platforms used in biological assays. Traditionally, synthetic experiments were large-scale and labor-intensive; however, the advent of miniaturized chemistry has allowed these experiments to be performed in a compact, plated format.[35, 36, 110-113, 134, 136, 140, 141] Additionally, the rise of robotics and liquid-handling technologies has made a significant impact on the field. These cutting-edge innovations have transformed synthetic protocols, making them more conducive to automation.[134] With the inclusion of such advanced techniques, the process has not only become more efficient but also more precise, greatly enhancing the quality of experiments.

The advantages of these advancements are not limited to increased efficiency and precision. In fact, they have paved the way for the collection of extensive, standardized reaction datasets. These comprehensive datasets provide an ideal basis for training predictive models using advanced data science techniques.[114, 125, 137] The ability to collect and analyze such large volumes of data enables researchers to make accurate predictions and gain in-depth insights. High throughput experimentation, with its capacity for large-scale data collection and analysis, holds immense promise for discovering numerous high-value transformations, which can be identified through reaction targeting.[101-104]

Automatable reaction protocols that function in miniaturized formats are revolutionizing the synthesis of pharmacophore libraries, streamlining processes that were once complex and time-consuming. This progress, combined with the development of direct-to-biology protocols, is significantly enhancing the effectiveness of high throughput experimentation campaigns in drug discovery. This chapter aims to present an overview of recent advancements in high throughput experimentation, as well as methodologies discovered and optimized using this technique. It also discusses the limitations of contemporary electronic lab notebooks in capturing the intricacies of plated reaction arrays, introducing the development of phactor, our laboratory's high throughput experimentation manager. Developed over several years and refined based on feedback from over 100 research chemists, phactor promises to significantly improve the management of high throughput experimentation.

Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery

This manuscript details the development of phactor, a software solution for high-throughput experimentation (HTE) in reaction discovery and stands out as a significant milestone. During its creation, we recognized the need for software to navigate data-rich experiments in chemical laboratories. As a result, phactor was designed to streamline HTE performance and analysis. It enabled researchers to rapidly design arrays of chemical reactions or direct-to-biology experiments using different wellplate formats, access online reagent data, and generate instructions for manual execution or automated handling. Importantly, phactor stores all chemical data, metadata, and results in machine-readable formats, fostering standardized data management and facilitating subsequent analysis. The project's impact became evident as phactor not only expedited experiment ideation and result interpretation but also fostered creativity and supported reaction discovery and optimization through the availability of standardized reaction data. Notably, the software's successful application in identifying a potent inhibitor of the SARS-CoV-2 main protease highlighted its efficacy. To encourage collaboration, we made phactor™ freely available for academic use in select wellplate formats.

Abstract

High-throughput experimentation (HTE) is an increasingly important tool in reaction discovery. While the hardware for running HTE in the chemical laboratory has evolved significantly in recent years, there remains a need for software solutions to navigate data-rich experiments. Here we have developed phactor™, a software that facilitates the performance and analysis of HTE in a chemical laboratory. phactor™ allows experimentalists to rapidly design arrays of chemical reactions or direct-to-biology experiments in 24, 96, 384, or 1,536 wellplates. Users can access online reagent data, such as a chemical inventory, to virtually populate wells with experiments and produce instructions to perform the reaction array manually, or with the assistance of a liquid handling robot. After completion of the reaction array, analytical results can be uploaded for facile evaluation, and to guide the next series of experiments. All chemical data, metadata, and results are stored in machine-readable formats that are readily translatable to various software. We also demonstrate the use of phactor™ in the discovery of several chemistries, including the identification of a low micromolar inhibitor of the SARS-CoV-2 main protease. Furthermore, phactor™ has been made available for free academic use in 24- and 96-well formats via an online interface.

Introduction

Miniaturized high-throughput experimentation (HTE) has emerged as an accessible, reliable, economical, and environmentally friendly technique for the rapid discovery of new reactivities.[35, 36, 108-113, 124, 134, 140-162] Curated HTE data has proven to be increasingly valuable for predictive models.[144-149] While in experimental practice, the community has gravitated towards liquid handling techniques in glass shell microvials with tumble stir dowels, or in plastic 384 or 1,536 wellplates[35, 36, 108-112, 134, 140-143], a standard for HTE data handling has yet to be established. The organizational load required to perform a simple 24-well reaction array is generally manageable by repetitive notebook entries or with spreadsheets, yet managing multiple reaction arrays in a single day, or running ultraHTE in 1,536 wellplates,[108] is challenging without information management software. Moreover, no readily available electronic lab notebook (ELN) can store HTE details in a tractable manner[150, 151] or provide a simple interface to extract data and results from multiple experiments simultaneously.[152, 153] Contemporary HTE software provide HTE solutions but are only commercially available.[163, 164] To continue developing HTE research and position data outputs for machine learning studies, detailed reaction data must be easily accessible for standardized rapid extraction and analysis.[154-156]

With these issues in mind, we developed the software phactor™ to streamline the collection of HTE reaction data. Our primary objective was to develop a robust yet generalizable HTE workflow solution that captures the nuances of chemical experimentation while reporting data in a standardized, machine-readable format. phactor™ minimizes the time and resources spent between experiment ideation and result interpretation. This enables creativity by freeing up time otherwise used thinking about experiment logistics, facilitates reaction discovery and optimization, and serves as a tool to bolster the amount of available reaction data reported in a standardized format.

We have provided phactor™ as a free web service to the academic community, currently supporting 24- and 96-well formats, which can be accessed at https://phactor.cernaklab.com.

Results

phactor™ workflow overview

The workflow of a typical high-throughput experiment involves design of the reaction array, preparation of reagent stock solutions, dosing of stock solutions according to the reaction array recipe (either by hand or with robotics), analysis of reaction outcome, followed by visualization and analysis of data and documentation of results. A standardized reaction template classifies substrates, reagents, and products (38-Fig. 1a). Interconnecting experimental results with online chemical inventories through this shared data format creates a closed-loop workflow for HTE-driven chemical research (38-Fig. 1b) and enables rapid reaction array design and analytics. While developing phactor™, we sought to maximize the automation of data movement and processing. Recognizing the rapidly accelerating chemical research software ecosystem,[28, 157-162, 165-178] the philosophy behind phactor™'s data structure was to record experimental procedures and results in a machine-readable yet simple, robust, and abstractable format to naturally translate to other system languages (Supplementary Fig. 8). As such, the inputs and outputs of phactor™ can be procedurally generated or modified with basic Excel or Python knowledge to interface with any robot, analytical instrument, software, or custom chemical inventory containing metadata (e.g., the organophosphorous ligand platform

Kraken[179]) such as reagent location, molecular weight, CAS number or SMILES string. Examples of interfacing phactor™ outputs with ORD,[149] XDL,[178] or EDBO+[176] are shown in Supplementary Figs. 14-16.

The event workflow of a typical phactor™ experiment is shown in 38-Fig. 1c. The user selects desired reagents from the inventory for automatic field population or enters specific reagent entries manually, such as for a custom substrate. Once all relevant chemicals are selected, the reaction array layout is designed automatically or manually, as the user prefers. Reagent distribution instructions are generated to be executed either manually or by an interfacing liquid handling robot. Last-minute changes in the face of unforeseen circumstances during reaction setup, such as poor chemical solubility, chemical instability, or the need to premix reagents before dosing can be made at any time. Stock solutions are prepared in vials or wellplates and distributed to their respective locations on the reaction wellplate. Once the reactions are complete, they are quenched and analysed. Any data with a well location map can be uploaded. This allows both data on reaction performance (e.g., UPLC-MS conversion) and biological assay results (e.g., bioactivity data) to be viewed in concert.

The workflows for executing HTE can vary depending on available equipment and desired throughput of the experiment. phactor™ incorporates these parameters into its user interface to ensure a consistent workflow experience agnostic of hardware capabilities (38-Fig. 1d). Examples herein demonstrate phactor™'s use with manual dosing as well as integration with the Opentrons OT-2 liquid handling robot for experiments of 384-well throughput or less, and the SPT Labtech mosquito® robot for 1,536-well ultraHTE (38-Fig. 1e). Regardless of instrumentation or throughput, all results are stored in the same format, facilitating analysis of results across multiple experiments. Reaction discovery and library synthesis campaigns utilizing standard 24-, 96-, 384- and 1,536-well experiments are described in detail.[1, 54, 102, 180-182]

Experimental Analysis

Phactor™ facilitates the discovery of new reactivity. Our lab is broadly interested in amine-acid coupling reactions[1, 102, 180-182] and particularly amine-acid C–C coupling reactions.[102, 181] Diverse chemistries discovered with the aid of phactor™ are shown in 39-Fig. 2. 39-Fig. 2a shows the discovery of a deaminative aryl esterification.[180] In the reaction array design, an amine, activated as its diazonium salt (**1**), a carboxylic acid (**2**), one of three transition metal catalysts, with one of four ligands, in the presence or absence of a silver nitrate additive were to be dosed to each reaction well in acetonitrile, then stirred at 60 ºC for 18 hours. phactor™ automatically designed the reagent distribution recipe by splitting the plate into a simple four-row and six column multiplexed array. At completion, a solution containing one molar equivalent of caffeine was added to each well as an internal standard. An aliquot of each reaction was transferred to a plastic wellplate, then diluted with acetonitrile for UPLC-MS analysis of the desired ester product (**3**). UPLC-MS output files were analysed by the commercial software Virscidian Analytical Studio, which provided a CSV file containing peak integration values for each of the 24 chromatographic traces. This file was fed into phactor™ to record the experimental outcome and produce the heatmap shown in 39-Fig. 2a. Analysis on phactor™ indicated an 18.5% assay yield when using 30 mol% CuI, pyridine and $AgNO_3$, and these specific conditions were triaged for further study.

In the example of 39-Fig. 2b, we optimized the penultimate step in our synthesis of umifenovir,[54] an oxidative indolization reaction between **4** and **5** to produce **6**. Inspired by the conditions of Glorius,[183] a reaction array was performed using copper catalysts and ligand/additive combinations. Four copper sources at 20 mol%: cuprous iodide, cuprous bromide, tetrakis(acetonitrile) copper(I) triflate, or cupric acetate, were distributed into the four rows while combinations of magnesium sulfate (0.0 equiv or 1.0 equiv) with 2-(1H-tetrazol-1-yl)acetic acid (**L1**), or 2,6-dimethylanilino(oxo)acetic acid (**L2**) at 40 mol% were distributed into the columns as DMSO solutions, with 3.0 equivalents of cesium carbonate added to every well as a suspension in DMSO. The reactions were manually arrayed in a glovebox, sealed, and stirred at 55 ºC for 18 hours. Well B3 (copper bromide with **L1** and no magnesium sulfate) was found to perform best, and a 0.10 mmol scale-up reaction produced desired indole **6** in 66% isolated yield.

In 39-Fig. 2c, the allylation of furanone **7** or furan **8** with reagents **9** or **10** was investigated. For each combination of nucleophile and electrophile, one of three ratios of Pd_2dba_3 to (*S*,*S*)-DACH-phenyl Trost ligand **L3** was added, followed by the addition or omission of potassium carbonate as a base. Each reaction was run in toluene for 24 hours at room temperature, quenched, and then analysed by UPLC-MS for conversion and selectivity. Multiplexed pie charts generated by phactor™ revealed that the conditions of well D3, with a 2:1 palladium catalyst to ligand loading and no base generated the desired γ-regioisomer with the greatest selectivity, along with α-allylation and its olefin isomer when **8** was used.

An organocatalyzed asymmetric Mannich reaction is detailed in Supplementary Fig. 17. Aldehyde **13**, *p*-anisidine (**14**), and ketone **15** are used to form desired product **16** via a solvent and catalyst reaction array that was analysed by TLC, which revealed the formation of undesired product **17**.

With phactor™, HTE becomes an exercise in workflow execution, with automation of the organizational aspect of the experiment. This allows chemists to focus on the design and analysis of the reaction array, rather than workflow details (40-Fig. 3). 40-Figs. 3a–3c display three examples of 24-well experiments. 40-Fig. 3a shows an amide coupling reaction array performed in preparation for an ultra-high throughput direct-to-biology assay (vide infra), with the aim of producing inhibitors of the SARS-CoV-2 main protease (M^{Pro}).[184] A screen of three anilines, two coupling agents HATU or DCC/HOBt and a carboxylic acid, with or without base, produced all desired products. Amide **18** was produced in high conversion using HATU, DMAP, and DIPEA. 40-Fig. 3b and 40-Fig. 3c detail the optimization of a recently published deaminative decarboxylative sp^2–sp^3 C–C coupling from an acid activated as an *N*-acyl-glutarimide and an amine activated as a Katritzky salt.[181] Reaction array 3b screened 24 ligands in NMP under nickel catalysis to generate product **19**, which was subsequently found with improved yield when using a bipyridine ligand and phthalimide as additive (40-Fig. 3c).

40-Figs. 3d-3f contain 96-well experiments designed with phactor™. 40-Fig. 3d and 40-Fig. 3e show the results from a recently reported deaminative-decarboxylative sp^3–sp^3 C–C coupling coupling[181] that generated products **20** and **21**. Three catalysts, four additives, four ligands, and two reductants were tested in 40-Fig. 3d, and it was found that **20** was afforded at 47% product/internal standard with $NiCl_2$•glyme as the catalyst, 4,4'-di-*tert*-butyl-2,2-bipyridine as ligand, zinc as the reductant, and lithium perchlorate as an additive. After additional optimization, the reaction array in 40-Fig. 3e was run to test the reaction's scope. Product **21** was afforded with optimal conditions of $NiBr_2$•glyme, 4,4'-bis-trifluoromethyl-2,2'-bipyridine, and manganese in 1:1 dioxane:acetonitrile. 40-Fig. 3f shows the results of a 12 ligand – eight metal catalyst reaction array further developing the analogous sp^2–sp^3 decarboxylative deaminative C–C coupling, optimizing the yield of model product **22**.[102]

40-Figs. 3g–3i show several ultraHTE reaction arrays. 40-Fig. 3g and 40-Fig. 3h are reaction arrays produced in the development of our aryl amine esterification reaction.[180] Both reaction arrays were substrate scope experiments, producing ester compounds such as **23** and **24.** In 40-Fig. 3g, lutidine and $Cu(MeCN)_4BF_4$ in benzonitrile showed good scope. Replacing lutidine with collidine expanded the reaction scope and reproducibility across quadruplicate measurements as shown in 40-Fig. 3h. Finally, 40-Fig. 3i contains the UPLC-MS results of a direct-to-biology reaction array towards the creation of SARS-CoV-2 M^{Pro} inhibitors, generating amides such as **25**. Files needed for all 24 and 96 well experiments shown in 40-Fig. 3 are provided as examples to learn the phactor™

workflow. The layouts for several reaction arrays are displayed in the Provided Examples section of the Supplementary Information.

Discovery of a low micromolar inhibitor of SARS-CoV-2 M^{Pro} via direct-to-biology assay.

An amide coupling experiment was planned based on an inventory of amines, largely anilines, and a tetrahydroquinoline carboxylic acid pharmacophore (**26**) found to be a potent inhibitor of M^{Pro}.[184] A preliminary 24-well amide coupling reaction array was performed to test the effectiveness of various amide coupling conditions for the acid and anilines (see 40-Fig. 3a and Supplementary Fig. 21). A diluted aliquot from each well was subjected to a RapidFire™ MS binding assay to determine concentration-response curves.[185] Curves were found to correlate with yield and literature reported IC_{50} values were replicated utilizing this assay.[184] Following an optimized direct-to-biology ultraHTE workflow (41-Fig. 4a), amide coupling reactions were executed with the aim of making diverse amides, which were directly tested for activity against M^{Pro} in a single experiment. Conducting an experiment of this complexity using spreadsheet software would be challenging. However, with the use of phactor™, ultraHTE and direct-to-biology experiments can be swiftly developed and assessed within a matter of minutes, especially if corresponding input or inventory files are in hand. Eight amide coupling conditions were tested in duplicate for each of the 80 amines, resulting in 1,280 reactions (41-Fig. 4b). A key step in this workflow is the distribution of the 1,536-well reaction plate into four 384-well analysis assay plates suitable for UPLC-MS or RapidFire™ analysis. As such, some of the wells in the 1536-well reaction plate were not utilized to account for the four control columns necessary in each of the four 384-well RapidFire™ assay plates to allow calculation of Z prime for the assay (0.961).[186] This distribution as well as the chemical and biological assay results are shown in 41-Fig. 4c and 41-Fig. 4d. Additional data analyses comparing chemical yield to biological response are shown in Supplementary Figs. 24 and 25. We note that reproducibility is a common concern in HTE and ultraHTE, and analyses of repeat experiments are provided in Supplementary Fig. 26. Both chemistry and biology assays are shown to be consistent, with 87% and 93% of data points having less than 10% error in the respective assays. From these analyses, three amides (**27**, **28**, and **29**) were chosen for scale-up and IC_{50} determination, two of which (**27** and **28**) were previously unreported in the literature. Compound **28** was found to have a IC_{50} of 5.06 μM (41-Fig. 4e), competitive with the best known M^{Pro} inhibitors in this series.[184] Notably, IC_{50} trends from pure compounds isolated on larger reaction scale are well matched to the percent inhibition trends obtained on the nanomole scale.

Methods

General phactor™ workflow for general use and to recreate the chemistry reported in this and other manuscripts.[1, 54, 102, 180-182]

The workflow for phactor™ contains six stages: Settings, Factors, Chemicals, Analysis, and Report (42-Fig. 5). The first stage, *Settings*, simply begins the plate development process when provided an experiment name, reaction volume, and throughput

(Supplementary Fig. 1a). The remaining stages can be optionally expedited with pre-generated workflow files that match a specified format.

The workflow files needed to reproduce the chemistry shown in 39-Figs. 2a-b, and 40-Figs. 3a-f are provided in an online repository. To replicate these studies, a 24- or 96-well reactor block, the corresponding amount of glass vials and stirbars, a hotplate with magnetic stirring, and autopipette will be required. The necessary solvents and reagents will also be required to the amounts as calculated on phactor™. Detailed specifications of hardware and chemical sources are provided in the Experimental Section of the Supplementary Information. Having these files in hand provide an expedited utilization of phactor™ – the user can directly drop any file that matches the chemical input specification directly into the Chemicals stage and click ‘Run’ to instantly get the procedural instructions for the experiments. Likewise dropping the provided corresponding analytical file into Analysis will display out reported results of the experiment.

Phactor™ can automatically distribute a list of chemicals into a multiplexed reaction array given that the chemicals are labeled as one of the reagent classes within the set {Electrophile, Nucleophile, Catalyst1, Catalyst2, Ligand1, Ligand2, Base/Acid, Reductant/Oxidant, Solvent1, Additive, or Other} and that the product of the amounts of each reagent class within the list is equal to or less than the wellplate size. For instance, a list containing six electrophiles, four nucleophiles, and a solvent can be automatically “plated” by phactor™. The desired plate design is specified in the *Factors* stage (Supplementary Fig. 1b).

Once the factors have been input and saved, the user can add reagents to be used in the experiment in the *Chemicals* stage (Supplementary Fig. 1c). These reagents can be either input by hand via textbox, where the user personally reports the reagent name, molecular weight, reagent class label, and density or reagents can be added directly from an online inventory. All chemicals are associated with a colour, which can be changed by clicking on the chemical’s corresponding colour bar in the table. If factors were specified on the previous stage, a checklist is provided to ensure the requirements for the automatic distribution are met. Once all lines are green, hitting run will design the plate on the following Grid stage. A third alternative to populating reagents is to upload a CSV file containing reagent data in a specified format. As mentioned, example files corresponding to the experiments in this manuscript are provided in an online repository. In the case where the input reagent file contains a distribution of reagents that would perfectly fill out the reaction plate, the factors are automatically updated and thus factors need not be specified on the factors stage.

Following the *Chemicals* stage is the *Grid* stage, which is prepopulated if the automatic distribution feature was utilized. On this stage, an interactive grid allows the user to add or remove reagents to wells in bulk. Individual wells in the digital reaction array can be selected to view reaction details. In addition to the reaction array grid, a table containing the recipes for stock solutions, which indicate the mass of reagent and volume of solvent to be added to each source vial. The volume and locations of aliquots from each source vial are shown as well, to be executed via autopipette or liquid handling robot. Finally, a suite of buttons interfacing the designed reaction array to a variety of hardware and software are provided. On the provided version, MassLynx and Virscidian files are preformatted and provided with prepopulated metadata to facilitate the characterization of the reaction array.

When the reaction array is fully designed, the recipe can be downloaded. Stock solutions are dosed according to the generated procedure, and the reaction plate is moved to a reactor with the appropriate atmosphere, pressure, and temperature. After the designated reaction time has elapsed, the reactions are quenched, and aliquots of each reaction are transferred to an analytical plate containing internal standard in a UPLC-MS compatible solvent. Once the analytical plate has been characterized, the resultant file containing well locations and output results can be dropped into the *Analysis* stage of phactor™. Here an interactive triptych display allows the user to view several output results in the same view. Clicking on a well displays the input and outputs of the reaction, as well as an image of the product if specified.

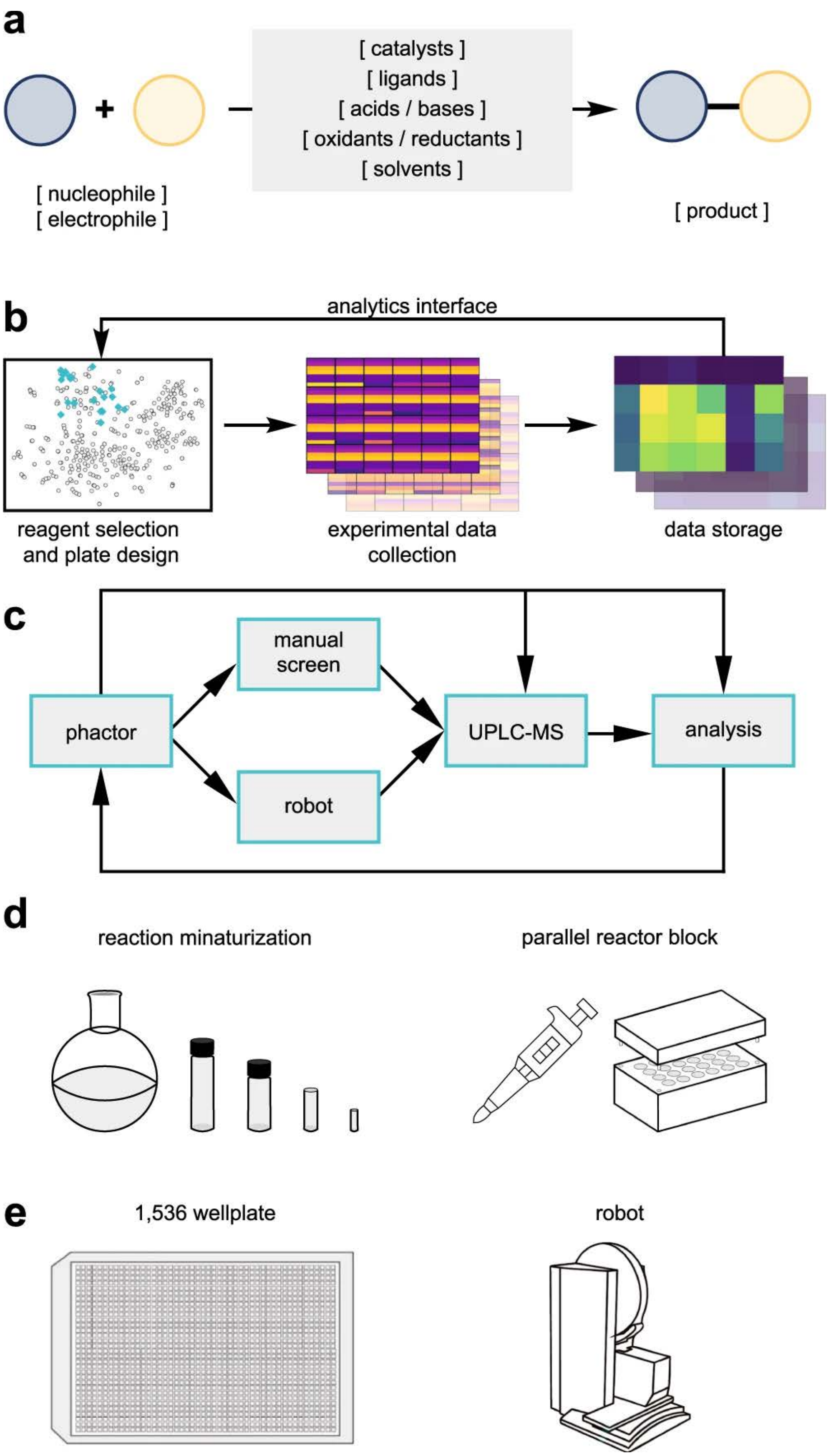
The final stage is *Report.* An overall one-page summary of the results is shown here via visualizations of inputs and outputs. Additionally, statistics of the results are calculated, and the top performing reagents are shown. Most importantly, here is where the user can download the compiled results of the entirety of the reaction array in a single CSV file. This file contains the inputs and metadata associated with each reaction of the reaction array, including reagent labels, SMILES, molarities, and solvents, as well as the output values associated with each reaction. This file can be trivially collated with other plate outputs to make supermassive reaction datasets.

Machine learning Python scripts for phactor™ output.

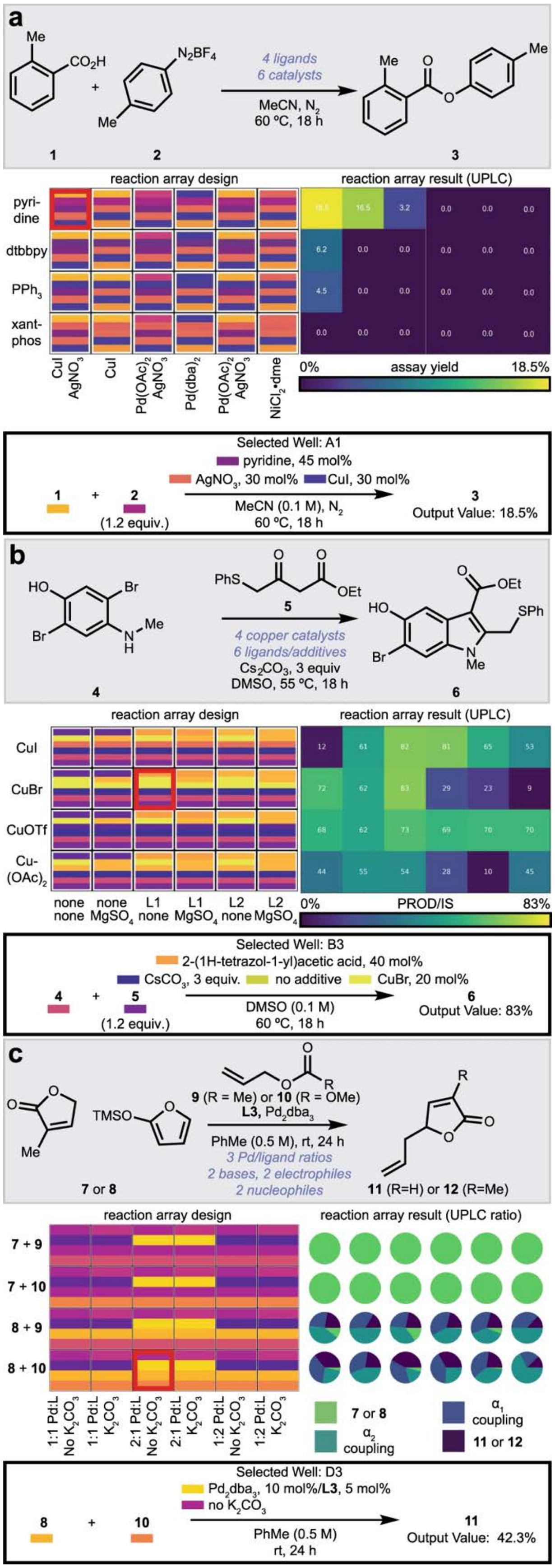
In the provided repository an example Python notebook is provided to facilitate downstream machine learning analyses from the phactor™ output file. In this example notebook, tSNE plots are generated from collated from a deaminative-decarboxylative sp^3–sp^3 C–C coupling optimization campaign dataset (43-Fig. 6). Product SMILES are encoded into fingerprints using RDKit, and fingerprint matrices are fed into the tSNE algorithm to reduce the dimensionality of the dataset into two dimensions. These are subsequently plotted and coloured by the average output value of each product to provide a rapid visualization of the reaction manifold. Areas of reaction success and failure are then readily identified in a singular display. Notebooks for conversion to heatmaps and kernel density estimate plots are also provided. Sample notebooks converting phactor™ output to ORD, EDBO+, and XDL are also provided within the repository. All the provided scripts can be easily modified by changing the input file string to work with other datasets.

In conclusion, we present the HTE ELN phactor™, which records all details of an experiment to allow for robust reproduction and accelerated discovery. phactor™ stores all experimental details in a machine-readable yet tractable and interpretable format with an SQL database and on the cloud to facilitate the use of downstream statistical analysis.

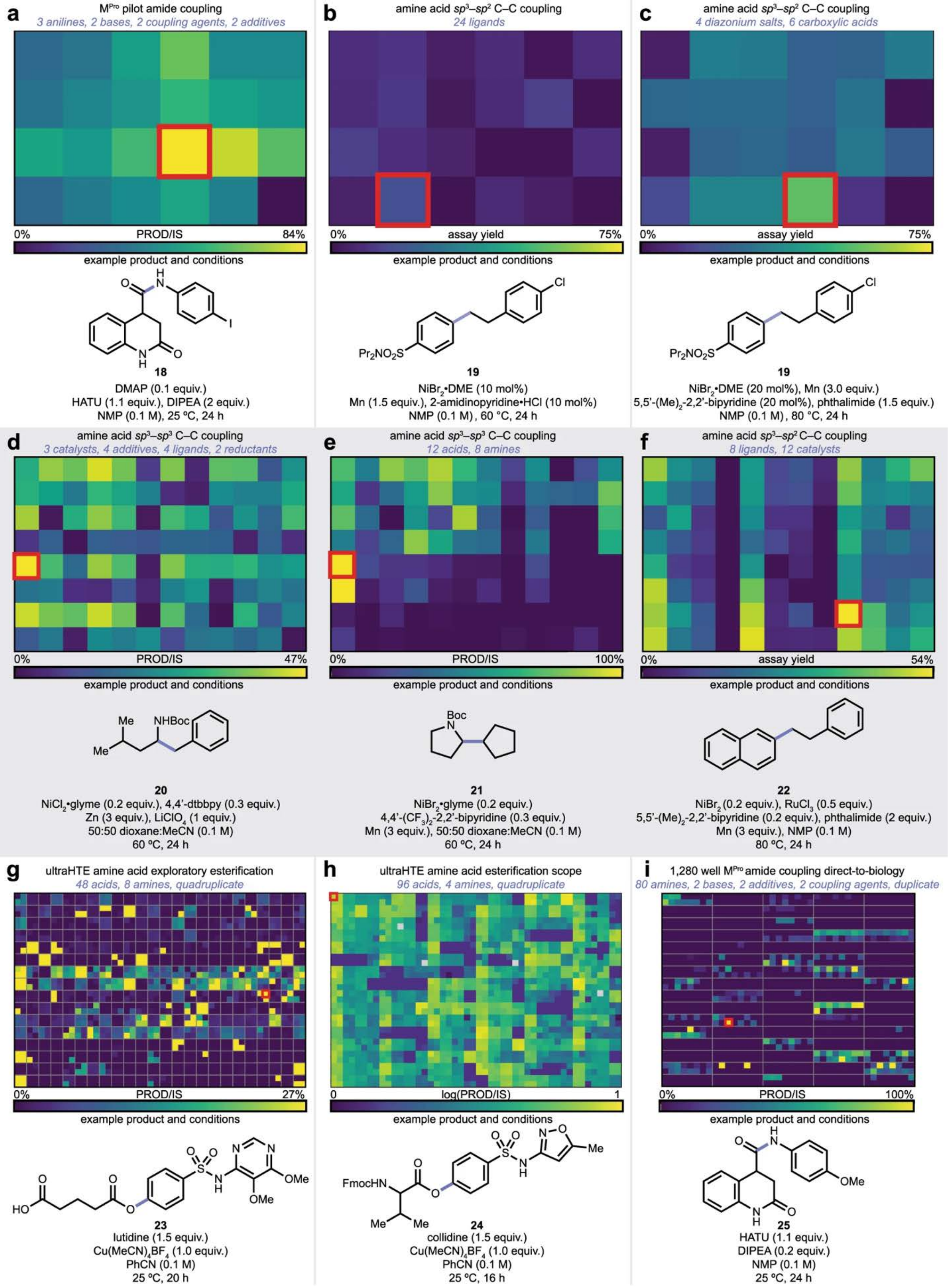
As all reaction arrays are stored in a centralized database, bulk analyses of multiple reaction arrays can be performed. phactor™ provides an exposed API that can be used to develop interfaces to other robots, assays, and software. Examples of various integrations and code infrastructures are shown in the Discussion section of the Supplementary Information. Furthermore, we are currently developing additional integrations with AI and ML models to generate and evaluate reaction arrays,[137] and to discover additional amine acid couplings via automated mining of relevant literature.[187-191] We hope that phactor™'s ease of use provides increased accessibility to HTE and HTE data in the chemistry community. Registration free and non-commercial use of phactor™ in 24- and 96-well formats is available through https://phactor.cernaklab.com/.


a
[catalysts]
[ligands]
[acids / bases]
[oxidants / reductants]
[solvents]
+
[nucleophile]
[electrophile]
[product]
b
analytics interface
reagent selection and plate design
experimental data collection
data storage
c
phactor
manual screen
robot
UPLC-MS
analysis
d
reaction minaturization
parallel reactor block
e
1,536 wellplate
robot

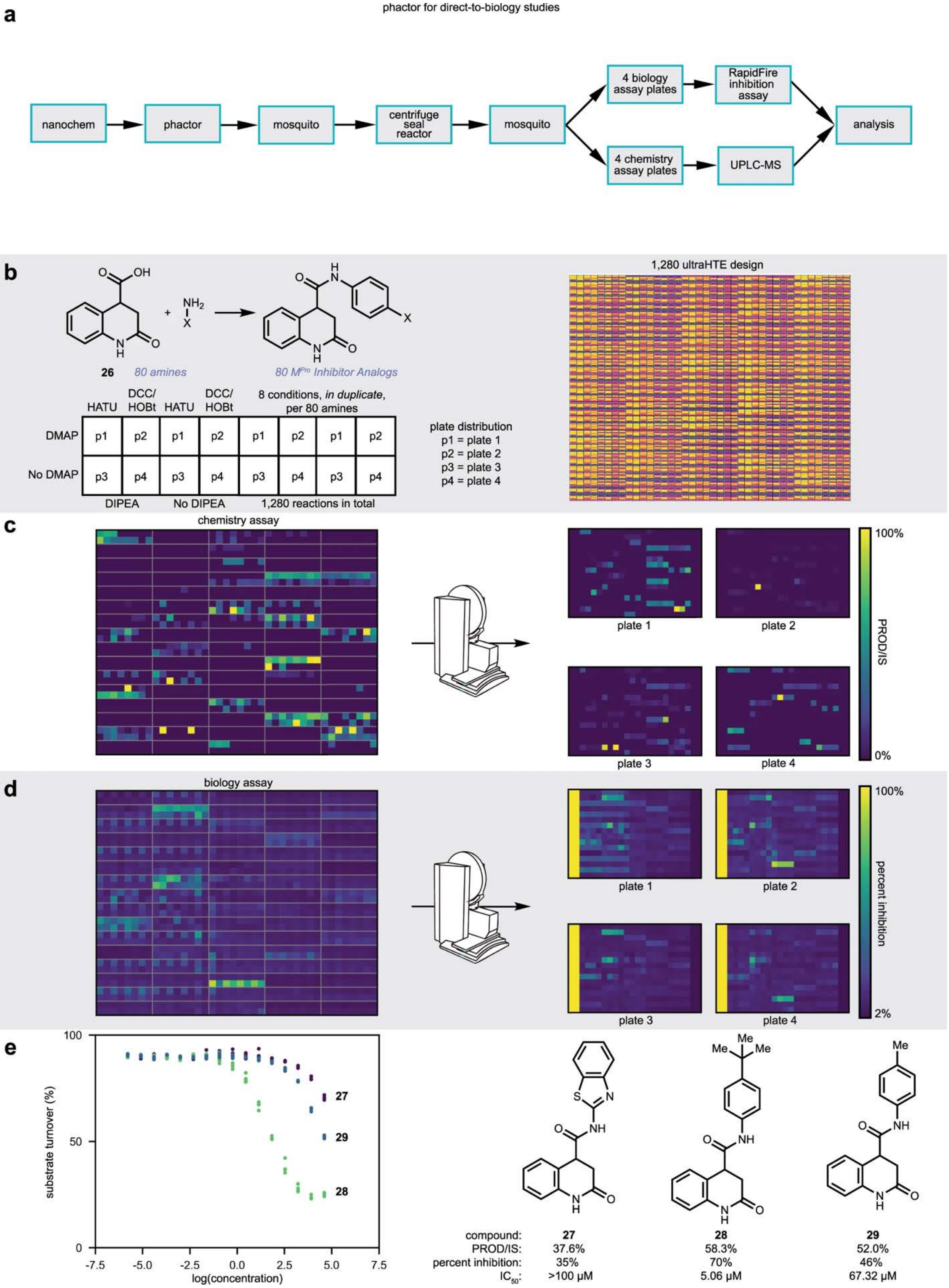
38-Fig. 1 | Overview of the phactor™ software. a Anatomy of a reaction as encoded by phactor™. **b** High-level software workflow of phactor™. Reaction arrays are designed from chemical inventories and reaction informatics. Resultant data is stored in delimited text (CSV) or in a relational database (SQLite3). phactor™ can convert results to Open Reaction Database (ORD)[149] and Chemical Description Language (XDL)[178] and is readily compatible with optimization programs such as EDBO+[176] and LabMate.ML[177]. **c** Workflow of phactor™. Once the reaction array is designed, phactor™ provides human-readable or machine instructions to execute the dosing manually or robotically (UPLC ultra-performance liquid chromatography). **d** phactor™ supports custom volumes allowing for reaction arrays to be performed at any scale. At a minimum, the hardware needed to execute a reaction array is an autopipette and an array reactor block. **e** phactor™ facilitates the design and execution of ultraHTE in 1536 wellplates.


a
Me
CO2H
+
N2BF4
Me
4 ligands
6 catalysts
MeCN, N2
60 ºC, 18 h
Me
O
O
Me
1
2
3
reaction array design
pyri-dine
dtbbpy
PPh3
xant-phos
CuI AgNO3
CuI
Pd(OAc)2 AgNO3
Pd(dba)2
Pd(OAc)2 AgNO3
NiCl2•dme
reaction array result (UPLC)
18.5
16.5
3.2
0.0
0.0
0.0
6.2
0.0
0.0
0.0
0.0
0.0
4.5
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0%
assay yield
18.5%
Selected Well: A1
pyridine, 45 mol%
AgNO3, 30 mol%
CuI, 30 mol%
1
+
2
(1.2 equiv.)
MeCN (0.1 M), N2
60 ºC, 18 h
3
Output Value: 18.5%
b
HO
Br
Br
N
H
Me
PhS
O
O
OEt
5
4 copper catalysts
6 ligands/additives
Cs2CO3, 3 equiv
DMSO, 55 ºC, 18 h
O
OEt
HO
SPh
Br
N
Me
4
6
reaction array design
CuI
CuBr
CuOTf
Cu-(OAc)2
none none
none MgSO4
L1 none
L1 MgSO4
L2 none
L2 MgSO4
reaction array result (UPLC)
12
61
82
81
65
53
72
62
83
29
23
9
68
62
73
69
70
70
44
55
54
28
10
45
0%
PROD/IS
83%
Selected Well: B3
2-(1H-tetrazol-1-yl)acetic acid, 40 mol%
CsCO3, 3 equiv.
no additive
CuBr, 20 mol%
4
+
5
(1.2 equiv.)
DMSO (0.1 M)
60 ºC, 18 h
6
Output Value: 83%
c
O
O
Me
TMSO
O
O
R
9 (R = Me) or 10 (R = OMe)
L3, Pd2dba3
PhMe (0.5 M), rt, 24 h
3 Pd/ligand ratios
2 bases, 2 electrophiles
2 nucleophiles
R
O
O
7 or 8
11 (R=H) or 12 (R=Me)
reaction array design
7 + 9
7 + 10
8 + 9
8 + 10
1:1 Pd:L No K2CO3
1:1 Pd:L K2CO3
2:1 Pd:L No K2CO3
2:1 Pd:L K2CO3
1:2 Pd:L No K2CO3
1:2 Pd:L K2CO3
reaction array result (UPLC ratio)
7 or 8
α1 coupling
α2 coupling
11 or 12
Selected Well: D3
Pd2dba3, 10 mol%/L3, 5 mol%
no K2CO3
8
+
10
PhMe (0.5 M)
rt, 24 h
11
Output Value: 42.3%

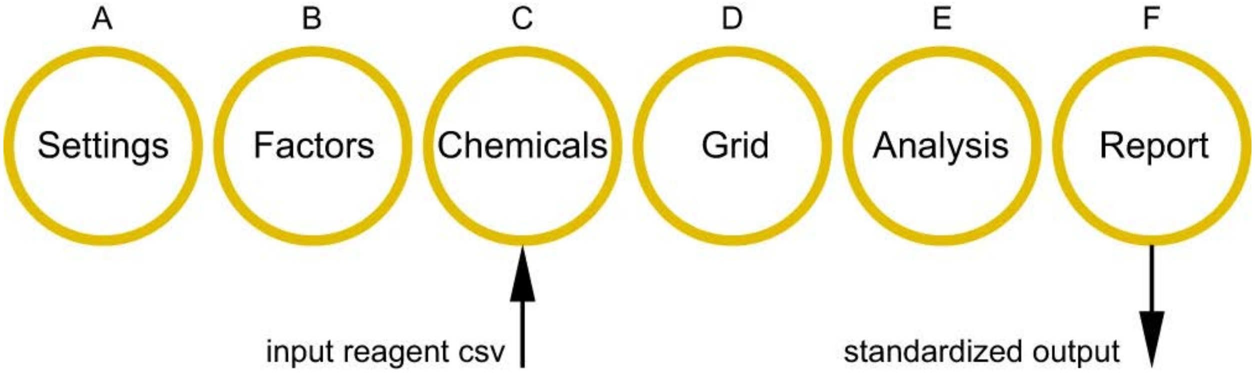
39-Fig. 2 | Reaction arrays executed with the phactor™ software. The reaction array design and results are shown here as displayed on phactor™. Colour bars adjacent to compound numbers correspond to the colour bars in the reaction array design grid generated by phactor™. Product/internal standard ratios are calculated using the observed UV-derived peak area, while assay yields account for differences in product absorptivity by calibrating to authentic samples of products. **a** Preliminary esterification hit leading to publication[180]. **b** Optimized oxidative indolization conditions towards the synthesis of umifenovir[54]. **c** Allylation catalyst/ligand concentration ratio and base reaction array analysed by conversion and selectivity.


a
M^Pro pilot amide coupling
3 anilines, 2 bases, 2 coupling agents, 2 additives
0%
PROD/IS
84%
example product and conditions
18
DMAP (0.1 equiv.)
HATU (1.1 equiv.), DIPEA (2 equiv.)
NMP (0.1 M), 25 ºC, 24 h
b
amine acid sp³–sp² C–C coupling
24 ligands
0%
assay yield
75%
example product and conditions
Pr₂NO₂S
Cl
19
NiBr₂•DME (10 mol%)
Mn (1.5 equiv.), 2-amidinopyridine•HCl (10 mol%)
NMP (0.1 M), 60 °C, 24 h
c
amine acid sp³–sp² C–C coupling
4 diazonium salts, 6 carboxylic acids
0%
assay yield
75%
example product and conditions
Pr₂NO₂S
Cl
19
NiBr₂•DME (20 mol%), Mn (3.0 equiv.)
5,5'-(Me)₂-2,2'-bipyridine (20 mol%), phthalimide (1.5 equiv.)
NMP (0.1 M), 80 °C, 24 h
d
amine acid sp³–sp³ C–C coupling
3 catalysts, 4 additives, 4 ligands, 2 reductants
0%
PROD/IS
47%
example product and conditions
Me
NHBoc
Me
20
NiCl₂•glyme (0.2 equiv.), 4,4'-dtbbpy (0.3 equiv.)
Zn (3 equiv.), LiClO₄ (1 equiv.)
50:50 dioxane:MeCN (0.1 M)
60 ºC, 24 h
e
amine acid sp³–sp³ C–C coupling
12 acids, 8 amines
0%
PROD/IS
100%
example product and conditions
Boc
N
21
NiBr₂•glyme (0.2 equiv.)
4,4'-(CF₃)₂-2,2'-bipyridine (0.3 equiv.)
Mn (3 equiv.), 50:50 dioxane:MeCN (0.1 M)
60 ºC, 24 h
f
amine acid sp³–sp² C–C coupling
8 ligands, 12 catalysts
0%
assay yield
54%
example product and conditions
22
NiBr₂ (0.2 equiv.), RuCl₃ (0.5 equiv.)
5,5'-(Me)₂-2,2'-bipyridine (0.2 equiv.), phthalimide (2 equiv.)
Mn (3 equiv.), NMP (0.1 M)
80 ºC, 24 h
g
ultraHTE amine acid exploratory esterification
48 acids, 8 amines, quadruplicate
0%
PROD/IS
27%
example product and conditions
HO
OMe
OMe
23
lutidine (1.5 equiv.)
Cu(MeCN)₄BF₄ (1.0 equiv.)
PhCN (0.1 M)
25 ºC, 20 h
h
ultraHTE amine acid esterification scope
96 acids, 4 amines, quadruplicate
0
log(PROD/IS)
1
example product and conditions
FmocHN
Me
Me
Me
24
collidine (1.5 equiv.)
Cu(MeCN)₄BF₄ (1.0 equiv.)
PhCN (0.1 M)
25 ºC, 16 h
i
1,280 well M^Pro amide coupling direct-to-biology
80 amines, 2 bases, 2 additives, 2 coupling agents, duplicate
0%
PROD/IS
100%
example product and conditions
OMe
25
HATU (1.1 equiv.)
DIPEA (0.2 equiv.)
NMP (0.1 M)
25 ºC, 24 h

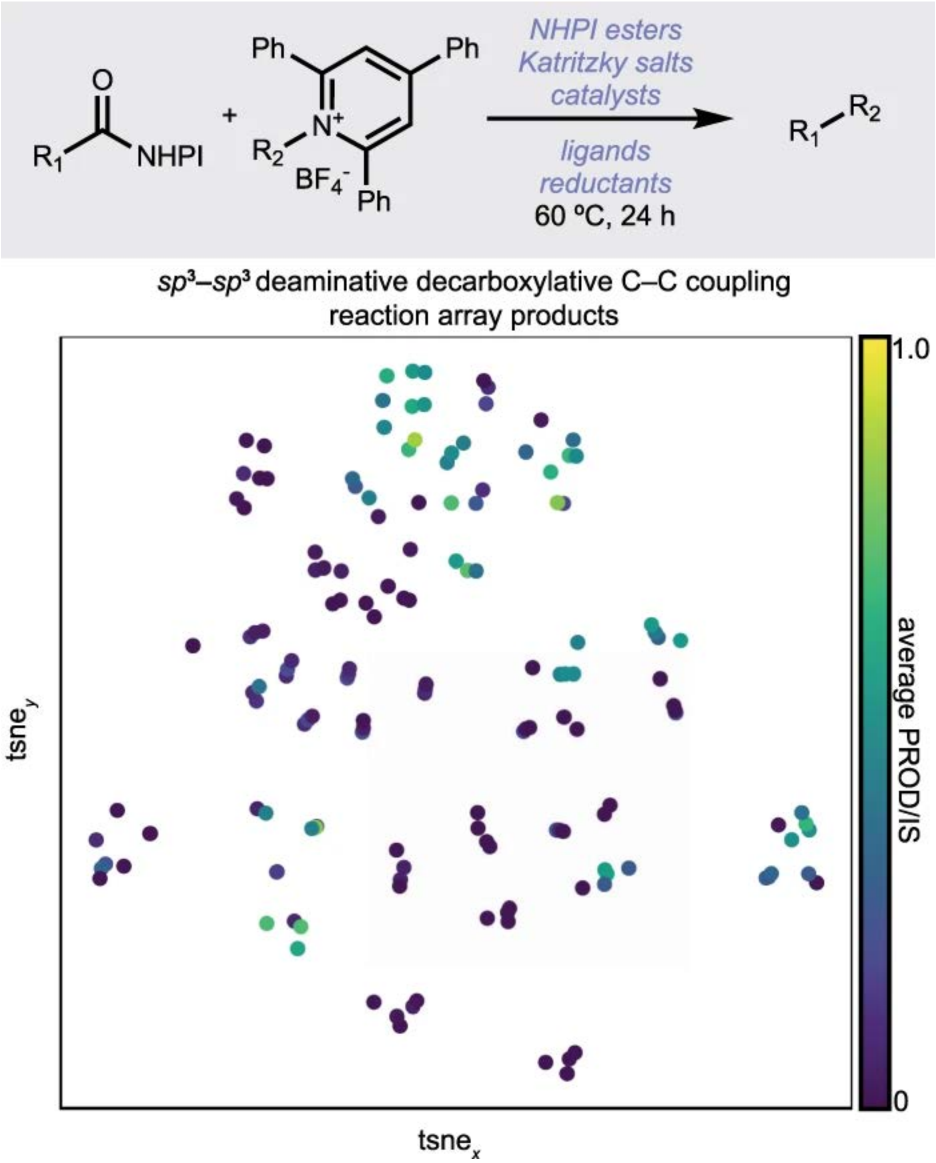
40-Fig. 3 | phactor™ has been utilized in a variety of synthetic campaigns. a–**i** Chemistries discovered via reaction arrays designed with phactor™. All input and output files used to produce reaction arrays (**a**–**f**) are provided via an online repository in addition to all compiled HTE results in a machine-readable format. Reaction schemes can be found in the "Selected screening examples" section of the Supplementary Information.


a
phactor for direct-to-biology studies
nanochem
phactor
mosquito
centrifuge seal reactor
mosquito
4 biology assay plates
RapidFire inhibition assay
4 chemistry assay plates
UPLC-MS
analysis
b
26
80 amines
80 M^Pro Inhibitor Analogs
HATU
DCC/HOBt
HATU
DCC/HOBt
8 conditions, in duplicate, per 80 amines
DMAP
No DMAP
p1 p2 p1 p2 p1 p2 p1 p2
p3 p4 p3 p4 p3 p4 p3 p4
DIPEA
No DIPEA
1,280 reactions in total
plate distribution
p1 = plate 1
p2 = plate 2
p3 = plate 3
p4 = plate 4
1,280 ultraHTE design
c
chemistry assay
plate 1
plate 2
plate 3
plate 4
100%
PROD/IS
0%
d
biology assay
plate 1
plate 2
plate 3
plate 4
100%
percent inhibition
2%
e
substrate turnover (%)
100
50
0
-7.5 -5.0 -2.5 0.0 2.5 5.0 7.5
log(concentration)
27
29
28
Me
Me Me
compound: 27 28 29
PROD/IS: 37.6% 58.3% 52.0%
percent inhibition: 35% 70% 46%
IC50: >100 μM 5.06 μM 67.32 μM

41-Fig. 4 | phactor™ facilitates ultraHTE direct-to-biology campaigns. **a** Event workflow for performing ultraHTE using phactor™ and a Mosquito robot. **b** Design of 1280 well amide coupling plate. 80 amines were selected to react with carboxylic acid **26**. Eight conditions were run in duplicate for each amine. **c** Results of the amide coupling are shown as a product/internal standard integration ratio from a 2-min LCMS injection of each well. The Mosquito robot is utilized to split the size 1536 plate into four sizes 384 plates for LCMS and bioassay analysis. **d** Percent inhibition of SARS-CoV-2 M^{Pro} when treated with a sample of the reaction mixture from the corresponding well. The 1280 plate is visually recreated. **e** IC_{50} curves for three scaled-up compounds chosen from the reaction array. Compounds **27**–**29** display a range of assay and inhibitory responses.



42-Fig. 5 | The six stages in the phactor™ workflow. Each stage is progressed sequentially. With an input reagent CSV, reaction arrays can be designed in seconds. Once the experiment has been executed, a standardized output can be downloaded on the report stage.



43-Fig. 6 | phactor™ enables rapid machine-learning analysis of multiple reaction arrays in tandem. Standardized output files can be rapidly merged to create massive datasets. Shown is a tSNE (t-distributed stochastic neighbour embedding) of all products made in the decarboxylative–deaminative sp^3–sp^3 C–C coupling detailed in ref. [181], coloured by average product/internal standard.

Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis

This review covers several important advancements that have led to contemporary ultrahigh throughput experimentation. The incorporation of data science has sparked a revolution in organic chemistry, enabling accurate prediction of reaction outcomes, computational planning of retrosynthetic routes, and design of sophisticated molecules. However, the scarcity of available reaction data poses a challenge for building predictive models. Literature mining provides some data, but it is often reported in text format, difficult for computers to process, and biased toward "productive" reactions. High-throughput experimentation (HTE) has emerged as a solution, offering systematic and consistent data points, including negative outcomes, in a machine-readable format. HTE, particularly ultrahigh-throughput experimentation (ultraHTE), using microliter-scale droplets in 1536-well microtiter plates, has generated a wealth of reaction data. This data has been instrumental in identifying optimal substrate-catalyst pairs for novel cross-coupling reactions. UltraHTE's compatibility with diverse solvents, catalysts, and biochemical assays, along with the development of supporting software, has paved the way for extensive exploration of reaction space and a holistic systems chemistry approach. This work is feature in *Accounts of Chemical Research*.[108]

Abstract

The incorporation of data science is revolutionizing organic chemistry. It is becoming increasingly possible to predict reaction outcomes with accuracy, computationally plan new retrosynthetic routes to complex molecules, and design molecules with sophisticated functions. Critical to these developments has been statistical analysis of reaction data, for instance with machine learning, yet there is very little reaction data available upon which to build models. Reaction data can be mined from the literature, but experimental data tends to be reported in a text format that is difficult for computers to read. Compounding the issue, literature data are heavily biased toward "productive" reactions, and few "negative" reaction data points are reported even though they are critical for training of statistical models. High-throughput experimentation (HTE) has evolved over the past few decades as a tool for experimental reaction development. The beauty of HTE is that reactions are run in a systematic format, so data points are internally consistent, the reaction data are reported whether the desired product is observed or not, and automation may reduce the occurrence of false positive or negative data points. Additionally, experimental workflows for HTE lead to datasets with reaction metadata that are captured in a machine-readable format. We believe that HTE will play an increasingly important role in the data revolution of chemical synthesis. This Account details the miniaturization of synthetic chemistry culminating in ultrahigh-throughput experimentation (ultraHTE), wherein reactions are run in ∼1 μL droplets inside of 1536-well microtiter plates to minimize the use of starting materials while maximizing the output of experimental information. The performance of ultraHTE in 1536-well microtiter plates has led to an explosion of available reaction data, which have been used to identify specific substrate–catalyst pairs for maximal efficiency in novel cross-coupling reactions. The first iteration of ultraHTE focused on the use of dimethyl sulfoxide (DMSO) as a high-boiling solvent that is compatible with the plastics most commonly used in consumable well plates, which generated homogeneous reaction mixtures that are perfect for use with nanoliter-dosing liquid handling robotics. In this way, DMSO enabled diverse reagents to be arrayed in ∼1 µL droplets. Reactions were run at room temperature with no agitation and could be scaled up from the ∼0.05 mg reaction scale to the 1 g scale. Engineering enhancements enabled the use of ultraHTE with diverse and semivolatile solvents, photoredox catalysis, heating, and acoustic agitation. A main driver in the development of ultraHTE was the recognition of the opportunity for a direct merger between miniaturized reactions and biochemical assays. Indeed, a strategy was developed to feed ultraHTE reaction mixtures directly to a mass-spectrometry-based affinity selection bioassay. Thus, micrograms of starting materials could be used in the synthesis and direct biochemical testing of drug-like molecules. Reactions were performed at a reactant concentration of ∼0.1 M in an inert atmosphere, enabling even challenging transition-metal-catalyzed reactions to be used. Software to enable the workflow was developed. We recently initiated the mapping of reaction space, dreaming of a future where transformations, reaction conditions, structure, properties and function are studied in a systems chemistry approach.

Introduction

Moore's law projected that computer power would accelerate because of the miniaturization of silicon transistors.[192] Through miniaturization, more information can be generated and processed per space unit: in terms of hardware, such as a computer chip, this leads to parallel processing. Nearly every advanced technology reaches a state where miniaturization and parallelization become necessary for advancement. Synthetic chemistry is now entering an era where new tools of data science are increasingly applicable[114] yet there are not enough synthetic chemistry data available to satiate modern computing power and scientific creativity. Miniaturization and parallel processing of chemical synthesis represent a strategy to generate more data. This Account details efforts to miniaturize synthetic chemistry experiments to the nanoscale and the implications of such a development for establishing an interface of chemical synthesis and data science. A significant portion of this work was performed at Merck & Co., Inc. Colleagues there have already reviewed the strategic application of HTE in pharmaceutical synthesis.[193] There have been a number of other reviews of HTE as well.[113, 194-197] This Account specifically covers reaction miniaturization to generate more reaction data per experiment, enabling a systems-level evaluation of synthesis. As the merger of chemical synthesis and data science enters the mainstream,[198] new tools to generate large chemical synthesis datasets will be needed. We advocate for miniaturized HTE as an ideal tactic to enable a systems chemistry approach to chemical synthesis.

The journey to miniaturize chemical synthesis to the nanoscale began during work on a diacylglycerol acyl transferase 1 (DGAT1) lead identification effort.[110] A series of 2-piperidinylpyridines, such as **3** (44-Fig. 1a), emerged as potent DGAT1 inhibitors and were prioritized for further investigation. However, challenges in the synthesis of these compounds significantly impeded the advancement of the lead series, with the key S_NAr reaction used to make them typically giving <10% (and often 0%) yields of the desired products. The challenges of producing new analogues in this series and the desire to produce libraries of analogues using parallel medicinal chemistry, where robust reactions with broad substrate tolerance are needed,[199] made it clear that improved reaction conditions would be required. Here we turned to high-throughput experimentation (HTE) to survey reaction performance systematically. Starting materials were precious, so it was decided to explore miniaturized HTE in 8 mm × 13 mm glass shell microvials charged with a 1.67 mm × 2.01 mm parylene coated stir dowel. Identifying mild reaction conditions to perform the key S_NAr reaction involved admixing electrophile **1** with nucleophile **2** in an array of four solvents and six bases to produce a library of 24 reactions (44-Fig. 1b). Each solvent–base combination was studied in an individual vial, and the 24 vials were housed in an aluminum parallel reactor block that is now commercially available. From this simple experiment we identified a variety of promising reaction conditions, including those using *N*-methylpyrrolidinone (NMP) or cyclopentyl methyl ether (CPME) as the solvent and *N*,*N*-diisopropylethylamine (DIPEA) or sodium bicarbonate as the base. One of the benefits of HTE is the ability to systematically interrogate multiple reaction parameters at once. Our survey of bases and solvents had been performed on a single substrate pair (**1** + **2**), but since a primary objective was to identify reaction conditions that were ideal for library synthesis, we elected to repeat the four best sets of conditions from this survey—NMP with $NaHCO_3$, NMP with DIPEA, CPME with $NaHCO_3$, and CPME with DIPEA—using 12 electrophile–nucleophile pairs. Surprisingly, while the best conditions for coupling **1** to **2** were CPME with DIPEA, these specific conditions were the least productive, in terms of average conversion to product, across all 12 electrophile–nucleophile pairs interrogated. This holistic analysis revealed NMP with $NaHCO_3$ to be the most successful reaction conditions across diverse substrate pairs, and these conditions were used to advance the chemical series. Indeed, multiple compounds were prepared via the optimized S_NAr reaction conditions as single reactions and in parallel synthesis campaigns using $NaHCO_3$ as the base in NMP as the solvent. The reaction conditions were incredibly robust and became widely adopted by the team, leading to many promising compounds that collectively advanced the 2-piperidinylpyridine series. For instance, one parallel synthesis campaign culminated in spirocycle **4**,[110] while another library array revealed biaryl electrophiles as viable precursors, resulting in **5**.[199] These compounds were critical stepping stones in the development of advanced compounds **6**[200] and **7** (44-Fig. 1d).[201] The registration of compounds bearing a 2-aminopyridine moiety accelerated considerably following the identification of NMP and $NaHCO_3$ as ideal promoters for the reaction (44-Fig. 1c). Indeed, an advanced compound reported from the series was discovered using an S_NAr reaction promoted by the NMP and $NaHCO_3$ conditions, highlighting the impact that a simple HTE campaign can have on a drug discovery program.

Our successful application of HTE in the hunt for DGAT1 inhibitors cemented the use of the technology on the frontlines of medicinal chemistry and emboldened us to explore other increasingly challenging chemistries. Around this time, an effort to perform late-stage functionalization (LSF) was launched to investigate the role of specifically targeting

C–H bonds for diversification in medicinal chemistry.[75] One such application from our studies was the development of ligand–solvent arrays for iridium-catalyzed borylation of aryl C–H bonds,[202-204] as shown in 45-Fig. 2 to convert **8** to **9**, **10**, and **11**.[205] Six diverse nitrogen- or phosphorus-containing ligands were interrogated in combination with tetrahydrofuran (THF), hexanes, dioxane, or NMP as the solvent. From these studies we identified conditions for productive bisborylation of C3 and C6 to provide **10**. Subsequent HTE studies surveying various acids and protic additives identified mild conditions for the selective deborylation of **10** to give **11** by heating in CPME with methanol as a mild proton source. These results are a single case study, but we routinely used HTE to enable LSF. This case study is just one example of how miniaturized HTE arrays were key to enabling the exploration of C–H functionalization on complex molecules, where typically substrate amounts were limited and product decomposition and undesired side reactions were common.

Our application of HTE in medicinal chemistry was consistently challenged by the availability of precious starting materials. On the front lines of drug discovery, substrates have typically been optimized, or partially optimized, toward multiple desired objectives such as desirable solubility, permeability, metabolic stability, and efficacy; therefore, these substrates tend to be structurally complex and typically can be accessed only by multistep synthesis. For this reason, in the exploration of new pockets of chemical space, there is rarely a significant amount of physical material with which to perform reaction condition studies. However, the more complex the molecule’s structure is, the more likely that reaction condition studies will be required to obtain meaningful amounts of a target compound. To address this issue, we became interested in miniaturizing synthesis beyond the 13 mm × 48 mm and 8 mm × 13 mm glass shell microvials that we were currently using for HTE. To miniaturize further, we elected to study the tools of the biochemical sciences, where high-throughput screening (HTS) and ultrahigh-throughput screening (uHTS) on fleeting amounts of chemical or biological reagents are routinely performed.

Reactions on Nanomole Scale

In 1950, Gyula Takátsy invented the 96-well microtiter plate[206] to miniaturize the amount of precious blood samples that would be required for clinical investigations. In the decades following, an entire industry of automated miniaturized experimentation had evolved to support advanced biochemical HTS and uHTS studies. While the 96-well microtiter plate was the workhorse of benchtop biochemical experimentation, higher-density automated liquid handling was easily achievable in 384-, 1536-, and 3456-well microtiter plates. We considered that synthetic chemistry could be performed using the tools of the biochemical trade, where miniaturization and parallel processing are the status quo. However, using biochemical tools required overcoming several engineering challenges. For instance, miniaturized biochemical experimentation typically takes place in plastic reaction vessels, in water, with reagents that dissolve well in water, at high dilution (nanomolar to micromolar reagent concentration) and at room temperature or 37 °C. In contrast, chemical synthesis typically takes place in glass reaction vessels, in anhydrous organic solvents that dissolve many plastics, with reagents that are frequently insoluble in the reaction medium, at low dilution (molar reagent concentration) across a broad range of temperatures (−78 to 200 °C). These differences in the experimental requirements of biochemistry versus synthetic chemistry presented major hurdles to the miniaturization of chemical synthesis. Nonetheless, it was hard to ignore the opportunity presented by the six decades of engineering and infrastructure that had gone into developing miniaturization and automation for the biochemical sciences, where it is possible to perform millions of biochemical reactions in a single day.

Miniaturizing chemical synthesis involved considerable experimentation. From the outset, our objective was to miniaturize synthetic chemistry experiments in such a way that they still approximated classic synthesis experiments. We did not want to perform reactions at high dilution or open to air since that would preclude the use of most transition-metal-catalyzed reactions. This was particularly important to us because our reaction miniaturization efforts were concurrent with our studies in LSF, where air-sensitive transition metal catalysts were commonly used. Many classic organic reactions like amide coupling, reductive amination, or sulfonamide formation with amines and sulfonyl chlorides are bread-and-butter reactions in medicinal chemistry.[30, 207, 208] These reaction classes can often work very well in an aerobic atmosphere, at high reaction dilution, and in a variety of solvents, including water. Reactions using transition metal catalysts, in contrast, often require an anaerobic environment and a high reaction concentration and may be heavily influenced by the choice of solvent. These reaction attributes presented engineering challenges to the miniaturization of chemical synthesis. For instance, our earliest studies were performed in microtiter plates made of polystyrene, which is a common plastic used in liquid handling consumables but has poor compatibility with organic solvents. Many initial experiments ended prematurely because the reaction vessel dissolved into the organic reaction medium being used. After some experimentation, it became clear that cyclic octane copolymer (COC) was a more robust material for organic chemistry experimentation. Luckily, 384- and 1536-well microtiter plates made of COC were commercially available. Subsequently, polypropylene was used because it is relatively inexpensive and robust toward many solvents. Glass 384- and 1536-well microtiter plates were also engineered to maximize chemical compatibility, but the high cost and fragility of these glass plates has to date been a hurdle to their wide adoption in reaction miniaturization. The lack of commercial availability of affordable 1536-well reaction vessels with high chemical compatibility and thermal stability remains a challenge for the field.

Beyond the reaction vessel, we invested considerable effort studying reaction solvents for miniaturization. An analysis of 240 years of chemical synthesis data revealed that dichloromethane, THF, and methanol are the three most frequently selected reaction solvents in chemical synthesis.[209] These solvents are challenging to use in miniaturization because they are too volatile. Miniaturized chemistry can in principle be performed with a diversity of solvents, even those with relatively low boiling points, but it becomes necessary to run reactions at high dilution to counteract evaporation.[38] The kinetics of many important reactions preclude their use at high dilution. Our target reaction volume was ~1 μL, such that 0.05 mg of a typical drug-like substrate[210] (500 g/mol) would produce a 0.1 M reaction concentration, which is a typical reaction concentration used in chemical synthesis. To achieve a reaction volume of ~1 µL, it was necessary to use high-boiling solvents, and we gravitated to the use of dimethyl sulfoxide (DMSO) and NMP. These solvents have been used in a diversity of chemistries and additionally are quite good at dissolving complex molecules, ensuring that reactions are homogeneous. Reaction homogeneity was a desirable experimental trait since we did not stir or agitate reaction mixtures at the outset. Homogenous liquids are also much easier to manipulate with liquid handling automation.

In addition to the engineering requirements of the reaction vessel and the solvent, the selection of the ideal automation technology required consideration. We aspired to execute chemistry inside an inert-atmosphere glovebox so that air-sensitive catalysts could be used. After evaluating a number of liquid handling systems, we arrived at the SPT Labtech mosquito as a preferred system for our needs. The small footprint of this robot enabled it to be easily enclosed inside a typical glovebox, allowing air-sensitive chemistries to be explored. As opposed to common air-displacement liquid handlers, the mosquito liquid dosing mechanism was based on positive displacement with a plunger in each pipet tip, which was distinct from other liquid handling technologies. The use of this positive-displacement liquid handling mechanism turned out to be highly enabling because multiple reagents could be loaded into each consumable tip of the liquid handler (46-Fig. 3b) and mixed by drawing the liquid up and down upon dosing into the receiving reaction plate. This dosing protocol ensured an even distribution of reagents throughout the reaction droplet. Additionally, the positive-displacement mechanism is based on the distance the plunger travels in the pipet tip and thus is generally agnostic of solution or suspension viscosity and heterogeneity, allowing us to handle diverse reagent or reaction mixtures, even those that did not dissolve.

Having navigated the challenges of hardware engineering, we were ready to explore miniaturized chemical synthesis. The first reactions we explored in 1536-well microtiter plates were Buchwald–Hartwig couplings (46-Fig. 3a,c),[36] which are among the most important reactions used in synthesis for drug discovery.[30, 207, 211] Santanilla and Dreher had discovered the perfect set of reaction conditions for ambient-temperature cross-coupling in a homogeneous reaction medium, utilizing the third-generation (G3) Buchwald precatalysts and P_2Et as the base in DMSO.[212] The reagents and substrates dissolved well in DMSO, a high-boiling solvent that was compatible with the COC plastic plates and could be easily dosed in 1 µL reaction droplets on a mosquito liquid handler (46-Fig. 3b) in an inert-atmosphere glovebox. With this novel setup, we were able to run 1536 reactions and at once survey the performance of catalyst and additive combinations with diverse electrophiles and nucleophiles. This realized our first ability to perform ultrahigh-throughput experimentation (ultraHTE) for chemical synthesis. Initial studies revealed a significant impact of matching specific ligand–base selections to various nucleophile–electrophile substrate pairs. These effects were observable because the miniaturized chemical synthesis platform allowed the collection of a high density of reaction data from each experiment. The biggest concern that we had in the development of a miniaturized reaction platform was the ability to repeat the miniaturized experiments, which were performed on a reaction scale of ~50 µg, on a meaningful reaction scale of at least tens of milligrams. Because the reactions were generally homogeneous with reagent and substrate mixtures that were soluble in DMSO, we anticipated that reactions would be reproducible on a larger scale. Indeed, reaction performance on the nanomole scale correlated well with reaction performance on the millimole scale, and in the instance of coupling **12** to **13** with ^{t}BuXPhos Pd G3 and P_2Et, the reaction was repeated on the gram scale with product **14** being isolated in 79% yield (46-Fig. 3a).

Having demonstrated the feasibility of performing transition-metal-catalyzed coupling chemistry on the nanomole scale in 1536-well plates, several opportunities for platform expansion based on engineering and technological improvements emerged.[141] The platform was focused on the use of reactions that perform well in DMSO at room temperature. The use of DMSO has significant advantages in that (i) pharmaceutical compound collections are generally stored in DMSO and could potentially be used as a source of substrates for miniaturized synthesis and (ii) many bioassays are tolerant of DMSO, paving the way to the potential merger of ultraHTE with in-line bioassays. Nonetheless, the menu of popular synthetic chemistries that perform well in DMSO at room temperature is currently limited. To expand the menu of usable solvents, subsequent studies augmented the capability through further engineering. While DMSO and other high-boiling polar aprotic solvents were straightforward to use on the mosquito, the use of lower-boiling solvents such as 1,4-dioxane required complementary liquid handling equipment that could rapidly dose in parallel with a 384-channel pipetting head. With a diversity of solvents and chemistries came a diversity of reagent mixture solubility characteristics, wherein many reactions were suspensions instead of solutions. Here we designed customized well plate holders so that we could use specialized glass 1536-well microtiter plates and a resonant acoustic mixer to agitate the reaction plates. These customized well plate holders could also be heated and outfitted with glass bottoms so that the miniaturized reaction mixtures could be illuminated with light from high-intensity light-emitting diodes for the execution of metallophotoredox catalysis experiments. In these studies, analytical chemistry was vastly accelerated through the incorporation of MALDI, allowing a 1536-well microtiter plate to be analyzed in 8–11 min.

We next explored the use of the ultraHTE capability in the development of new cross-coupling strategies in partnership with the Buchwald lab.[213] Isolated oxidative addition complexes (OACs) had been demonstrated as viable coupling partners for protein modification.[214] We considered that OACs derived from pharmaceuticals might improve cross-coupling performance in a drug discovery setting. Pharmaceutically derived OACs proved to be excellent coupling partners with diverse nucleophiles. While the use of a stoichiometric equivalent of palladium is an atypical strategy, cost analyses showed that the complex pharmaceutical substrates we used were orders of magnitude more expensive than commercial palladium complexes, suggesting that on a discovery scale of dozens of milligrams the OAC approach can be highly economical if it improves the chances that synthesis will succeed on the first attempt. A library of 384 nucleophiles, such as amines, alcohols, and thiols, were subjected to coupling with rivaroxaban (**15**) (47-Fig. 4a) under catalytic conditions using 5 mol % *t*BuXPhos Pd G3 and P_2Et. Among 384 diverse amines, only 39 coupling products were observed by ultrahigh-performance liquid chromatography–mass spectrometry (UPLC–MS). Simultaneously, the analogous *t*BuXPhos OAC **16**, prepared from **15** and a stoichiometric equivalent of a palladium complex with ^{t}BuXPhos as a supporting ligand, was used as a coupling partner with the same 384 amines and P_2Et as the base. In this instance, 206 of the 384 diverse coupling products were observed by UPLC–MS (47-Fig. 4c), and key reactions were successfully reproduced on traditional reaction scale to yield dozens of milligrams each of **17**–**22** and many other congeners (47-Fig. 4b).

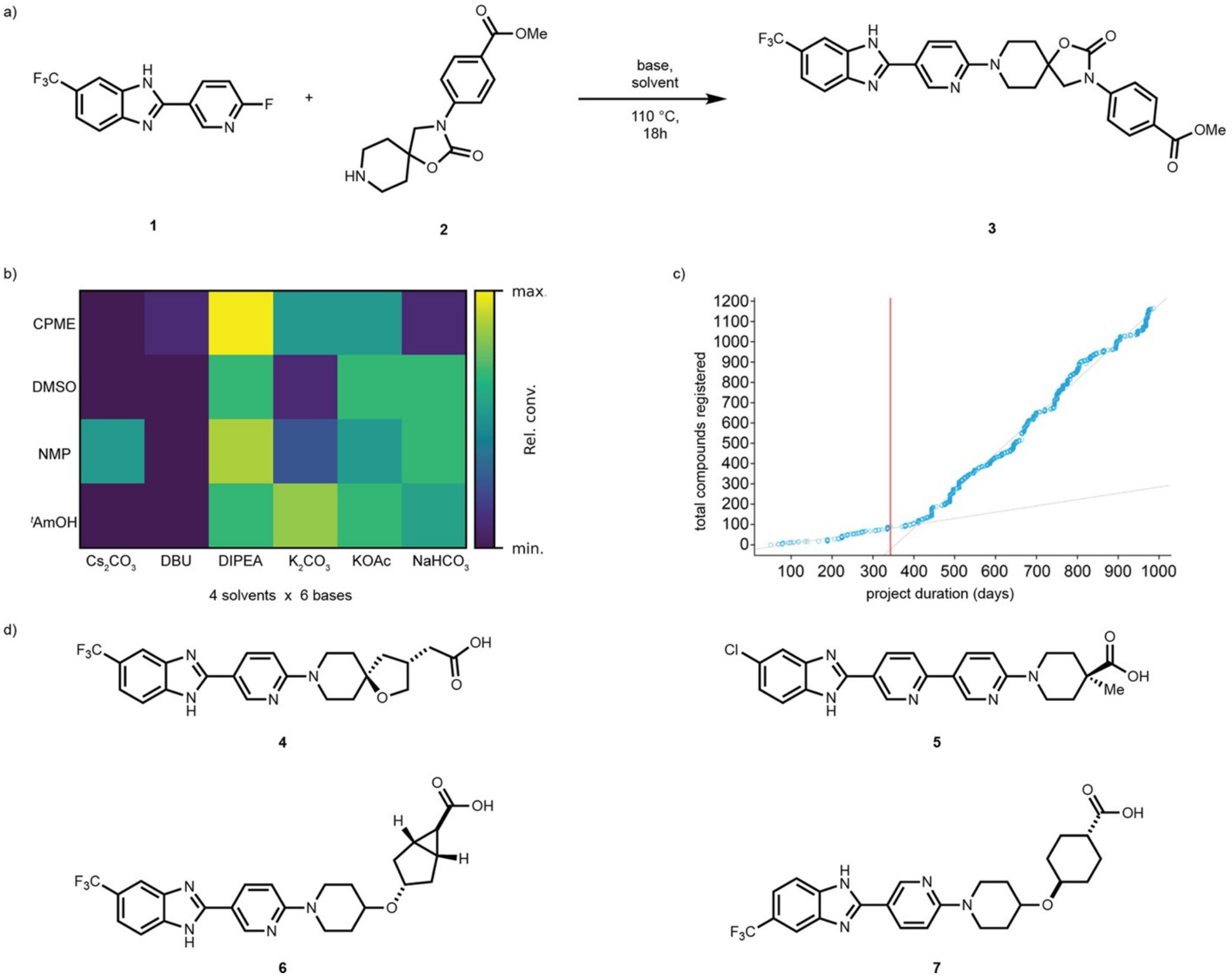
It quickly became clear that the miniaturized platform we had developed was ideal for merging chemical synthesis with in-line biochemical studies. The synthesis of small molecules and testing of their reversible affinity to protein targets of interest are two of the core operations of drug invention. In the interest of accelerating this process, we sought to merge our ultraHTE platform with an in-line high-throughput biochemical assay. Each reaction we ran contained ~0.05 mg of theoretical product, which is ample material to run multiple in vitro biochemical assays. Following consideration of various bioassay platforms, we gravitated to affinity-selection mass spectrometry (ASMS).[215-217] In ASMS, the affinity of a small molecule for a protein target of interest is registered through the observation of the small molecule's mass-to-charge ratio following incubation with the target protein, size-exclusion chromatography to separate unbound ligands from protein–ligand complexes, denaturation of the protein–ligand complexes, and mass spectrometric analysis. The technique is commonly used to identify protein ligands from mixtures of small molecules, suggesting that it would be possible to differentiate protein-binding compounds from other reaction components such as residual phosphine ligands, catalysts, and other chemical reagents as long as they had distinct molecular formulae. Indeed, we found that crude reaction mixtures from ultraHTE could be subjected to ASMS analysis with facile separation of compounds that had high affinity for specific proteins from those that had little or no affinity.[35] To increase throughput, reaction mixtures were pooled, while ensuring that no isobaric products were combined into the same well. A technique was developed to rank protein affinity by titrating the protein concentration while holding the concentration of the compound test mixture constant and monitoring the disappearance of the small molecule's mass spectral signal. Those compounds whose mass spectral signal persisted at the lowest protein concentration were expected to have the highest affinity to the protein target, and this was confirmed in control studies. Multiple libraries were compared wherein crude reaction products from nanoscale synthesis (~0.05 mg) were compared to chromatographically purified compounds prepared on a traditional reaction scale (~50 mg). A library of 384 nucleophiles were cross-coupled with simultaneous screening of four reaction conditions (48-Fig. 5). The most productive reactions were submitted directly to the bioassay. Excellent fidelity between the crude and purified products was observed by the affinity ranking experiments, confirming that the same conclusions could be reached even when reactions were run on a nanomole scale and subjected to bioassays without compound purification. As an example, checkpoint 1 (CHK1) kinase inhibitor **25** was readily differentiated from structurally related analogues **23** and **24**, which had no observable affinity for CHK1. The real estate on the well plate allows for positive and negative control experiments to be included. A series of known CHK1 inhibitors ranging in CHK1 activity (IC_{50}) from 19 to >10 000 nM were spiked into the ASMS sample pools, providing an estimation of the test compounds' CHK1 activities. This merger of chemical synthesis and biochemical testing significantly streamlines the drug discovery process and simultaneously provides useful data for learning about reaction optimization and protein affinity.

To facilitate HTE and ultraHTE research in our academic laboratories at the University of Michigan, we have developed phactor,[218] a full-stack software that allows bench chemists to design complex reaction arrays in just a few button clicks. In addition to experimental identification of winning reaction conditions, an advantage of HTE is the capture of systematically performed reaction data. In phactor, HTE and ultraHTE results are recorded in a machine-readable format for future statistical analyses. Designing experiments along with analyzing, visualizing, and reporting the results takes just a few button clicks in phactor, thus allowing the chemist to focus on idea generation, data interpretation, and reaction mechanism instead of experimental logistics. An example of phactor's reaction array design tool is shown in 49-Fig. 6.

Using phactor to facilitate our exploration of amine–acid coupling space, we explored an esterification coupling (50-Fig. 7a) as a direct complement to the venerable amide coupling where the N–H is replaced with an oxygen atom (**28**). Experimentally, we investigated the coupling of *o*-toluic acid (**26**) with diazonium salt **27** using transition metals [CuI, $Pd(OAc)_2$, $Pd(dba)_2$, $NiCl_2$•diglyme] and common ligands (pyridine, di-*tert*-buylbipyridine, triphenylphosphine, and xantphos), with silver nitrate included in selected wells as an additive. The heat map of UPLC–MS results indicates that CuI, $AgNO_3$, and pyridine were the best conditions (50-Fig. 7b).

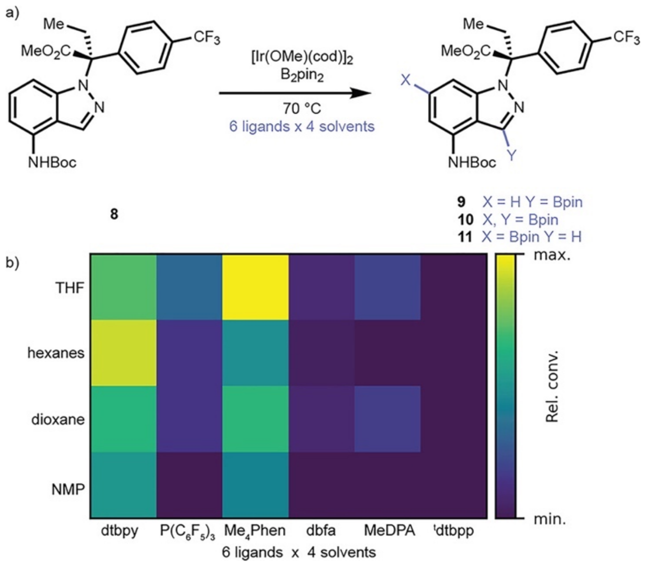
Since our preliminary report in this space,[36] there has been considerable innovation in the field of miniaturized HTE.[35, 38, 54, 110, 114, 141, 142, 197, 205, 213, 219-224] Identifying viable reaction conditions for any transformation is a multidimensional problem. When all of the combinations of catalysts, ligands, additives, and solvents that could be investigated are considered, the search space becomes nearly infinite. Through the rational design of arrays of reaction conditions, one can accelerate the navigation of the experimental search space. Exploration and exploitation are possible since larger amounts of systematically captured data are available. Looking forward, we see ultraHTE as a critical experimental technique to enable a systems chemistry approach (51-Fig. 8). We envision a future in which chemical transformations, reaction conditions, molecular structures, and physicochemical properties are directly linked to molecular function in the form of systems chemistry, much in the same way that systems biology links metabolites, genes, proteins, and pathways to disease. To lay a foundation for explorations in systems chemistry and to complement our ultraHTE experimental strategy, we have recently developed a system for mapping chemical transformations.[219] To date we have shown that amines and carboxylic acids can be coupled in hundreds of ways beyond the classic amide coupling and that each amine–acid coupling imprints a unique physicochemical footprint on the product. The use and rapid invention of new reactions to control molecular properties and function can be realized only through the rapid generation of reaction data, and thus, ultraHTE is uniquely suited to enable systems chemistry.

Much as Moore’s law describes the acceleration and advancement of information transfer in computer science on the basis of hardware miniaturization, the miniaturization and parallelization of chemical synthesis leads to more reaction information. As data science plays an increasingly prominent role in chemical synthesis,[198] the ability to generate larger amounts of reaction data will be needed. The development of ultraHTE allows a systems chemistry approach at the nexus of transformations, reaction conditions, chemical structure, properties, and ultimately molecular function.

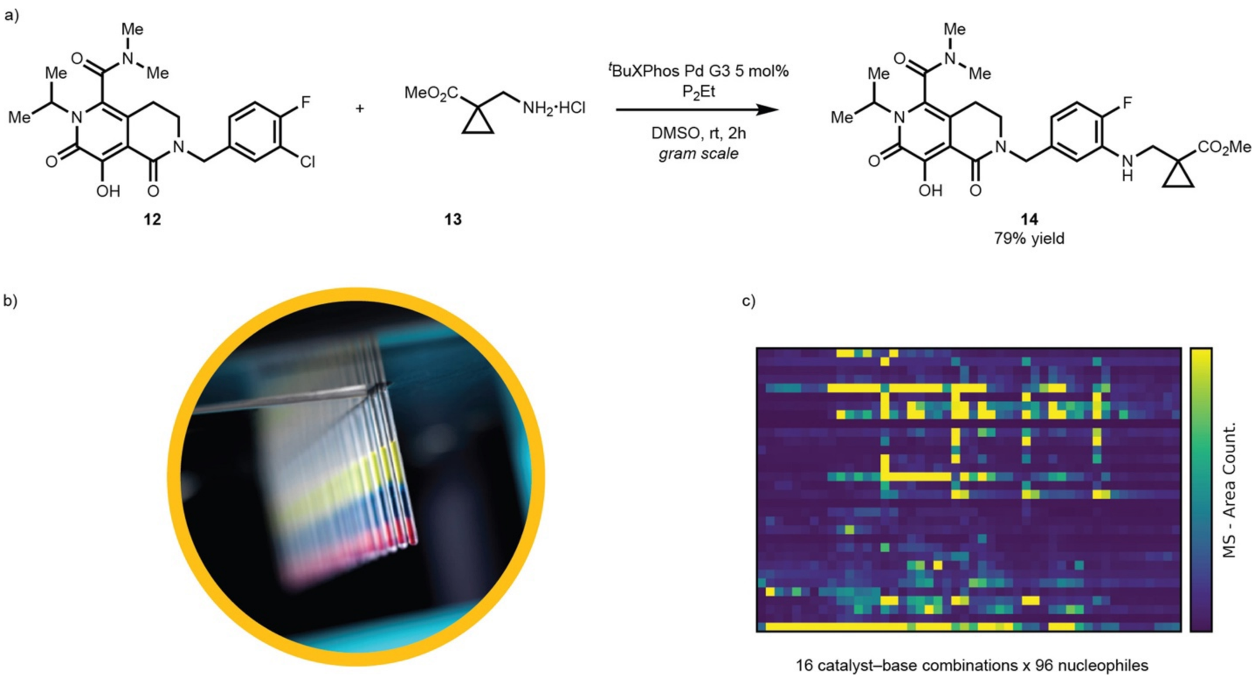


44-Fig. 1 | (a) HTE was used to develop reaction conditions for a key S_NAr coupling. (b) A first survey of four solvents and six bases revealed several productive conditions for the coupling of **1** and **2** to form **3**. Among the best conditions from this screen, the use of $NaHCO_3$ in NMP was revealed as the most robust set of conditions through subsequent parallel synthesis studies (not shown). (c) Registration of compounds with the 2-

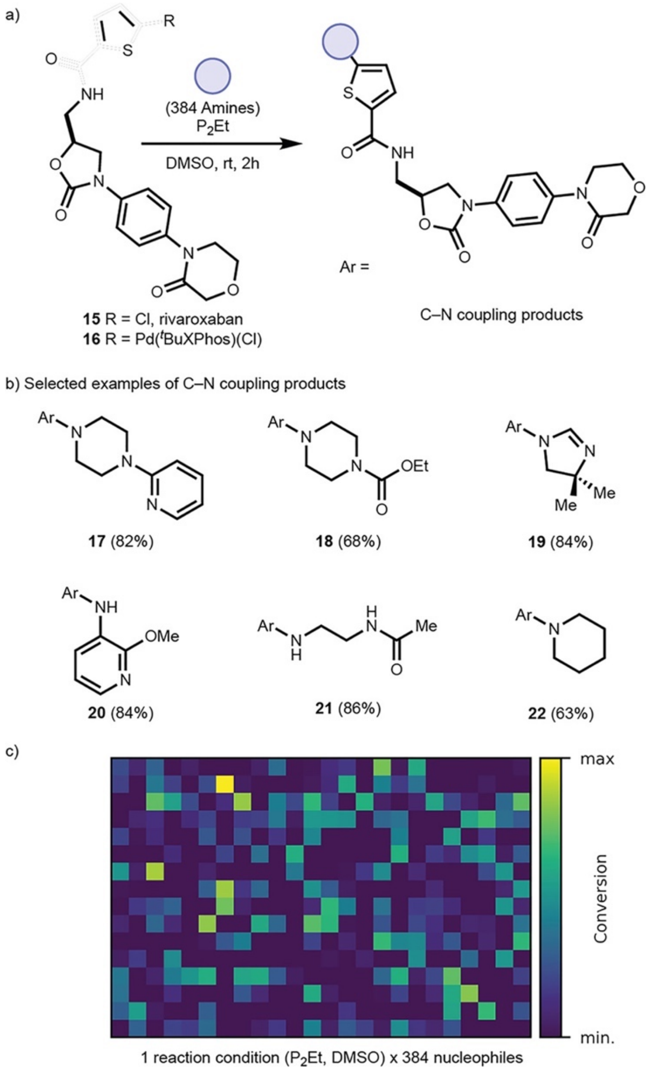
aminopyridine motif on the DGAT1 program accelerated after the day that the screen was run (red line). (d) The availability of these reaction conditions enabled hit expansion on a series of DGAT1 inhibitors, culminating in key compounds such as **3**–**7**.



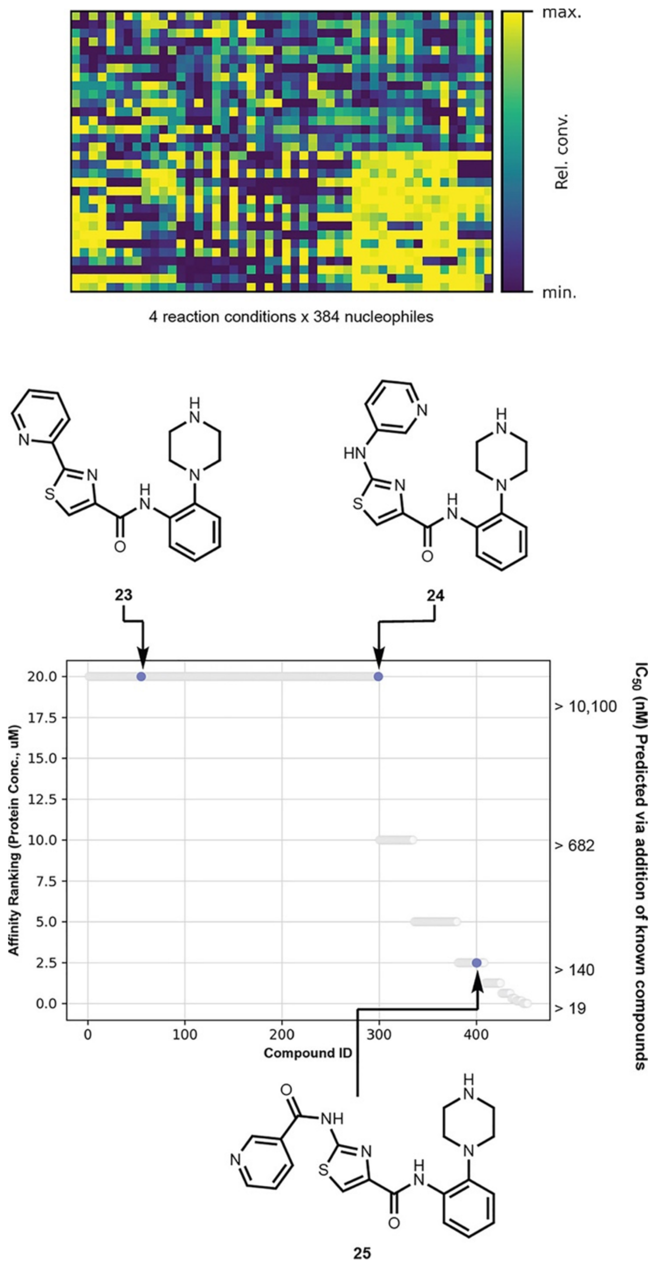
45-Fig. 2 | (a) Miniaturized reaction exploration on a milligram scale was a key enabling tool in the late-stage functionalization of medicinal compounds. (b) A survey of ligands for the iridium-catalyzed C–H borylation enabled the selective functionalization of **8**. Abbreviations: dtbpy = 4,4′-di-*tert*-butyl-2,2′-bipyridine; $P(C_6F_5)_3$ = tris(pentafluorophenyl)phosphine; Me_4Phen = 3,4,7,8-tetramethyl-1,10-phenanthroline; dbfa = (*E*)-*N*,*N*-dibenzyl-*N*′-(pyridin-2-yl)formimidamide; MeDPA = *N*-methyl-*N*-(pyridin-2-yl)pyridin-2-amine; ^{t}dtbpp = tris(2,4-di-*tert*-butylphenyl) phosphite.



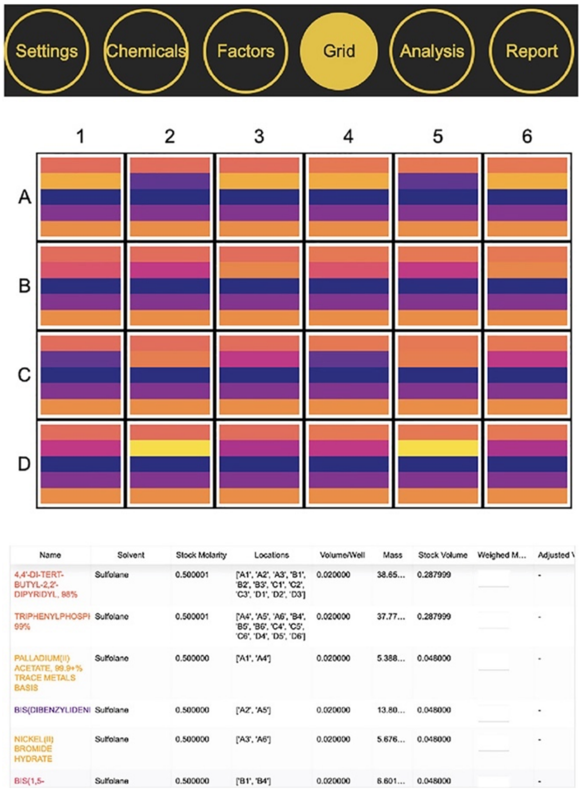
46-Fig. 3 | Development of ultraHTE. Chemical synthesis was performed in 1536-well plates. (A) Complex medicinally relevant molecules such as **14** could be synthesized on the gram scale using air-sensitive catalysts that were synthesized under conditions determined on the nanomole scale. (B) SPT Labtech mosquito. (C) Heat map of reaction performance for 1536 cross-coupling reactions on complex substrates.



47-Fig. 4 | (a) Use of ultraHTE to explore the substrate scope of cross-couplings on isolated palladium oxidative addition complexes (OACs). (b, c) Exploration of the substrate scope (b) on the nanoscale (c) highlights that OACs are robust coupling partners, with product observed by UPLC–MS for 54% of 384 nucleophiles attempted, compared with 10% using complementary catalytic conditions for the same library of nucleophiles.

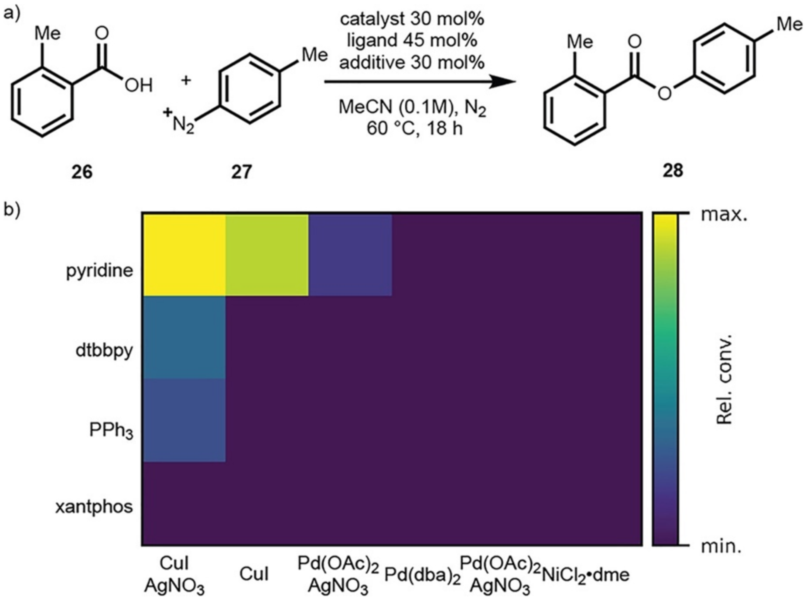


48-Fig. 5 | Nanoscale synthesis and affinity ranking using ASMS analysis of crude ultraHTE reaction mixtures. A series of known CHK1 inhibitors were used as benchmarking compounds (right ordinate). Compound **25** was readily observable as a molecule with higher affinity for CHK1 than **23** and **24** based on protein titration affinity ranking.

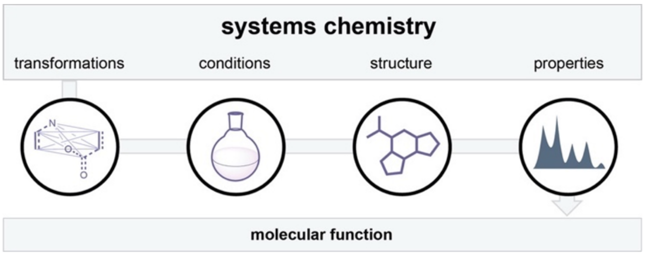


Name	Solvent	Stock Molarity	Locations	Volume/Well	Mass	Stock Volume	Weighed M...	Adjusted
4,4'-DI-TERT-BUTYL-2,2'-DIPYRIDYL, 98%	Sulfolane	0.500001	['A1', 'A2', 'A3', 'B1', 'B2', 'B3', 'C1', 'C2', 'C3', 'D1', 'D2', 'D3']	0.020000	38.65...	0.287999		-
TRIPHENYLPHOSPI 99%	Sulfolane	0.500001	['A4', 'A5', 'A6', 'B4', 'B5', 'B6', 'C4', 'C5', 'C6', 'D4', 'D5', 'D6']	0.020000	37.77...	0.287999		-
PALLADIUM(II) ACETATE, 99.9+% TRACE METALS BASIS	Sulfolane	0.500000	['A1', 'A4']	0.020000	5.388...	0.048000		-
BIS(DIBENZYLIDENI	Sulfolane	0.500000	['A2', 'A5']	0.020000	13.80...	0.048000		-
NICKEL(II) BROMIDE HYDRATE	Sulfolane	0.500000	['A3', 'A6']	0.020000	5.676...	0.048000		-
BIS(1,5-	Sulfolane	0.500000	['B1', 'B4']	0.020000	6.601...	0.048000		-

49-Fig. 6 | phactor is a high-throughput experimentation design software. The reaction array design tool is shown.



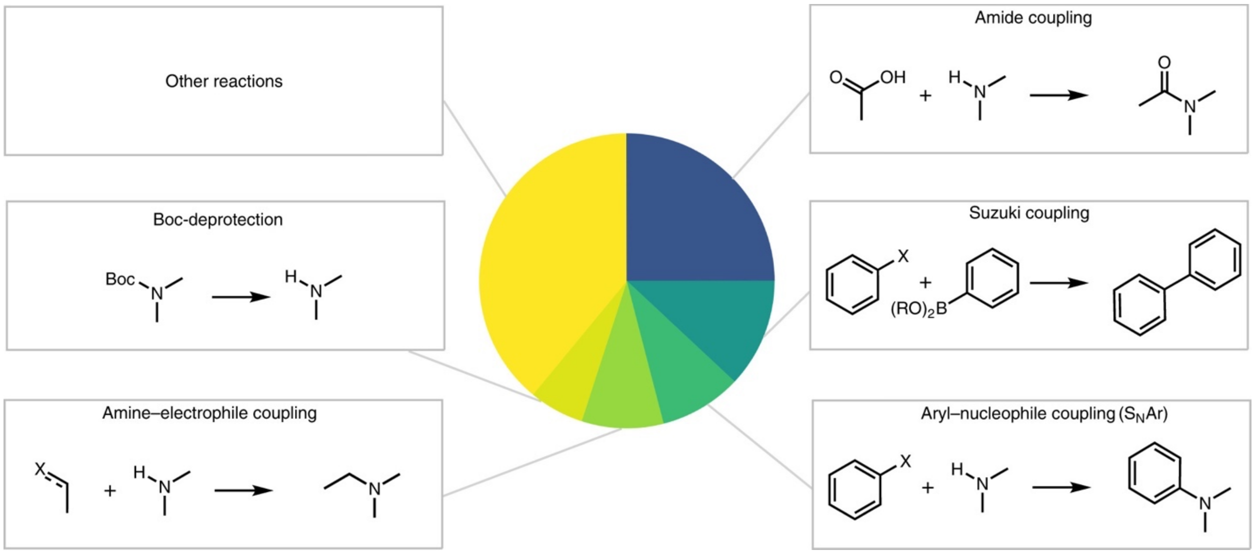
50-Fig. 7 | HTE screen of transition metals and ligands, designed in phactor, to realize a deaminative esterification.



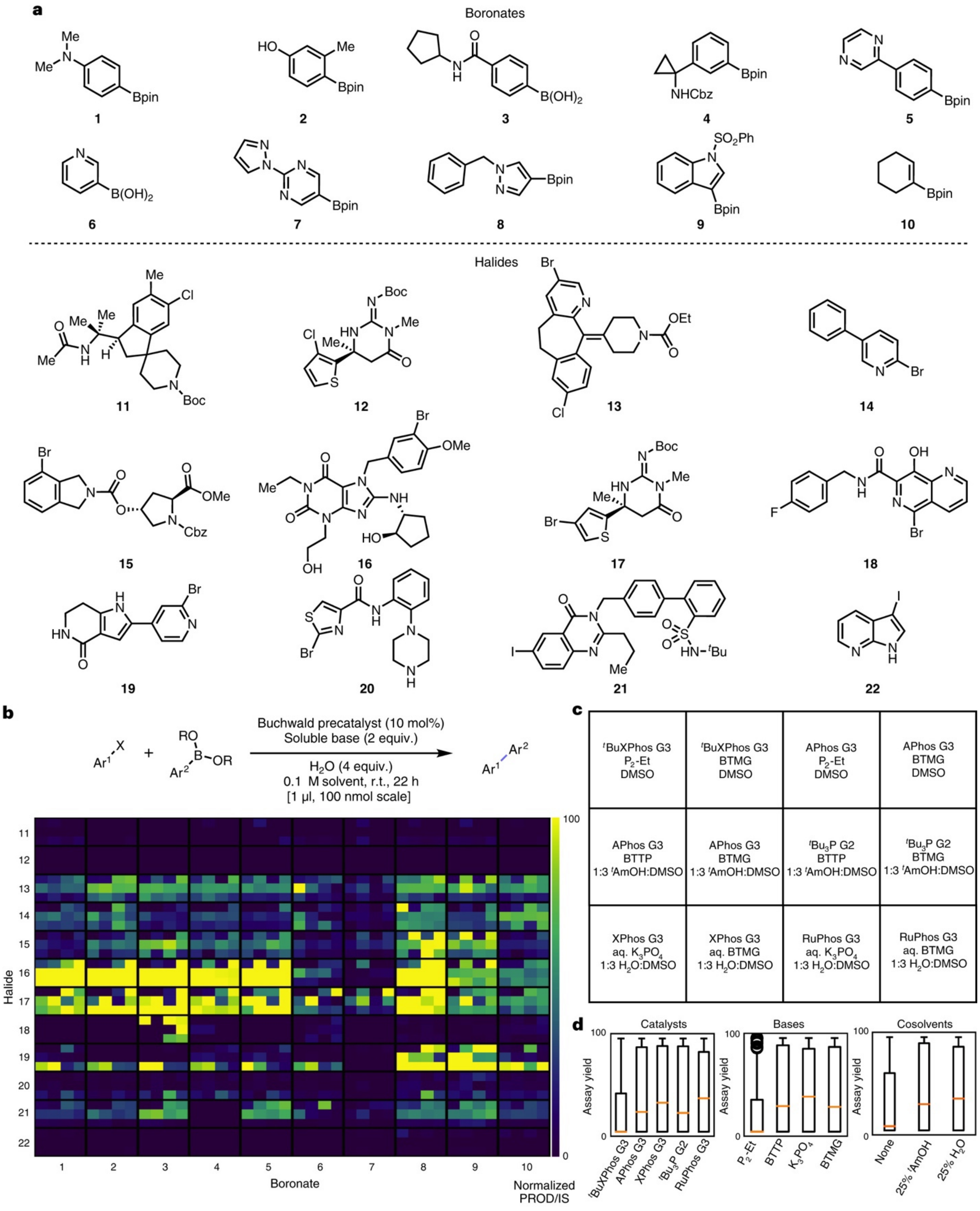
51-Fig. 8 | A systems chemistry approach that links chemical transformations and reaction conditions to structure, properties, and molecular function is enabled by ultraHTE.

Miniaturization of Popular Reactions from the Medicinal Chemists' Toolbox for Ultrahigh-Throughput Experimentation

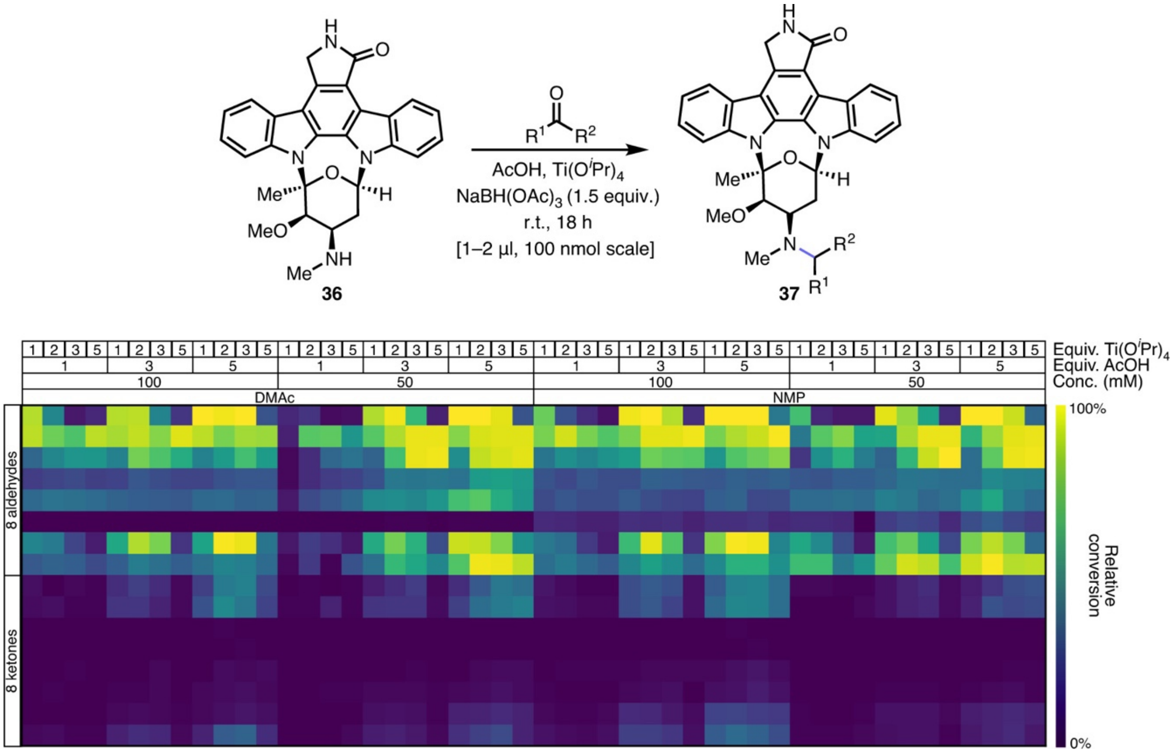
In this work, I compiled and visualized various datasets obtained through the development of miniaturized reactions that are popular in medicinal chemistry. The data was organized to inform the community the effectiveness of these reactions at nano and micro scale. This report is featured in *Nature Synthesis*.[134]



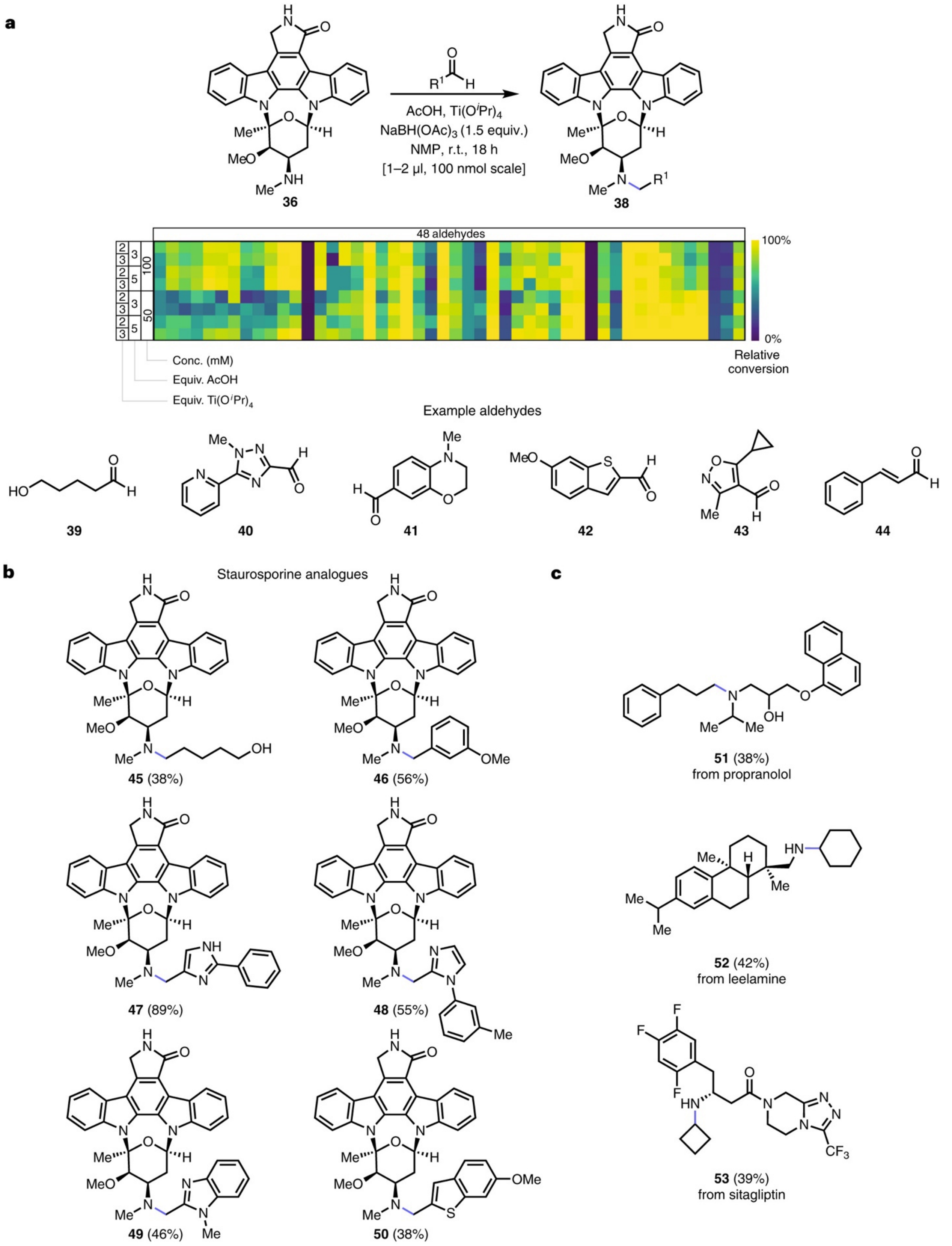
52-Fig. 1 | Popularity of common reactions in the synthesis of pharmaceuticals.



53-Fig. 2 | Nanoscale ultraHTE Suzuki coupling reaction performance.a, Boronates (**1**–**10**) and halides (**11**–**22**) used in the study. b, Heatmap of conversion to product as determined by UPLC–MS relative to an internal standard at λ = 254 nm. c, Plate map. d, Box-plots comparing the effects of base, cosolvent and catalyst on reaction performance (black circles indicate outlying points). Bpin, 4,4,5,5-tetramethyl-1,3,2-dioxaborolane; Cbz, benzyloxycarbonyl; tAm, tert-amyl; r.t., room temperature; PROD/IS, product/internal standard; aq., aqueous.



54-Fig. 3 | Nanoscale reaction optimization for the reductive amination of staurosporine (**36**) with eight aldehydes, eight ketones and 48 reaction conditions. See Supplementary Section 5.1 for full list of aldehydes and ketones. Relative percentage conversion was determined by UPLC–MS at λ = 254 nm. Conc., concentration.



45 (38%)

46 (56%)

47 (89%)

48 (55%)

49 (46%)

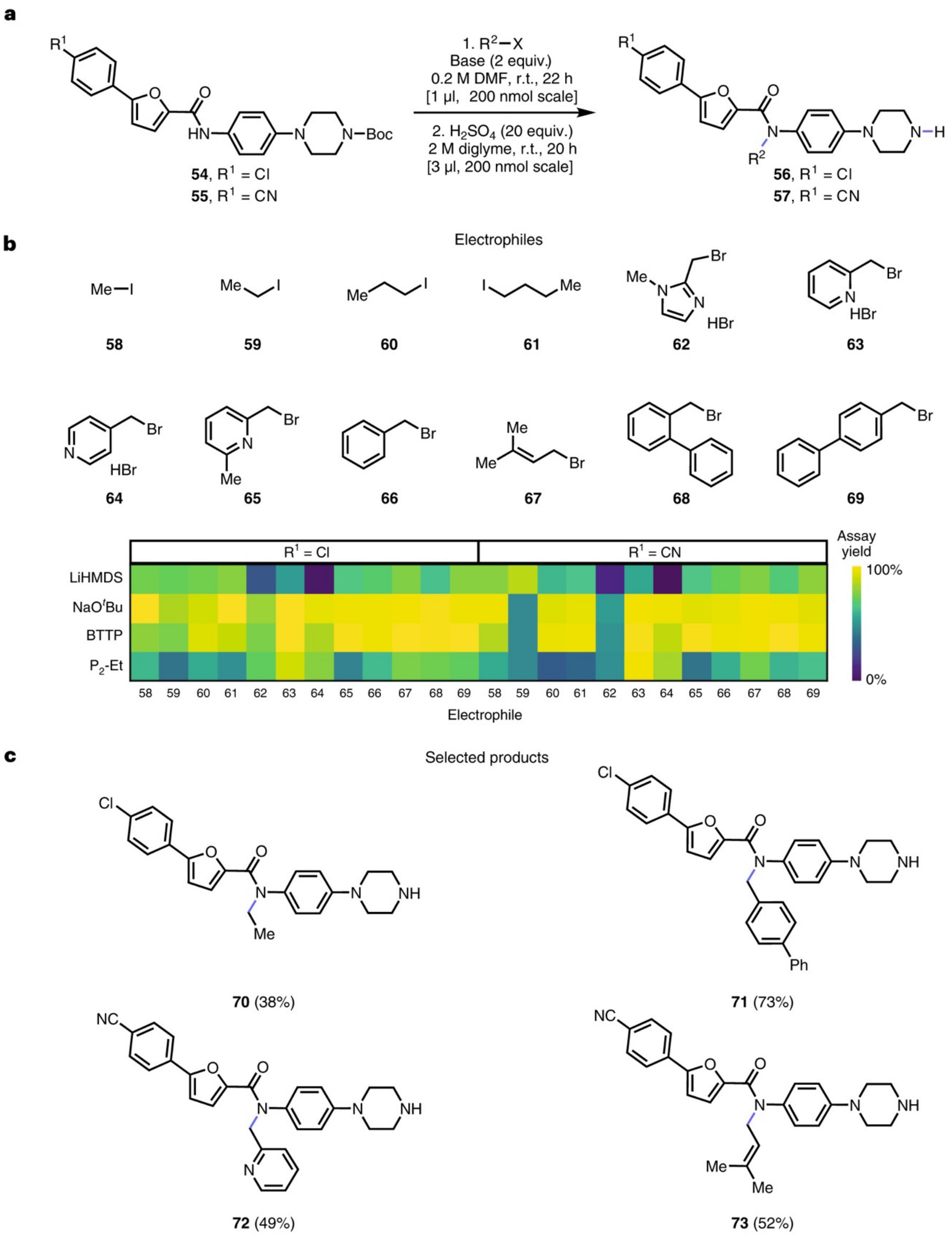
50 (38%)

51 (38%)
from propranolol

52 (42%)
from leelamine

53 (39%)
from sitagliptin

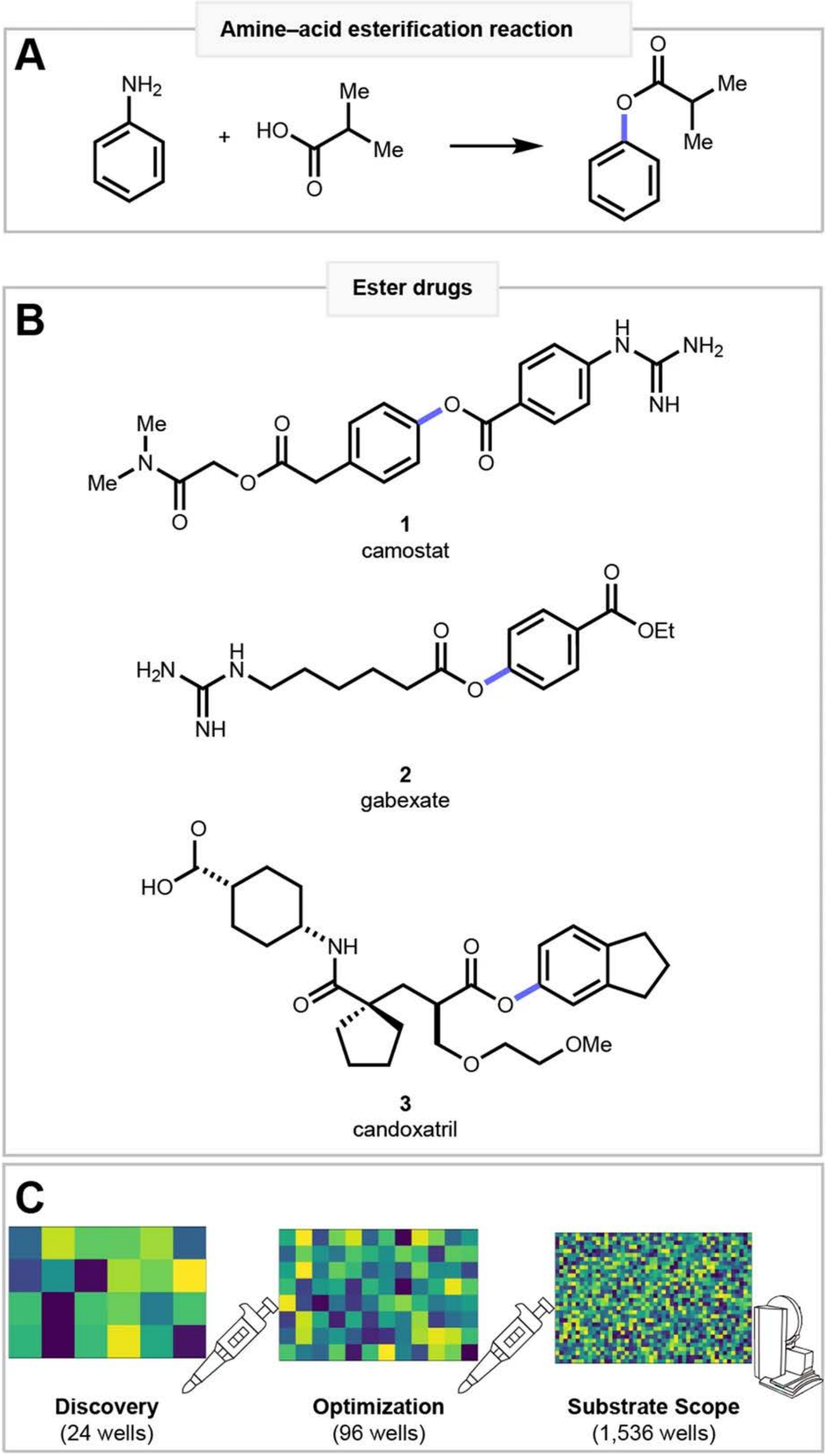
55-Fig. 4 | Nanoscale staurosporine analogue library synthesis via reductive amination. a, Heatmap of conversion to product for the reductive amination of staurosporine **36** with 48 aldehydes using eight reaction conditions. A representative example of aldehydes is shown (see Supplementary Section 5.2 for full list). Relative conversion was determined by UPLC–MS at λ = 254 nm. b,c, Selected reactions were repeated on 50 μmol scale giving **45**–**50** and S1 (b) or 0.30 mmol scale giving **51**–**53** (c). Yields are of purified products.



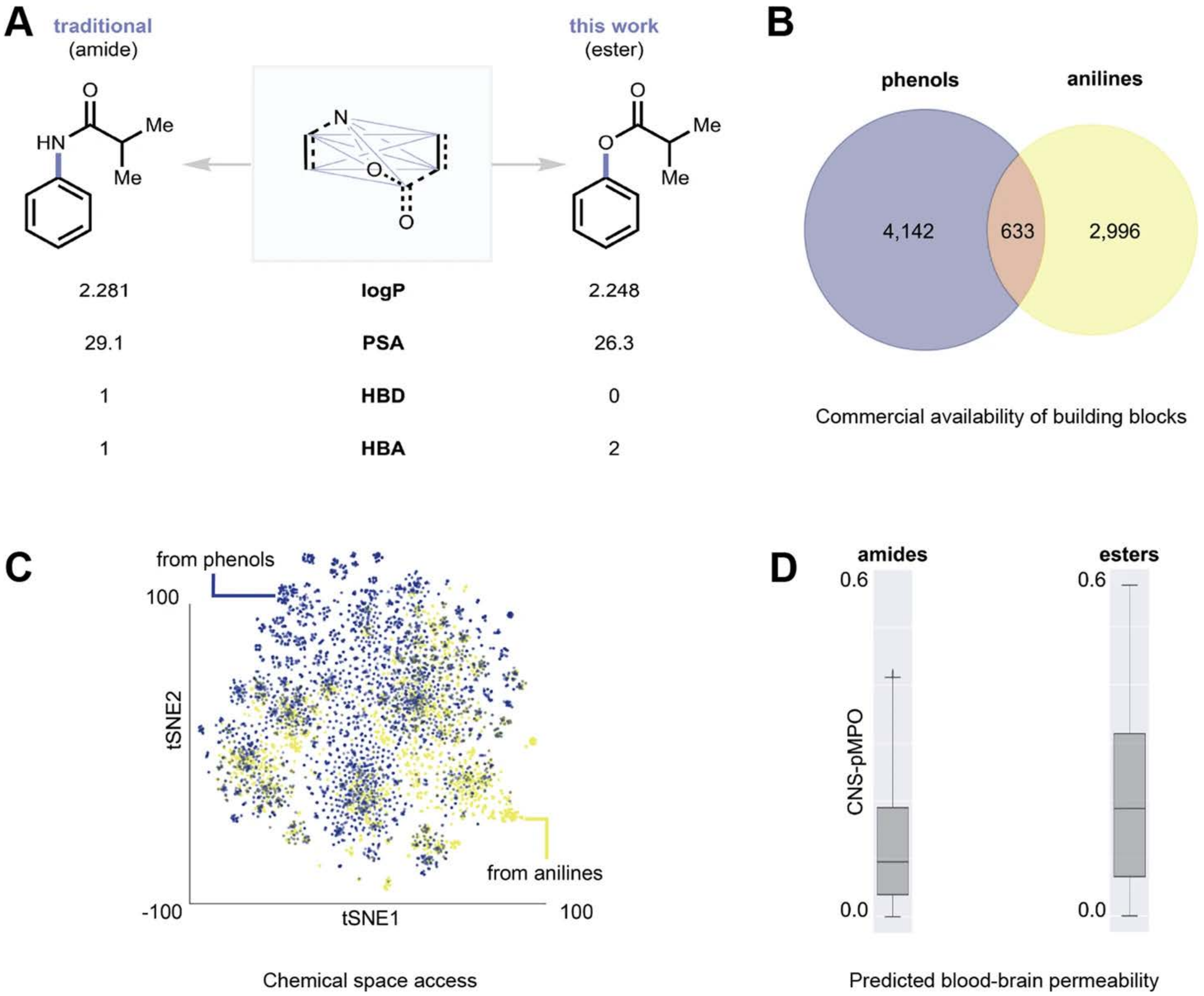
56-Fig. 5 | Two-step nanoscale synthesis of MK2-inhibitor analogues (**56**) and (**57**) via N-alkylation followed by Boc-deprotection. a, Electrophiles **58**–**69** were coupled to **54** and **55**. b, Heatmap of reaction performance determined by UPLC–MS at λ = 254 nm relative to standard curves created from isolated products. c, Selected examples **70–73** repeated on 40 μmol scale (see Supplementary Section 6.5 for full scope). Yields are of purified products. LiHMDS, lithium bis(trimethylsilyl)amide.

Development of copper-catalyzed deaminative esterification using high-throughput experimentation

In this work led by Yuning Shen, a copper-catalyzed deaminative esterification was developed. In this transformation, an amine activated as a diazonium salt reacts with a carboxylic acid to form various esters. The reaction was optimized to perform well in ultrahigh throughput format and is amenable to a various of substrates. We show that amine substrates provide better access to novel space compared to their phenol counterparts. This work is featured in *Chemical Communicatons*.[180]

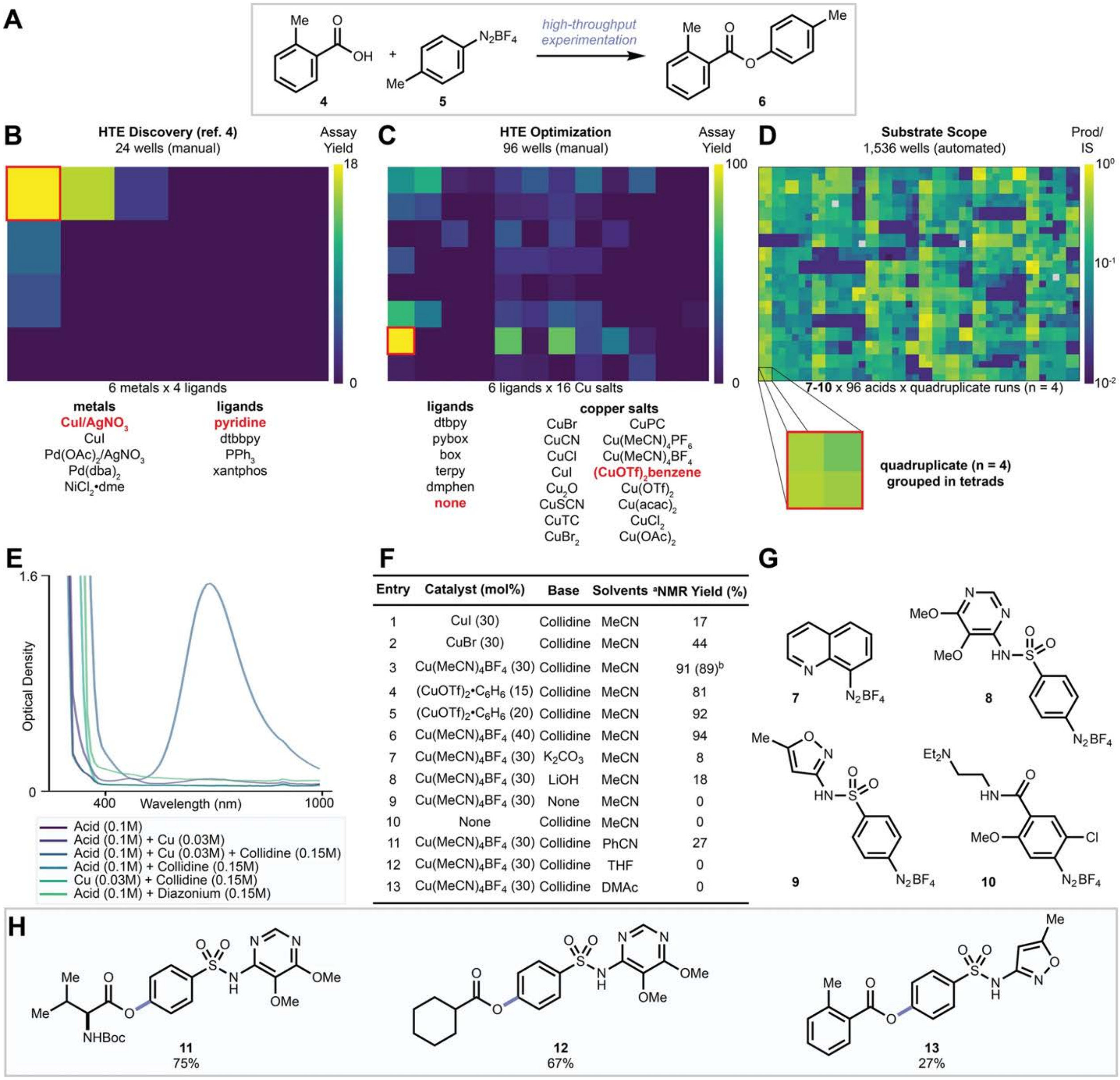

A
Amine–acid esterification reaction
NH2
HO
Me
Me
O
+
O
O
Me
Me
B
Ester drugs
Me
Me
N
O
O
O
O
O
O
H
N
NH2
NH
1
camostat
H2N
H
N
NH
O
O
O
OEt
2
gabexate
HO
O
NH
O
O
O
OMe
3
candoxatril
C
Discovery
(24 wells)
Optimization
(96 wells)
Substrate Scope
(1,536 wells)

57-Fig. 1 | HTE enabled amine–acid esterification and its application in drug molecules. A The amine–acid esterification. **B** Esters are an important medicinal functionality as in drugs **1**–**3**. **C** Reaction discovery, optimization, and scope profiling is achieved using HTE.



58-Fig. 2 | Physicochemical properties of amides and esters and complementary chemical space of anilines to phenols. A An amine and carboxylic acid can be coupled in a variety of transformations beyond the traditional amide coupling, including the esterification reaction developed here, which gives a unique property footprint. **B**

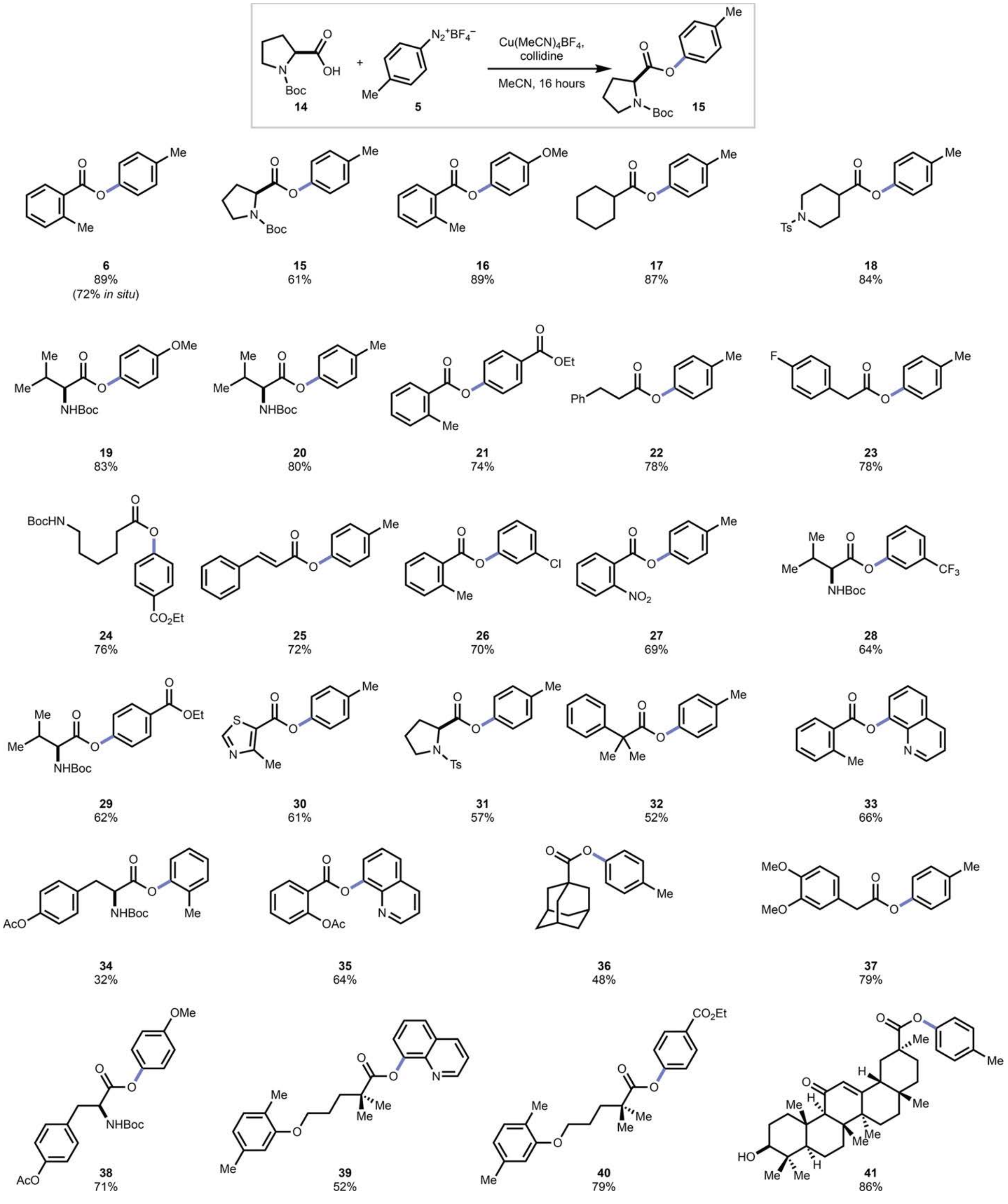
Venn diagram showing the complementarity of phenol to aniline building blocks available in the MilliporeSigma catalog. **C** tSNE analysis showing the complementarity of ester products generated by coupling aspirin to phenol (blue dots) versus aniline (yellow dots) building blocks available in the MilliporeSigma catalog. **D** Ester products have a higher predicted blood–brain permeability (CNS-pMPO score) than classic amide products produced from the same set of anilines when coupled to the drug metaclopramide.



Entry	Catalyst (mol%)	Base	Solvents	ᵃNMR Yield (%)
1	CuI (30)	Collidine	MeCN	17
2	CuBr (30)	Collidine	MeCN	44
3	Cu(MeCN)$_4$BF$_4$ (30)	Collidine	MeCN	91 (89)^b
4	(CuOTf)$_2$•C$_6$H$_6$ (15)	Collidine	MeCN	81
5	(CuOTf)$_2$•C$_6$H$_6$ (20)	Collidine	MeCN	92
6	Cu(MeCN)$_4$BF$_4$ (40)	Collidine	MeCN	94
7	Cu(MeCN)$_4$BF$_4$ (30)	K$_2$CO$_3$	MeCN	8
8	Cu(MeCN)$_4$BF$_4$ (30)	LiOH	MeCN	18
9	Cu(MeCN)$_4$BF$_4$ (30)	None	MeCN	0
10	None	Collidine	MeCN	0
11	Cu(MeCN)$_4$BF$_4$ (30)	Collidine	PhCN	27
12	Cu(MeCN)$_4$BF$_4$ (30)	Collidine	THF	0
13	Cu(MeCN)$_4$BF$_4$ (30)	Collidine	DMAc	0

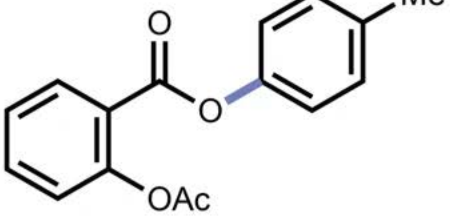
59-Fig. 3 | Reaction discovery and profiling with HTE. A General deaminative esterification reaction. **B** Heatmap showing the discovery of the esterification reaction in 24 glass microvials. Dtbpy = 4,4′-di-*tert*-butyl-2,2′-dipyridyl, xantphos = 4,5-bis(diphenylphosphino)-9,9-dimethylxanthene. **C** Heatmap showing optimization of esterification in 96 glass microvials (see Supplementary Information for details). Pybox = 2,6-bis[(*4 S*)-(–)-isopropyl-2-oxazolin-2-yl]pyridine, box = 2,2-Bis((*4 S*)-(–)-4-

isopropyloxazoline)propane, terpy = 2,2′:6′,2″-terpyridine, dmphen = 4,7-dimethoxy-1,10-phenanthroline, CuTC = copper(**I**) thiophene-2-carboxylate, CuPC = copper(**II**) phthalocyanine. **D** Heatmap showing the reproducibility of quadruplicate data for a library of 96 acids coupled to **7**–**10** in plastic 1536-well plates. The overall average standard deviation across the quadruplicate data was 5.5%. The wells with the gray color indicated that the dosage of the internal standard was missing. **E** UV-Vis absorbance data show an interaction between $Cu(MeCN)_4BF_4$, **4** and 2,4,6-collidine. **F** Reaction optimization. a ^1H-NMR yield with 1,3,5-trimethoxybenzene as internal standard, b Isolated yield. **G** Diazonium salts used in the 1536 screen shown in **D**. **H** Select reactions from those shown in **D** were repeated on a 0.300 mmol scale using $Cu(MeCN)_4BF_4$ (100 mol%) to produce **11**, **12**, and **13** in isolated yield shown.

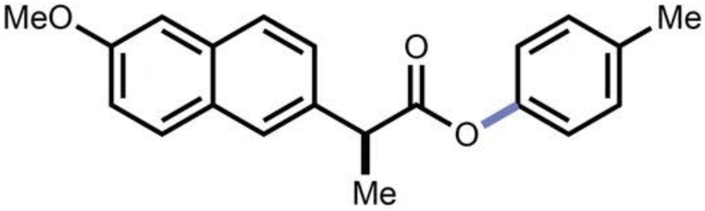


60-Fig. 4 | Substrate scope. Reactions were run with carboxylic acids (0.300 mmol), diazonium salts (1.5 equiv.), $Cu(MeCN)_4BF_4$ (30 mol%), and 2,4,6-collidine (1.5 equiv.)

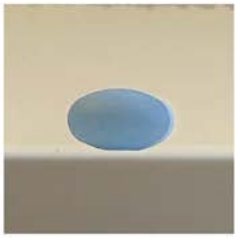
in anhydrous acetonitrile (0.1 M) at room temperature, generally over 16 h. Reported yields are isolated yields of purified products.

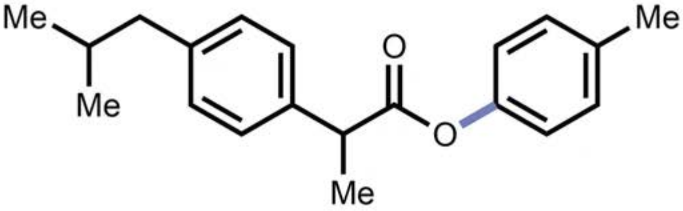


42 (Aspirin)
92% yield (reagent grade)
80% yield (pill)

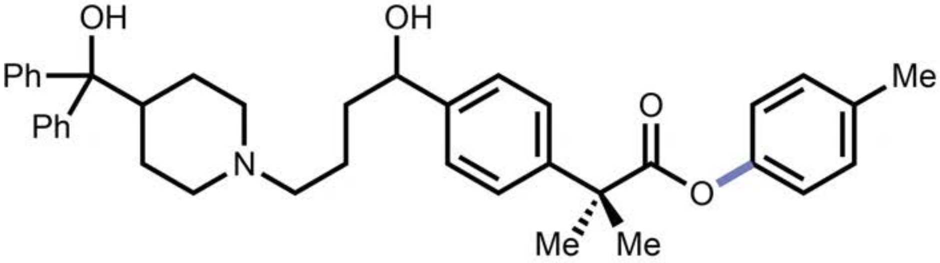


43 (Naproxen)
88% yield (reagent grade)
61% yield (pill, Na salt)





44 (Ibuprofen)
64% yield (reagent grade)
61% yield (pill)



45 (fexolanadine)
27% yield (reagent grade)
15% yield (pill, HCl salt)

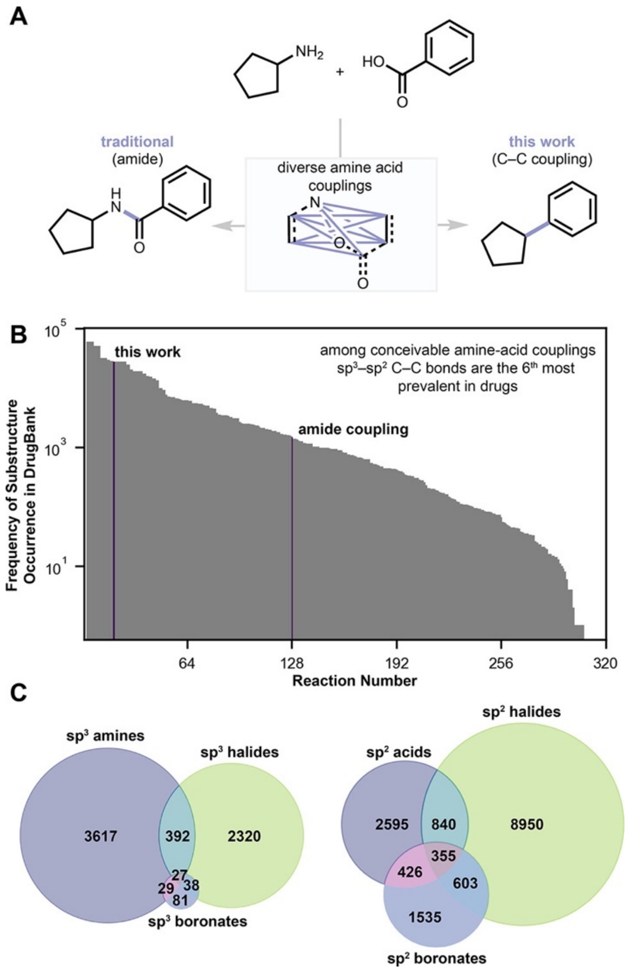
61-Fig. 5 | The amine–acid esterification applied directly to over-the-counter pills.

Ester products **42**–**45** are derived from both reagents and OTC pills

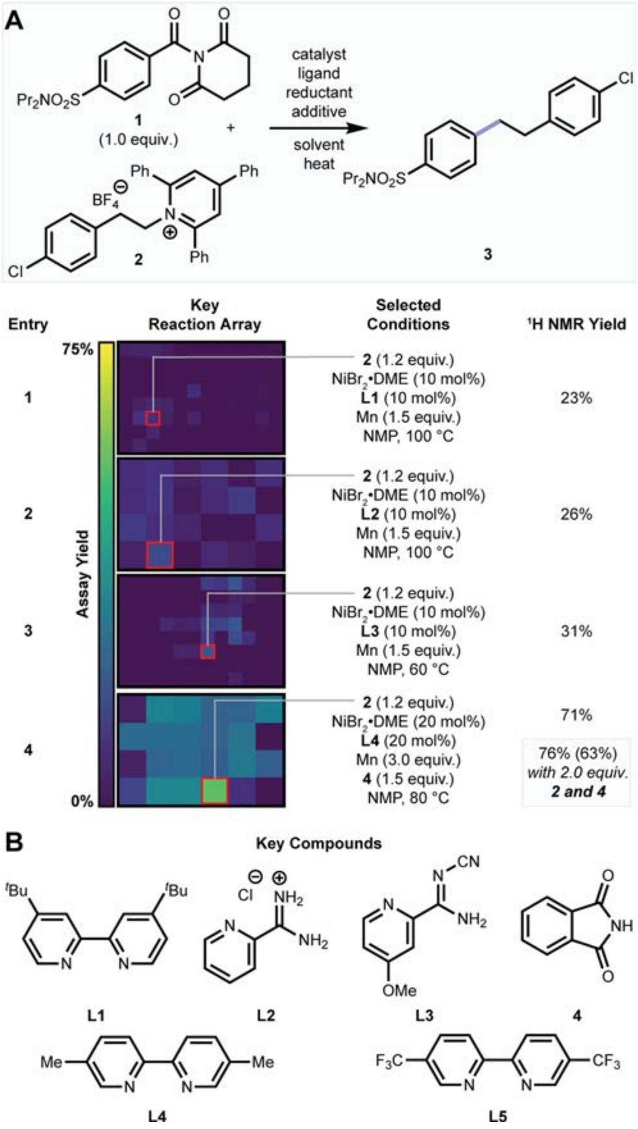
(See Supplementary Information for details).

The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp^3–sp^2 Carbon–Carbon Bonds

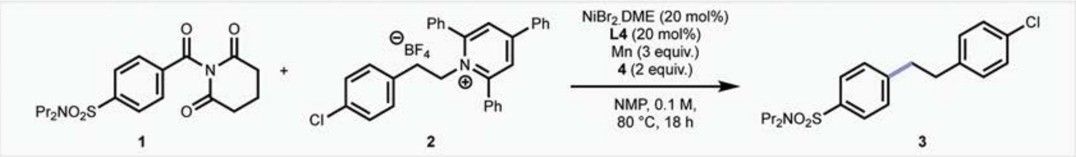
In this work led by James Douthwaite, a nickel-catalyzed deaminative decarboxylative C–C bond forming reaction was developed. In this transformation, an amine activated as a Katritzky salt reacts with a carboxylic acid activated as a twisted amide to form various C–C bonds. The reaction was optimized using phactor through a serious of high throughput experiments and is amenable to a various of substrates. We show that amine and acid substrates provide better access to novel space compared to their boronic and halide counterparts used in the Suzuki reaction. This work is featured in *JACS*.[102]

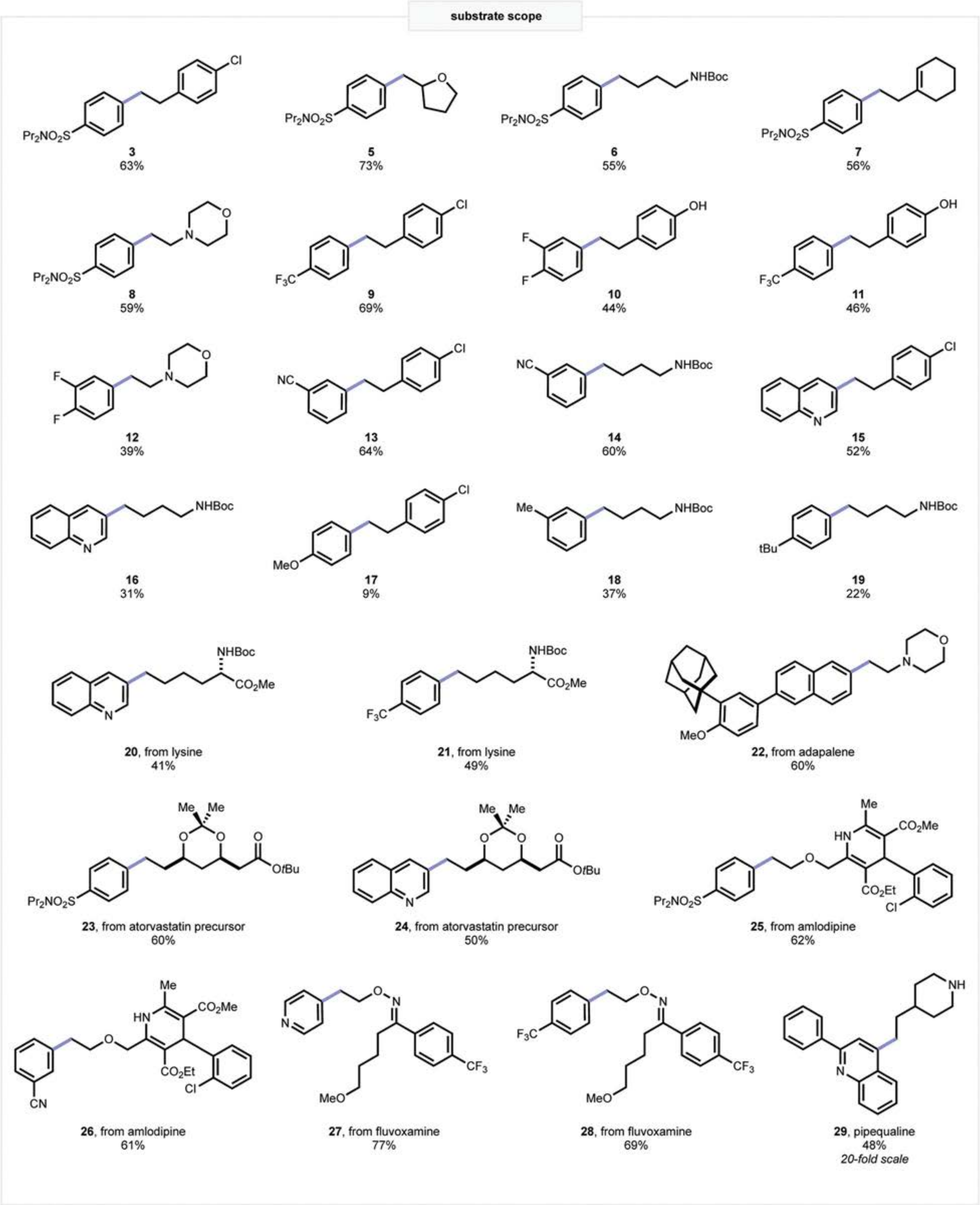


62-Fig. 1 | (A) Cross-coupling of amines and carboxylic acids to form C(sp^3)–C(sp^2) bonds as a complement to amide coupling. (B) Substructure search of 320 possible products from amine–acid coupling in the DrugBank database. (C) Analysis of building block availability from MilliporeSigma. Unions represent exact substructures differing only by the indicated functionality.

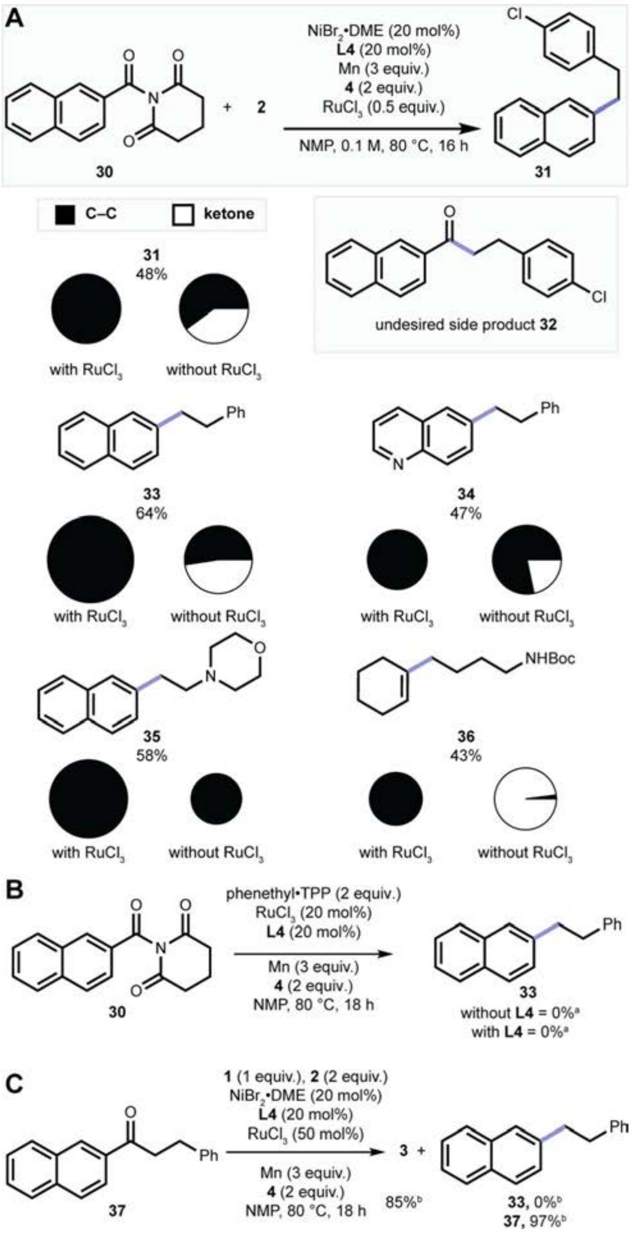


63-Fig. 2 | (A) HTE optimization on 10 μmol scale in 24- and 96-reaction arrays. Assay yields were determined by UPLC-MS. Selected conditions were repeated on the 0.25 mmol scale and analyzed by ^{1}H NMR with an internal standard. Isolated yield in parentheses. (B) Key compounds.


Pr$_2$NO$_2$S
1
+
BF$_4$
Ph
Ph
Ph
Cl
2
NiBr$_2$ DME (20 mol%)
L4 (20 mol%)
Mn (3 equiv.)
4 (2 equiv.)
NMP, 0.1 M,
80 °C, 18 h
Cl
Pr$_2$NO$_2$S
3

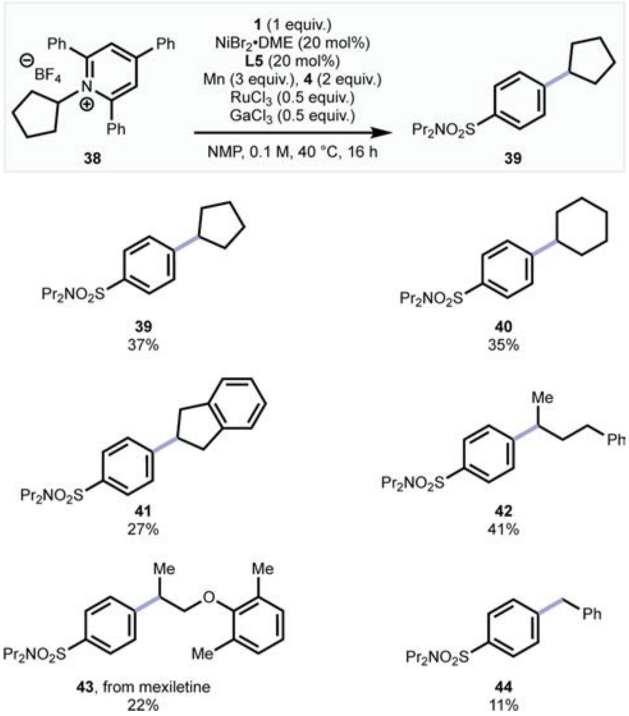

substrate scope
3
63%
5
73%
6
55%
7
56%
8
59%
9
69%
10
44%
11
46%
12
39%
13
64%
14
60%
15
52%
16
31%
17
9%
18
37%
19
22%
20, from lysine
41%
21, from lysine
49%
22, from adapalene
60%
23, from atorvastatin precursor
60%
24, from atorvastatin precursor
50%
25, from amlodipine
62%
26, from amlodipine
61%
27, from fluvoxamine
77%
28, from fluvoxamine
69%
29, pipequaline
48%
20-fold scale

64-Fig. 3 | Substrate scope of C(sp^3)–C(sp^2) coupling of *N*-acyl-glutarimides with primary alkylpyridinium salts on 0.10–0.25 mmol scale. Isolated yields are shown.

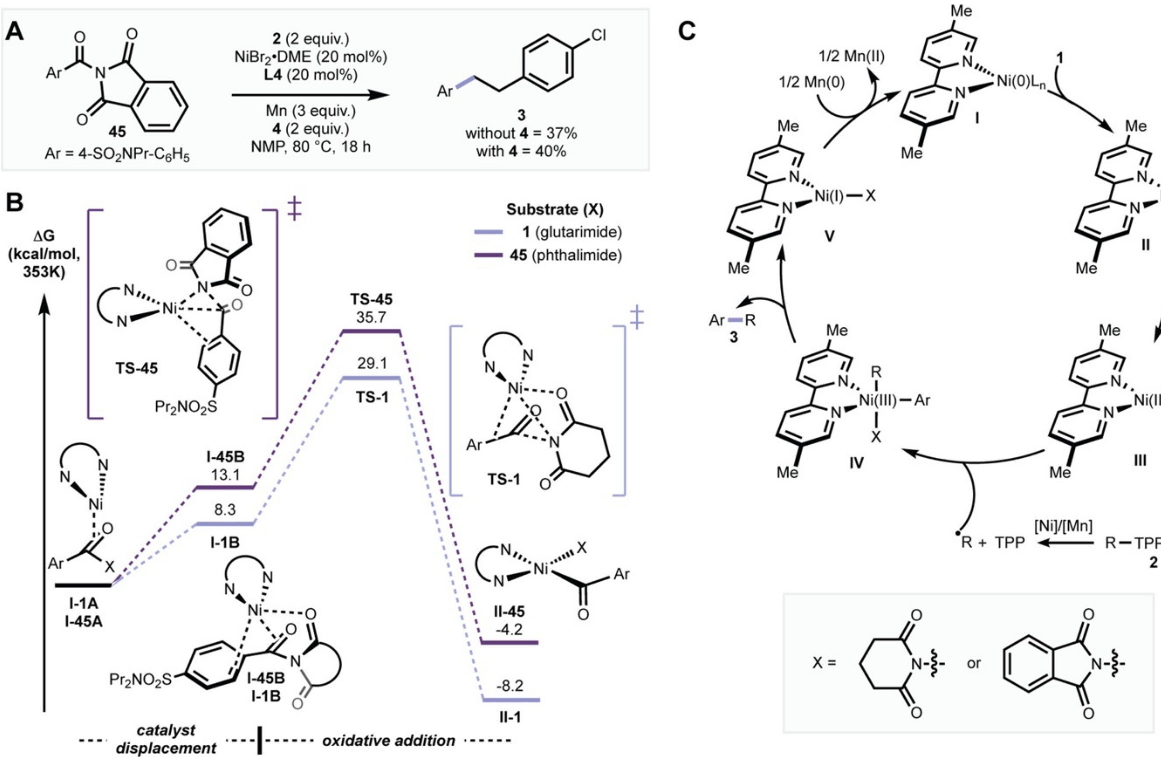


65-Fig. 4 | (A) Effect of $RuCl_3$ on selectivity. Isolated yield on 0.25 mmol shown. Pie charts determined from ^1H NMR analysis of crude reaction with reference to an internal standard. (B) Effect of $RuCl_3$ in the absence of $NiBr_2 \cdot DME$. TPP = triphenylpyridinium.

(C) Coupling of **1** and **2** in the presence of ketone **37** and $RuCl_3$. [a]Not detected by UPLC-MS. [b]Yield from ^1H NMR.



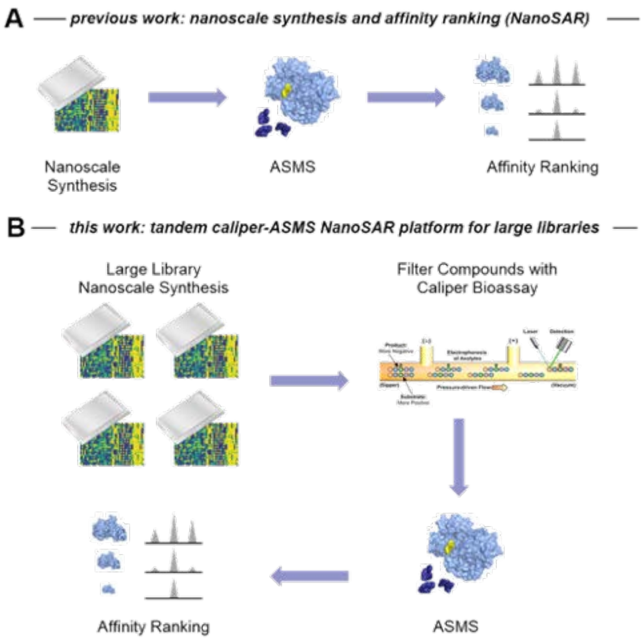
66-Fig. 5 | Substrate scope of $C(sp^3)$–$C(sp^2)$ coupling of *N*-acyl-glutarimides with secondary and benzyl alkylpyridinium salts on 0.25 mmol scale. Isolated yields are shown.



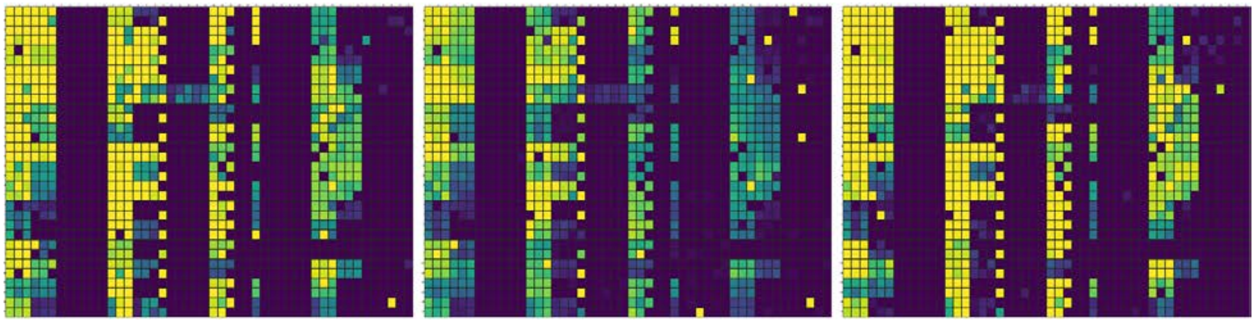
67-Fig. 6 | (A) Replacement of **1** with *N*-acyl phthalimide **45**, ^{1}H NMR yields shown. (B) DFT comparison of oxidative addition of **1** and **45**. (C) Proposed catalytic cycle.

Expansion of NanoSAR for the Deconvolution of Large Compound Libraries

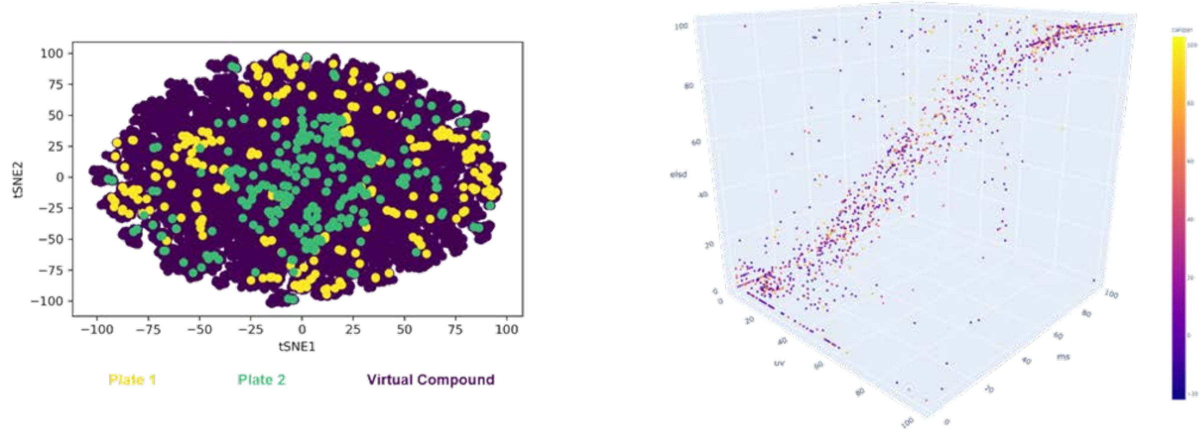
In this work we collaborated with the company Relay Therapeutics to expand the NanoSAR platform.[35] An amide coupling was developed to work in ultrahigh throughput format and a chemical library was developed for direct-to-biology studies. In this project, aliquots from reactions were pooled before being tested in affinity selection mass spectrometry. Much of the analysis relied on the deconvolution mass hits found and matching compounds to reaction wells.


A — previous work: nanoscale synthesis and affinity ranking (NanoSAR) —
Nanoscale Synthesis
ASMS
Affinity Ranking
B — this work: tandem caliper-ASMS NanoSAR platform for large libraries —
Large Library Nanoscale Synthesis
Filter Compounds with Caliper Bioassay
Affinity Ranking
ASMS

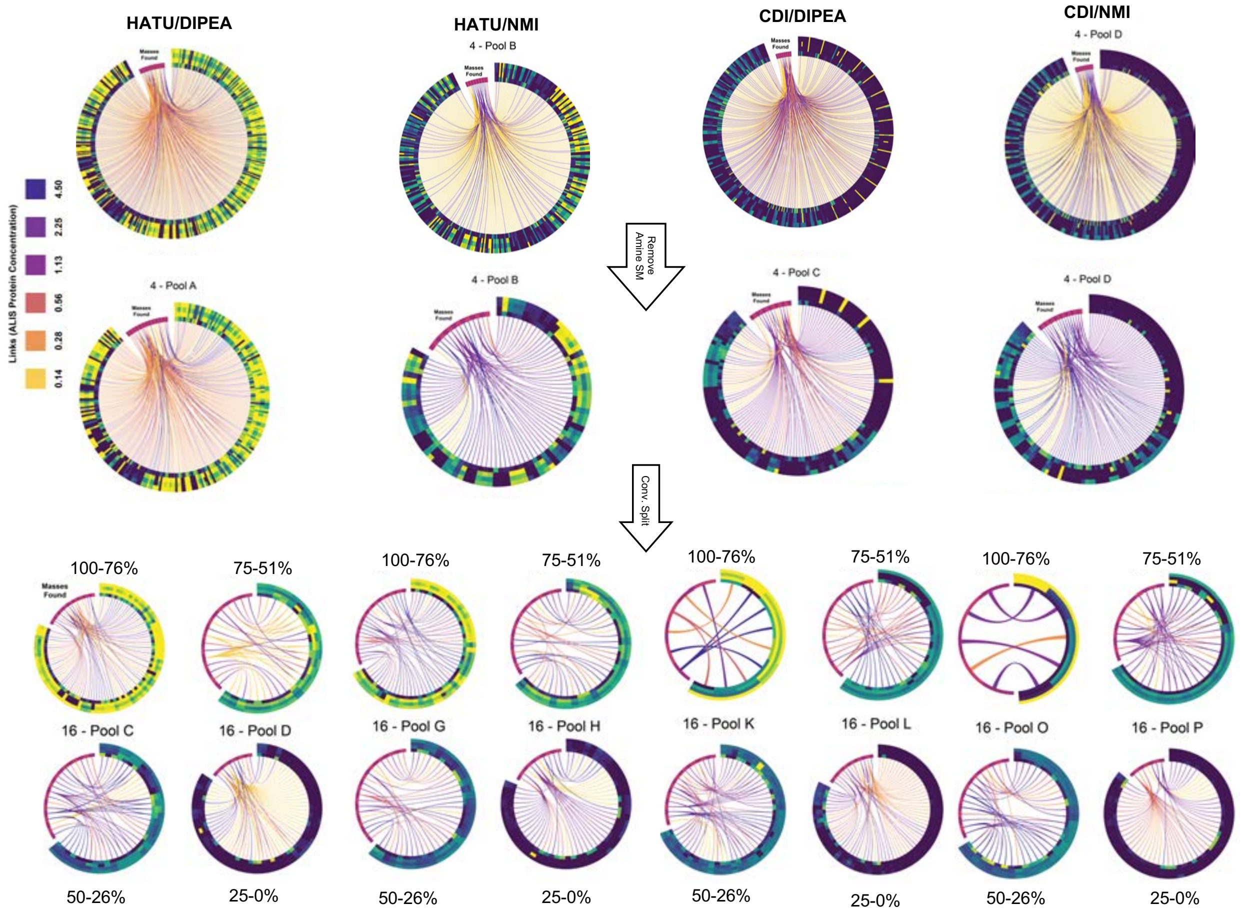
68-Fig. 1 | Approaches to nanoscale synthesis and affinity ranking (NanoSAR); (a) Previously, crude nanoscale reactions were assayed by ASMS with affinity ranking achieved by decreasing the concentration of target protein to induce competition among compounds; (b) introduction of a Caliper bioassay into the workflow allows larger libraries of thousands of compounds to be evaluated and reduces the need for mass encoding.



69-Fig. 2 | Chemical assay results of ultrahigh throughput experiment performed for the project. Heatmaps represent the UV, MS, and ELSD integrations of the product over internal standard respectively.



70-Fig. 3 | Library design and synthesis (left) tSNE comparing targeted compounds to virtual compounds generated from enumeration of the enamine acid catalog. The targeted compounds cover a large portion of chemical space. (right) Relative conversion to product for the ultraHTE amide coupling of pharmacophore amines with 128 different acids as determined by UPLC-MS. Three detection methods, UV-absorbance, MSTIC(+) and ELSD were utilized. Each point represents one reaction, colored by percent inhibition observed when subjected to Caliper biochemical assay.



71-Fig. 4 | Affinity Selection Mass Spectroscopy chord diagrams for each pool of reactions. All reactions were initially split into four pools, based on reagents, resulting in 768 reactions per pool. Chords connect reactions to compound masses found in that reaction. Each reaction may be connected to multiple masses, and heatmaps representing chemical and biological assay results are shown on the periphery of the reaction arc. Chords are colored by the lowest ALIS protein concentration that a compound in that reaction bound to. In the second iteration, 16 pools were made where reactions were split by conversion.

Chapter IV. Reagent Discovery

High-throughput experimentation is a valuable technology that allows for the rapid testing of various experimental conditions in chemical reactivity. However, due to the vast number and combinations of reagents available for testing, it becomes necessary to intelligently select the experimental conditions.[114, 125] While algorithms have been developed to optimize continuous variables such as concentration and temperature, the sampling of discrete variables or combinations of discrete variables still presents a challenging problem.

Expert chemists traditionally rely on literature searching and high-throughput experimentation to optimize reactivities, but there is a growing desire to automate this experimental design process in order to create self-driven laboratory workflows and optimize the reaction development process. The goal is to develop algorithms that can efficiently navigate the complex landscape of possible experimental conditions and intelligently select the most promising ones to test. By leveraging machine learning and artificial intelligence techniques, researchers aim to develop algorithms that can effectively explore the vast chemical space and guide the discovery of new reactions.[41, 144, 161, 176, 225, 226]

The automation of experimental design holds great potential for the discovery of conditions that enable novel reactivity. Currently, this is achieved through transfer learning from relevant literature and reaction datasets. Expert chemists leverage their knowledge and experience to identify key factors and parameters that contribute to desired reactions. By training algorithms on large datasets of known reactions,[149] it becomes possible to extract patterns and correlations that can guide the optimization of experimental conditions. This approach enables the efficient exploration of a vast chemical space, accelerating the discovery of new reactivity and facilitating the development of innovative chemical processes. Through the integration of high-throughput experimentation, machine learning, and automation, researchers aim to revolutionize the field of chemical discovery and enable the creation of self-driven laboratory workflows.[136, 157]

Designing Chemical Reaction Arrays using phactor and ChatGPT

The year 2023 brought the emergence of near lifelike artificial intelligence that could be spoken to naturally in a conversational manner. In this work, the artificial intelligence ChatGPT was interrogated to design reaction arrays to optimize common reactions between various substrates. In conjunction with phactor, these experiments were executed, and all designs resulted in the formation of the desired product. The artificial intelligence was then directly incorporated into phactor.

Abstract

High throughput experimentation (HTE) is a common practice in the optimization of chemical synthesis. Chemists design reaction arrays to optimize the yield of couplings between building blocks. Popular reactions used in pharmaceutical research include the amide coupling, Suzuki coupling, and Buchwald-Hartwig coupling. We show how the artificial intelligence (AI) language model ChatGPT can automatically formulate reaction arrays for these common reactions based on the literature corpus it was trained on. Critically, we showcase how ChatGPT results can be directly translated into inputs for the management software phactor, which enables automated execution and analysis of assays. This workflow is experimentally demonstrated, with modest to excellent yields of products obtained in each instance on the first attempt.

Introduction

Chemical synthesis is a primary bottleneck in drug development. High throughput experimentation (HTE) is a widely practiced method for the discovery and optimization of reaction conditions in medicinal chemistry campaigns.[35, 36, 108, 140, 141] Chemists typically design reaction arrays based on conditions found in the literature, using search tools such as Google, SciFinder, or Reaxys. The automated generation of reaction arrays to optimize or discover a coupling reaction between two substrates is a contemporary problem.[124, 125, 161, 176, 225, 227, 228] Recently, generative transformers, a form of artificial intelligence (AI), have emerged as interactive language models that can interpret and answer scientific questions via verbal human input.[229, 230] Herein we demonstrate how the general-purpose language model ChatGPT (version 3.5) can be utilized to generate initial-guess reaction array designs for specific substrate pairs. Due to the flexibility of the AI model, the output can be directly translated into input files for the HTE management software phactor.[136] We showcase several case studies of using ChatGPT to aid in designing reaction arrays for phactor, specifically for transformations that are most commonly used in pharmaceutical chemistry.[30] With phactor, we execute the arrays designed by ChatGPT experimentally leading to viable first-pass reaction conditions from simple prompts that are easy to devise by non-expert users.

Experimental

To test the effectiveness of reaction arrays designed by ChatGPT, a workflow to automatically generate reagent proposals and execute reaction arrays for popular reactions was developed. A typical workflow can be summarized in three steps (72-Fig. 1):

- First, have ChatGPT generate reaction array designs for specific substrates based on simple text prompts.
- Then, translate the output from ChatGPT into an input file for the HTE management software phactor. An interfacing script written in python is provided online.
- Finally, use phactor to create stock solutions and distribute the chemicals into the reaction array, manually or robotically, and then analyze its results.

ChatGPT can be interrogated during the design step to elaborate on experimental details or reasonings and was asked to clarify experimental details at times. Each product was scaled up using the best conditions identified for its respective reaction array and isolated.

Results

ChatGPT is a newly released general-purpose AI language model developed by OpenAI.[231] It serves as a conversational model where the user can ask a series of questions and receive text answers based on the context of the conversation. While not directly a model for chemistry, ChatGPT has been trained on a large corpus of scientific literature. As such, in its own words, ChatGPT has "knowledge of basic chemistry concepts, such as the periodic table, chemical reactions, acids and bases, and thermodynamics. It can also provide information on more advanced topics, such as organic chemistry, biochemistry, and physical chemistry." We demonstrate how ChatGPT can be asked to generate reaction arrays of viable reagents and catalysts for common reaction classes for specific substrates. For each class of reactions (amide, Suzuki, and Buchwald-Hartwig couplings), we ask ChatGPT to develop an experimental design for various pairs of substrates. While we hypothesize that the model is exercising little, in any, physical and chemical intuition in its designs, its ability to select popular reagents and catalysts associated with reaction-type keywords leads to viable and interesting proposals for array recipes. Critically, the merger with phactor exploits the strength of ChatGPT to propose several plausible answers, then sample them systematically using

HTE as opposed to relying on a single “correct” answer. Ultimately, this merger of ChatGPT and phactor led to successful reaction conditions in every case interrogated.

1. Amide Coupling

In the first conversation, we requested the generation of a reaction array to optimize an amide coupling between 2-methylbenzoic acid (**1**) and *p*-toluidine (**2**) to form amide **3** (73-Fig. 2).

Prompt: “Generate a reaction array with 3 coupling agents, 2 bases, 2 solvents and 2 additives to optimize an amide coupling between 2-methylbenzoic acid and p-toluidine. Add each reagent into a single table, where each row has a column for the reagent class and a column for the reactant name. As well, include specific recommendations and values for other reaction parameters for this screen. Respond succinctly. Finally, cite some literature references from which this screen was designed with DOIs.”

ChatGPT Response:

Reagent Class	Reactant Name
Coupling Agent	HATU
Coupling Agent	EDC
Coupling Agent	PyBOP
Base	DIPEA
Base	TEA
Solvent	DMF
Solvent	DCM
Additive	HOAt
Additive	DMAP

For reaction parameters, some specific recommendations for this amide coupling reaction could include:

- Temperature: 0-25°C (refrigeration)
- Time: 2-24 hours
- Stoichiometry: 1.2-2.0 equivalents of coupling agent, 1.0-1.5 equivalents of base, and 1.0-1.2 equivalents of amine relative to carboxylic acid
- Concentration: 0.1-0.5 M
- Inert atmosphere: nitrogen or argon

References:

- Li, H.; Banerjee, S.; Liao, L.; Pei, X.; Chen, X.; Miao, Q.; Huang, W. An efficient one-pot approach for the synthesis of unsymmetrical ureas via domino reaction of isocyanides with amines using EDC as the coupling agent.
- Kamber, D. N.; Jeon, J.; Lee, S. H.; Kim, S. K. Efficient N,N'-dimethylformamide catalyzed synthesis of amides from carboxylic acids using HATU as a coupling reagent.
- Hwang, S. H.; Park, S. H.; Yang, J. E.; Yoon, Y. J.; Kim, D. S.; Lee, D. W. High yield synthesis of peptide-PEGylated sunitinib via EDC-mediated conjugation.

DOIs have been omitted from the shown response as the DOI and citation references did not match the article titles. Furthermore, many of the references are, to the best of our knowledge, not real. This is because ChatGPT is a language model rather than a knowledge model, which has been reported to hallucinate citations. Despite this, the array design is reasonable. Further examples of “references” for each subsequent array design are included in the Supporting Information. We also note that the model struggled to accurately recreate the SMILES of compounds when asked (see Supporting Information), but structures were derived from the IUPAC name output, which were always reasonable.

The experimental execution of this reaction array precisely followed the suggestions from ChatGPT with resultant reaction metadata such as concentrations, volumes and well locations designed by phactor (74-Fig. 3). Two-thirds of the array produced hits with moderate assay yield, while reactions 1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) as a coupling agent failed entirely. Well D5, the best hit using benzotriazol-1-yloxytripyrrolidinophosphonium hexafluorophosphate (PyBOP) and N,N-diisopropylethylamine (DIPEA) with (4-dimethylaminopyridine) DMAP in dichloromethane (DCM), was scaled up and resulted in an 94% isolated yield of product **3**.

2. Amide Coupling on Complex Molecule Sitagliptin

Next, we explored how the conversation can be continued with a more complex substrate for the amide coupling. In the same dialog, we asked ChatGPT to refer to the previous question but to instead optimize the amide coupling between sitagliptin (**4**) and carboxylic acid **2** to form amide **5** (75-Fig. 4).

Prompt: “Generate a similar reaction assay that instead optimizes the coupling of Sitagliptin with 2-methylbenzoic acid. Respond in the same format.”

ChatGPT Response:

Reagent Class	Reactant Name
Coupling Agent	HBTU
Coupling Agent	TBTU
Coupling Agent	PyAOP
Base	DIEA
Base	TBD
Solvent	DMF
Solvent	DCE
Additive	DMAP
Additive	HOAt

For other reaction parameters, the following values are recommended:

- Temperature: 0-25°C (refrigeration)
- Time: 2-24 hours
- Stoichiometry: 1.2-2.0 equivalents of coupling agent, 1.0-1.5 equivalents of base, and 1.0-1.2 equivalents of amine relative to carboxylic acid
- Concentration: 0.1-0.5 M
- Inert atmosphere: nitrogen or argon

Again, the reaction array generated by ChatGPT was executed using the recipe designed by phactor (76-Fig. 5) Seemingly, the results of this array performed better than the original amide coupling, with only 3 failed hits, but with lower overall yields. The best hit was well A6, which produced **5** in 62% yield when using 7-Azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (PyAOP), triazabicyclodecene (TBD), and 1-Hydroxy-7-azabenzotriazole (HOAt) in dimethylformamide (DMF). When scaled up, the reaction yielded 62% isolated product.

3. Suzuki Coupling

We then attempted to optimize a Suzuki coupling reaction. A new chat was initiated with ChatGPT to propose new inputs for phactor. Here we asked for an optimized reaction array to form biaryl **8** from the Suzuki coupling between boronate **6** and chloride **7** (77-Fig. 6). A typical ligand versus catalyst screen is generated, as requested. The model performed well at pulling ligands and catalysts from the literature corpus, as well as generating additional parameters needed for the assay. These probabilistic proposals for ligands and reagents are the perfect input for phactor, which then enables systematic testing of all combinations.

Prompt: "Generate a 6 ligand by 4 catalyst reaction array to optimize an Suzuki coupling between o-tolylboronic acid and 1-chloro-4-methylbenzene. Separate each parameter into a single table, where each row has a column for the reagent class and a column for the reactant name. As well, include specific recommendations and values for other reaction parameters for this screen. Respond succinctly. Finally, cite some literature references from which this screen was designed with DOIs."

ChatGPT Response:

Reagent Class	Reactant Name
Ligand	XPhos
Ligand	SPhos
Ligand	RuPhos
Ligand	BrettPhos
Ligand	Buchwald ligand
Ligand	DavePhos
Catalyst	$Pd_2(dba)_3$
Catalyst	$Pd(OAc)_2$
Catalyst	$PdCl_2$
Catalyst	$Pd(dppf)Cl_2$
Substrate	o-tolylboronic acid
Substrate	1-chloro-4-methylbenzene

Conditions: The reaction should be performed under nitrogen atmosphere, and solvents such as DMF, DMSO, or toluene could be used. The reaction should be performed at 100-120°C for 12-24 hours, with a 1:1 ratio of the two substrates, and 2 mol% of Pd catalyst. The ratio of the Pd catalyst to the ligand should be 1:2.

We note that the model does not discuss the addition of a base, so we directly ask in the same conversation:

Prompt: “What base should I use?”

The model responds: A base is typically not necessary for Suzuki coupling reactions since the reaction proceeds via a transmetalation step rather than a deprotonation step. However, if a base is needed, a common choice is potassium carbonate (K_2CO_3) or cesium carbonate (Cs_2CO_3) at a loading of 1.2-2 equiv.

The Suzuki coupling between **6** and **7** proved to be more challenging for ChatGPT, with slightly over half the plate failing to produce a hit (78-Fig. 7). $Pd(dppf)Cl_2$ performed consistently well and DavePhos produced a hit whenever used. The unusual combination of $Pd(dppf)Cl_2$ and BrettPhos, which has not been previously reported, to our knowledge, generated the best hit with a 25% product/internal standard integration ratio. When scaled up, 66% of the product was isolated. Interestingly $Pd_2(dba)_3$ has been previously reported to yield over 95% with these substrates with when run with a bisimidazolyl ligand[232] or the WK-Phos ligand.[233] These reports were seemingly missed by the AI.

We noted that ChatGPT did not mention the addition of water as an additive in the Suzuki coupling reaction conditions, which is typically required to activate boronic acids.[234] When asked, the model responds that water may or may not be beneficial to the yield of the reaction (see Supporting Information). We repeated the design above with base added as an aqueous solution to compare the results, finding that the yields are significantly improved across the plate (79-Fig. 8).

4. Buchwald-Hartwig Coupling

Finally, we explored the ability for the model to optimize a Buchwald-Hartwig coupling. With this transformation, the interplay of ligands and base effects is often more complex. Again, a new conversation was started. In this final example, we asked the model to generate a three factorial screen to optimize the C–N coupling between aryl chloride **9** and aniline **2** to form compound **10** (80-Fig. 9).

Prompt: “Generate a 3 ligand by 4 catalyst by 2 base reaction array to optimize an Buchwald-Hartwig coupling between 1-chloro-2-methylbenzene and p-toluidine. Add each reagent into a single table, where each row has a column for the reagent class and a column for the reactant name. As well, include specific recommendations and values for other reaction parameters for this screen. Respond succinctly. Finally, cite some literature references from which this screen was designed with DOIs.”

ChatGPT Response:

Reagent Class	Reactant Name
Ligand	XPhos
Ligand	SPhos
Ligand	RuPhos
Catalyst	$Pd_2(dba)_3$
Catalyst	$Pd(dppf)Cl_2$
Catalyst	$P(t\text{-}Bu)_3$
Catalyst	$Pd(OAc)_2$
Base	K_3PO_4
Base	NaO^tBu

Other recommended reaction parameters:

- Solvent: DMF
- Temperature: 100°C
- Reaction time: 24 hours
- Stoichiometry: 1:1 ratio of 1-chloro-2-methylbenzene and p-toluidine
- Concentration: 0.1 M

We note the incorrect output “$P(t\text{-}Bu)_3$”, which we manually corrected to the palladium complex $Pd(Pt\text{-}Bu_3)_4$ in our experimental execution. We asked additional questions to clarify the experimental details.

Prompt: What catalyst, ligand, and base concentrations should I use for this reaction?

ChatGPT Response: The specific concentrations of catalyst, ligand, and base to use for the reaction depend on various factors such as the desired reaction rate, reaction yield, and the specific reactants being used. However, as a general guideline, a common catalyst concentration range for Buchwald-Hartwig couplings is 0.5-5 mol%. Ligand concentrations can range from 1-10 mol%. Base concentrations can range from 0.5-2 equiv. It's important to note that these concentrations may need to be optimized for specific substrates and conditions.

Prompt: How about the order of addition for the reagents?

ChatGPT Response: The order of addition for the reagents can also be important for optimizing the reaction conditions. In a typical Buchwald-Hartwig coupling, the base is usually added first to the reaction mixture, followed by the catalyst and the ligand. The organic halide and the amine are then added to initiate the reaction. However, depending on the specific reactants and conditions, the order of addition may need to be adjusted to achieve the desired reaction outcome. Therefore, it's always a good idea to consult the literature or perform some preliminary experiments to determine the optimal order of addition for the specific reaction being performed.

Prompt: How would I quench and work up this reaction?

ChatGPT Response: After the Buchwald-Hartwig coupling reaction is complete, it can be quenched by adding a quenching agent to the reaction mixture. Common quenching agents include ammonium chloride or hydrochloric acid. The choice of quenching agent may depend on the specific reaction conditions and the nature of the products.

To work up the reaction, the mixture is typically cooled and then diluted with an appropriate solvent such as water or ethyl acetate. The reaction mixture is then filtered to remove any solid catalyst or byproducts. The organic layer is separated and washed with water or brine to remove any remaining inorganic salts. The organic layer is then dried over anhydrous sodium sulfate or magnesium sulfate and concentrated under reduced pressure to yield the crude product. Purification of the product can be done using techniques such as column chromatography or recrystallization.

We note that the quench of the reaction should be performed after cooling for safety. A follow up question was asked to interrogate if ChatGPT would identify the single best answer from the array performed with ChatGPT and phactor.

Prompt: What do you think the single best combination of ligand and catalyst would be for this reaction, given the reaction array you generated earlier? Respond with a single best guess.

ChatGPT Response: It's difficult to predict the single best combination of ligand and catalyst for the Buchwald-Hartwig coupling reaction between 1-chloro-2-methylbenzene and p-toluidine without performing experiments or conducting computational studies. However, based on the reaction array I provided earlier, a possible combination of a ligand and a catalyst that could work well together for this reaction is SPhos as the ligand and $Pd_2(dba)_3$ as the catalyst. This is just a best guess, and it's important to note that the optimal combination of ligand and catalyst may depend on various factors, such as the specific reaction conditions, and could require further optimization to achieve the best results.

This reaction array was executed experimentally with the exact specifications given by ChatGPT and exact array design given by phactor. The best hit was repeated on a 0.2 mmol scale. (81-Fig. 10). The array produced many hits with moderate yield with respect to internal standard. Furthermore, the absence of impurities and byproducts simplified purification. Well B5 was scaled up and resulted in an 84% isolated yield of product **10**. The best conditions for coupling this substrate pair reported in the literature yielded 98% of **10** when using a N-heterocyclic carbene-palladium(II) chloride complex and potassium tert-butoxide as reaction conditions.[235]

The model's output seemed to select catalysts and ligands that have performed well in the literature but may be less aware of how they are used together. Moreover, phactor necessarily combines the answers from ChatGPT, which were not likely given with an understanding of combinatorial enumeration. As such, the resultant array includes unusual combinations of palladium sources and ligands, such as $PdCl_2$(dppf) and RuPhos wherein two phosphine ligands are present. Furthermore, DMF and DMSO are less typical than ethereal solvents for this reaction. Despite this, the assay still produced positive results.

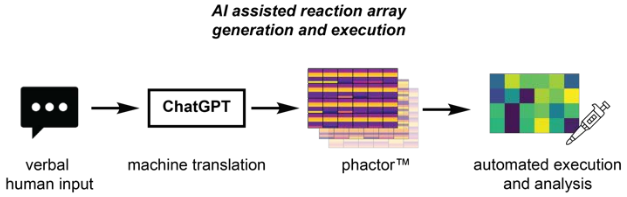
The combination of Pd(dppf)Cl_2 and RuPhos is a surprising result, since there is conceivably an equilibrating mixture of Pd(dppf)L_n and Pd(RuPhos)L_n complexes. We suspect the proposal to mix these compounds arises as an artifact of making a combinatorial array (phactor) out of popular Pd complex and ligand choices from the literature (ChatGPT). Nonetheless, the observation that this "cocktail" of ligands was the most productive result could be supported by related reports of "cocktail" catalysis in the Buchwald-Hartwig coupling.[236-239]

Discussion

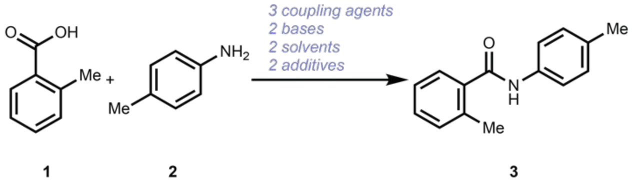
With several reaction arrays executed and an understanding of effective prompt engineering ChatGPT needs to generate effective arrays, we sought to integrate ChatGPT directly into the interface of phactor. The design goal was to allow chemists to generate relevant reagents for specific substrate pairs and a given transformation, even if they were non-experts. The GPT-3.5 API provided by OpenAI, enabled integration on the “chemicals” tab of phactor (82-Fig. 11). Here, the user selects substrates already added to the design stage in phactor and describes the desired reaction. Given the user inputs, the prompt template is automatically filled with ChatGPT’s proposals. The user retains the ability to modify the prompt as desired to add or lessen constraints. Once the prompt is submitted, the AI responds with a given number of reagents specified by the user. These compounds can be instantly added to phactor with a molecular weight and SMILES by clicking the “add” button, as long as they appear in PubChem.

Conclusion

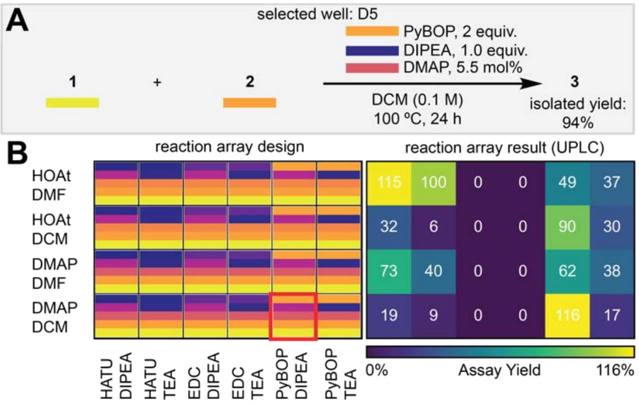
The software ChatGPT was utilized to generate reaction array designs with the HTE design software phactor for several popular reactions. The desired product was observed in every instance tried on the first attempt, and isolated yields ranged from 62%–94%. This initial study showcases the impact that large language model predictions can have on chemical synthesis when coupled to HTE.


AI assisted reaction array generation and execution
ChatGPT
verbal human input
machine translation
phactor™
automated execution and analysis

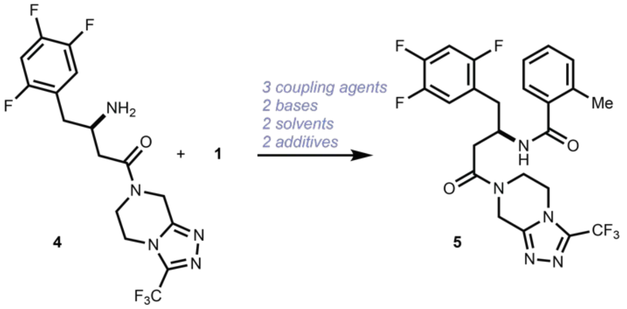
72-Fig. 1 | Overview of the ChatGPT to phactor workflow. Verbal input is given by a human to have ChatGPT generate a reaction array design for a particular coupling and substrate pair. The output can be fed directly into phactor, creating an assay recipe to be executed robotically or manually.



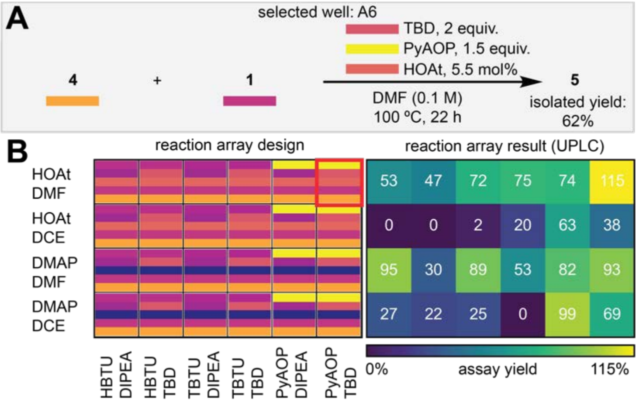
73-Fig. 2 | The amide coupling between **1** and **2** ChatGPT was asked to optimize. The reagent classes are specified in the prompt, but the specific species are generated by ChatGPT.



74-Fig. 3 | Executed reaction array and UPLC assay results of the screen designed by ChatGPT to perform the amide coupling between **1** and **2**. The top performing reaction condition was repeated on 0.2 mmol scale to yield **3** in 94% isolated yield.

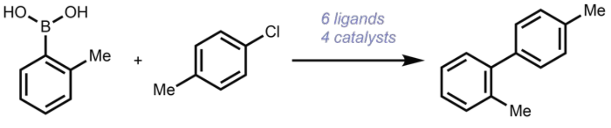


75-Fig. 4 | The amide coupling between **4** and **1** ChatGPT was asked to optimize. The resultant design is different than with the original simpler substrates.



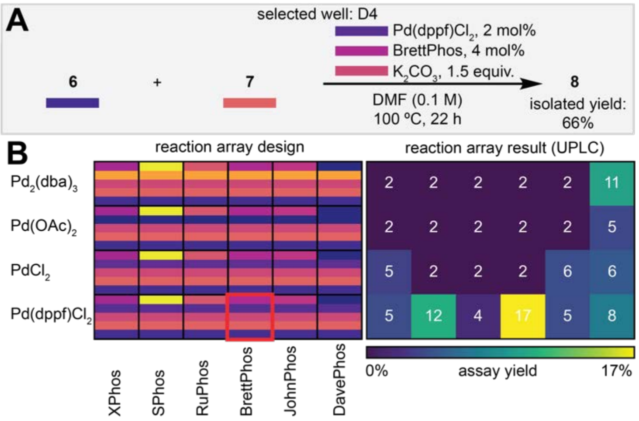
	HBTU DIPEA	HBTU TBD	TBTU DIPEA	TBTU TBD	PyAOP DIPEA	PyAOP TBD
HOAt DMF	53	47	72	75	74	115
HOAt DCE	0	0	2	20	63	38
DMAP DMF	95	30	89	53	82	93
DMAP DCE	27	22	25	0	99	69

76-Fig. 5 | Results of the amide coupling to produce **5** designed by ChatGPT and phactor when executed experimentally. The top performing reaction condition was repeated on 0.4 mmol scale to yield **5** in 62% isolated yield.



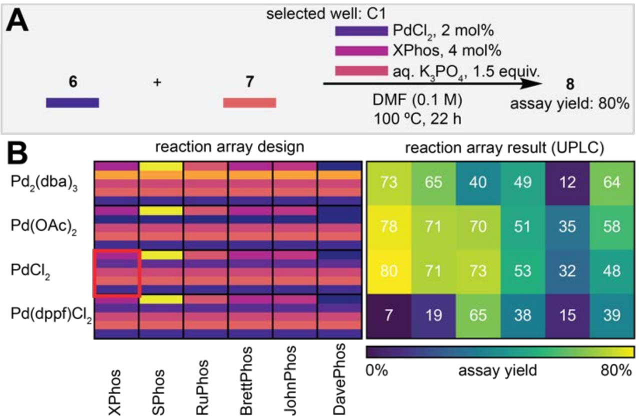
6 **7** **8**

77-Fig. 6 | The Suzuki coupling between **6** and **7** ChatGPT was asked to optimize.

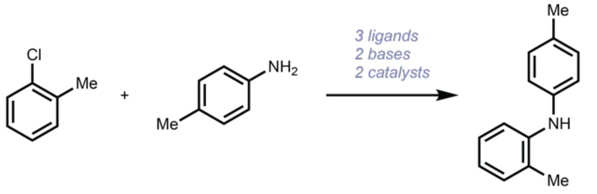


	XPhos	SPhos	RuPhos	BrettPhos	JohnPhos	DavePhos
Pd$_2$(dba)$_3$	2	2	2	2	2	11
Pd(OAc)$_2$	2	2	2	2	2	5
PdCl$_2$	5	2	2	2	6	6
Pd(dppf)Cl$_2$	5	12	4	17	5	8

78-Fig. 7 | Results of the Suzuki coupling reaction array designed by ChatGPT and executed with phactor. The top performing reaction condition was repeated on 0.2 mmol scale to yield **8** in 66% isolated yield.

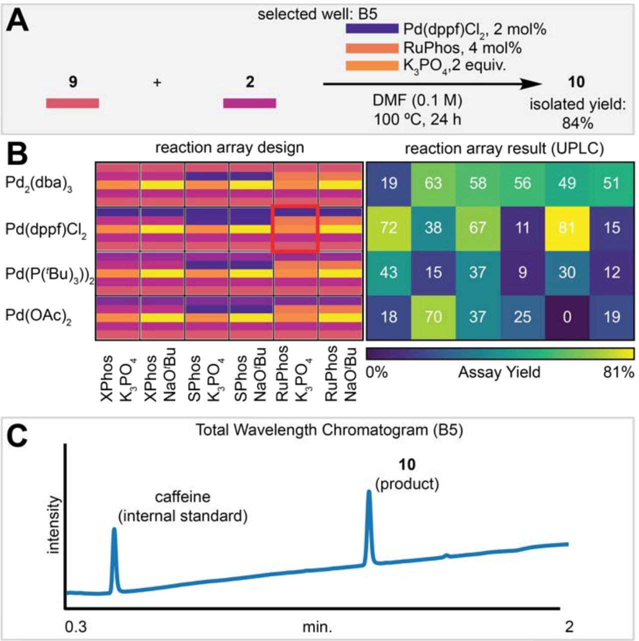


79-Fig. 8 | Results of the repeated Suzuki coupling reaction array designed by ChatGPT when 20% water is added.



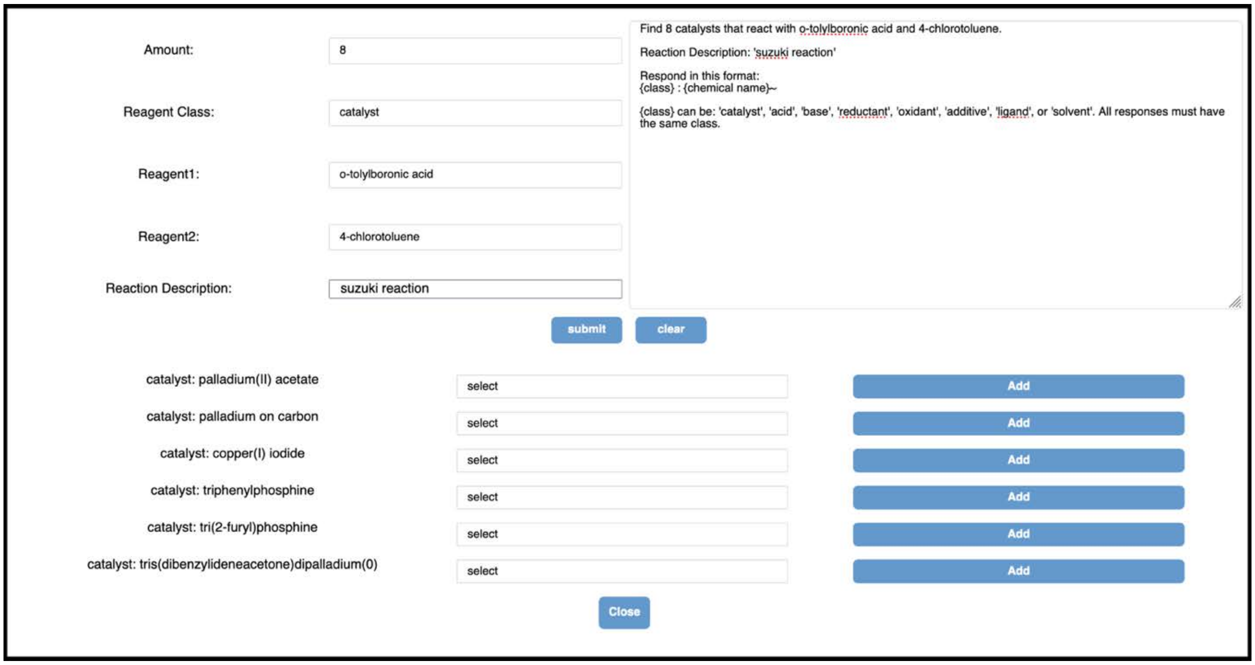
9 **2** **10**

80-Fig. 9 | The Buchwald-Hartwig coupling ChatGPT is asked to optimize.



	XPhos K_3PO_4	XPhos NaO^{t}Bu	SPhos K_3PO_4	SPhos NaO^{t}Bu	RuPhos K_3PO_4	RuPhos NaO^{t}Bu
$Pd_2(dba)_3$	19	63	58	56	49	51
Pd(dppf)Cl_2	72	38	67	11	81	15
Pd(P(^{t}Bu)$_3$))$_2$	43	15	37	9	30	12
Pd(OAc)$_2$	18	70	37	25	0	19

81-Fig. 10 | (A) 0.2 mmol scale-up result of the best hit from the reaction array proposed by ChatGPT-phactor. The combination of palladium catalyst complex Pd(dppf)Cl_2 and RuPhos gave the best result. (B) The reaction array design and results. (C) Crude UPLC trace for selected well B5.



Amount: 8

Reagent Class: catalyst

Reagent1: o-tolylboronic acid

Reagent2: 4-chlorotoluene

Reaction Description: suzuki reaction

Find 8 catalysts that react with o-tolylboronic acid and 4-chlorotoluene.

Reaction Description: 'suzuki reaction'

Respond in this format:
{class} : {chemical name}~

{class} can be: 'catalyst', 'acid', 'base', 'reductant', 'oxidant', 'additive', 'ligand', or 'solvent'. All responses must have the same class.

catalyst: palladium(II) acetate — select
catalyst: palladium on carbon — select
catalyst: copper(I) iodide — select
catalyst: triphenylphosphine — select
catalyst: tri(2-furyl)phosphine — select
catalyst: tris(dibenzylideneacetone)dipalladium(0) — select

82-Fig. 11 | Integrated GPT interface in phactor. Based on the substrates selected by the chemist and a description of the desired transformation, chemicals suggested by the AI can be automatically added into the reaction array design by phactor.

Weighted Reaction Fingerprints for Visualizing Reactivity Cliffs and Generality

In this project, a simple algorithm was developed to visualize reaction datasets. By simply adding the fingerprints of the various components in a reaction for each reaction in a dataset, the resultant matrix could be embedded into a two-dimensional manifold. Multiplying certain components by a scaling factor enabled manipulation of the manifold. With this technique, reactivity cliffs, generality, and other reactivity trends are quickly identified.

Abstract

Visualization of reaction space is a critical step in improving human understanding of bulk chemical reaction data. We present weighted reaction fingerprinting, a simple method to rapidly analyze and evaluate the results of large reaction corpuses which we have found to be helpful in the analysis of high-throughput experimentation (HTE) campaigns. Weighted reaction fingerprints can be utilized to rapidly identify successful and failing conditions and systems for chemical transformations. Reactions are encoded into a standard "reaction anatomy" comprising popular reagent components such as nucleophile, electrophile, catalyst, ligand, and solvent. Each reaction is converted into a typical fingerprint matrix and multiplied by a weight vector to generate the weighted reaction fingerprint. These fingerprints are fed into dimensionality reduction algorithms such as principal component analysis (PCA) or t-stochastic neighbor embedding (t-SNE) to create visualizable 2-D manifolds that visually reveal reactivity trends such as catalysts with high substrate generality or reactivity cliffs.

Introduction

Statistical modeling of chemical reactions requires large datasets of chemical reactions from high-throughput experimentation (HTE) campaigns or curations of published data. Our group has been using HTE[1, 35, 36, 102, 103, 108, 110, 134, 136, 140-142, 181, 240] to explore chemical reactivity. The primary goals of HTE analysis include statistical data profiling, allowing chemists to rapidly identify best, worst, and average performing reaction conditions, to understand which reaction conditions work best for certain substrate pairs, and to inspire the generation of ideas for new experimental space to explore. HTE reaction array visualizations as simple rectangular heatmaps are common ways to show the physical location on a wellplate where each reaction took place, but these rectangular heatmaps cannot easily communicate trends such as catalyst generality against substrate pairs or reactivity cliffs. Similarly, common machine learning or artificial intelligence algorithms for reaction informatics are opaque and challenge human-interpretable analyses. As reaction data[136] is deposited into centralized databases,[149] visual tools that communicate reaction performance in HTE campaigns, or related reaction datasets, are needed to identify patterns and trends.

Machine-readable molecular representations[241] are critical in understanding trends of chemical reactivity from large datasets. Graph representations[135] and molecular fingerprints have been used for the prediction of chemical properties, similarity searching, and structure optimization. In a reaction context, embeddings of the molecules from chemoinformatic, chemometric and quantum descriptors in conjunction with reaction outcomes are used to train models to predict reactivity and elucidate mechanisms.[124, 125] In particular, fingerprinting methods, such as the Morgan instantiation[242] of extended connectivity fingerprints (ECFP),[243] provide a fast and computationally non-intensive method to analyzing chemical data and the influx of reported reaction data in a standardized format. In recent literature, fingerprinting methods for reactions have included the concatenation of reagent fingerprints,[161, 244] feature binning fingerprints,[245] and reaction difference fingerprints.[22]

Herein, we demonstrate the utility of weighted reaction fingerprinting (83-Fig. 1) – a simple abstractable method applicable to any large reaction corpus stored in a standardized format. Reaction discovery campaigns are analyzed to evaluate the results of multiplexed reaction arrays and large reaction datasets. We show that this visual tool can be used to analyze and draw conclusions from large corpuses of reaction data.

Methods

We introduce reaction array fingerprinting with a dataset consisting of 2,786 Suzuki reactions.[134] Each reaction is composed of an electrophile, nucleophile, catalyst, base, and solvent. In 84-Fig. 2, the Morgan Fingerprint with radius 4 and 2,048 bits was calculated for each substance, then summed such that the five 2,048-bit binary fingerprints for each reaction become a single 2,048-bit count reaction array fingerprint – a one-dimensional array representation of the coupling. This array of 2,786 reaction fingerprints was fed into a t-SNE algorithm with no hyperparameter tuning to generate 84-Fig. 2A, where reaction points are colored by the product yield for the reaction – high, medium, low, and 0% – based on product/internal standard (PROD/IS) integrations. The observed clusters are chemically distinguishable, with each cluster composed of similar or identical topological fragments in the reaction mixture. This reaction fingerprinting method (Figure S1) contrasts with currently published methods that include reagent concatenation,[161, 244] reaction feature binning,[245] and reaction difference fingerprints,[22] although weighing can at times be used in other reaction fingerprints to a similar effect. Results of different dimensionality reduction algorithms such as PCA, UMAP, and SOMs as well as different fingerprint representations, including one-hot encoding, are shown in the Supporting Information (Figure S2). Effects of modifying the perplexity of the t-SNE for this analysis is shown in Figure S3 and the same analysis is shown with molecular access system (MACCS) fingerprints[246] in Figure S4. Results of unsupervised versus supervised UMAP for the dataset is shown in Figure S5, and comparisons of weighted fingerprints against concatenated and difference fingerprints are shown in Figure S6. The six plots of 84-Fig. 2C display the same embedding shown in 84-Fig. 2B with six alternative color scales representing different features. The first plot's points are colored by the exact product/internal standard value for each reaction as calculated in the dataset. The remaining five plots are each colored by reagent per specific reagent class as defined in the template (this reaction dataset consists entirely of electrophile, nucleophile, catalyst, base, and solvent components.)

An example of a reaction cluster identified from the manifold is shown in 84-Fig. 2D. All reactions in the dataset using bromide electrophile **1** and boronate nucleophile **2** exist within this cluster. It is rapidly identified that this substrate pair was tested with 16 different catalysts and six different bases. 84-Fig. 2E showcases a simple extension of the fingerprinting algorithm using the data from this cluster. When summing the fingerprints of the individual components, a weight factor can be multiplied into a reagent's fingerprint to influence the clustering within the manifold. The 96 reactions between **1** and **2** were encoded as reaction fingerprints in two different formats. Once where the catalyst fingerprint was weighed by a factor of three, and again where the base fingerprint was weighed by a factor of three. These two datasets were then fed into the t-SNE reduction algorithm, and three plots colored by product/internal standard integrations, catalyst, and base for each of the two datasets are displayed in 84-Fig. 2E. As revealed by the color encoded reagents, elevating the catalyst weight produces manifolds with catalyst clusters, and similar behavior is seen with base clusters when elevating the base weight. This fingerprinting method can be used to identify reagents with high generality. For the same dataset, weighing is used to cluster electrophile and nucleophile pairs (Figure S7). When weighing the catalyst component, RuPhos Pd G3 was found as a high generality palladium catalyst and resulted in a 0.49 PROD/IS integration across 20 electrophile/nucleophile pairs. Similarly, 2-tert-butyl-1,1,3,3-tetramethylguanidine was found as a high generality base and a mixture of water and DMSO was found as a high generality solvent that worked consistently across the substrate scope used in the campaign.

Results and Discussion

85-Fig. 3 shows a reaction landscape from the discovery of a *sp*^{3}–*sp*^{3} deaminative–decarboxylative carbon–carbon cross coupling reaction over multiple HTE experiments. In this study, the fingerprints of all acid electrophiles were multiplied by 3 ($w_e = 3$) before summing with the fingerprints of the other reagents. This trivial modification resulted in clusters of reactions based on electrophile.

As shown in 85-Fig. 3, with the acid electrophiles having an elevated weight, the clusters within the manifold are split into three distinct regimes reflecting the stages in the discovery and development of the chemistry. The discovery campaign began with using the free acid **4** as the model substrate (85-Fig. 3 – Initial Reactions) but after a limit to the reactivity was realized, efforts moved to the NHPI activated acid **5** (85-Fig. 3 – Methodology Development). In this case, the activated acid electrophile NHPI ester N-Boc-proline was used as the model substrate for the reaction, resulting in a large cluster with few substrates and many conditions. Thus, most of the reactions in this dataset fall within the NHPI activated (**5**) and free acid (**4**) N-Boc proline clusters colored in black (85-Fig. 3B – Initial Reactions) and yellow (85-Fig. 3B – Methodology Development) respectively. In the PROD/IS manifold shown in 85-Fig. 3C, a direct path can be followed as experiments drive the product output from 0% to close to 100% within the NHPI activated N-Boc proline cluster (Methodology Development). Once ideal conditions were developed for N-Boc proline, this reaction system was tested with a variety of other acid electrophile substrates, creating a tight cluster of many electrophile acids and few conditions (**6**-**26**, Scope Exploration). The efficacy of this system on these substrates is shown in the t-SNE with clusters of various acid electrophile substrates forming in different locations, each with their own PROD/IS distributions. Average PROD/IS results of each acid electrophile tested are shown below the plots in 85-Fig. 3C as well as the number of reactions they were tested in.

In 86-Fig. 4, a case study analyzing the Suzuki dataset reveals reactivity cliffs – clusters that contain some reactions that fail and some that are successful, indicating minor changes to the system that cause the reaction to “flip” on or off.[247] An electrophile nucleophile substrate pair is identified to have a reactivity cliff in 86-Fig. 4A. The plots in

86-Fig. 4A are reaction array fingerprints only containing the fingerprints for the electrophile and nucleophile to emphasize focus on the substrate flexibility of the Suzuki reaction as opposed to the specifics of the reagents. Using this technique, a cluster containing a single nucleophile/electrophile pair was identified. On further analysis of this cluster with reactions that use electrophile **27** and nucleophile **28** to form **29**, it is revealed that this substrate pair was tested with four different solvents, six bases, and eight catalysts. In 86-Fig. 4B, we plot the reaction array fingerprints for these reactions, producing clean clusters that separate all components and producing a humanly interpretable explanation of the behavior behind the reactivity. Since experimental chemistry is rife with reactivity cliffs, oftentimes as subtle as a switch in solvent or order of reagent addition, it is critical to be able to visualize and interpret this behavior. From the color-coded solvent plot, it is clear that the reactivity of the substrate pair **28** and **29** is controlled by the solvent system used. The manifold directly identifies failing and working solvents when traversing the space from 1:9 ^{t}AmOH:NMP to 1:3 water:NMP. The addition of water in the solvent system is found to be critical in achieving desired reactivity as shown by another reactivity cliff between 1:3 water:NMP and pure NMP. A chi-squared analysis is reported in the Supporting Information to validate this finding (Figure S8). Thus, it can be rapidly identified which solvent systems poison the reaction even though a variety of reagents are being changed. This analysis is compared to boxplots of the solvent system used (Figure S9).

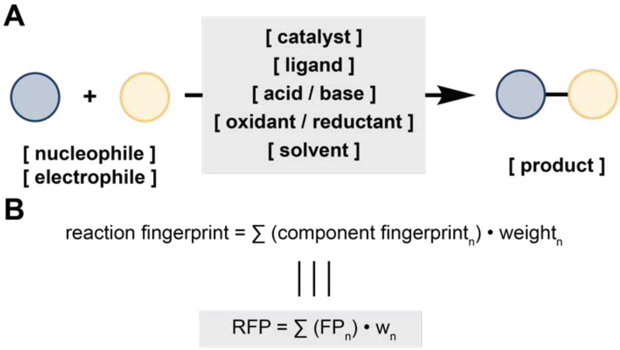
The creation of reaction array fingerprint manifolds is automatable. When using a HTE management system such as phactor or obtaining machine readable reaction data in a standardized format, reaction arrays can be rapidly analyzed using this method. In 87-Fig. 5, we show the automated creation of six reaction fingerprint t-SNEs utilizing output files procedurally generated from phactor. Hyperparameters including t-SNE perplexity and reagent weights (w_x) were optimized to best illuminate reactivity trends (see Supporting Information). In all cases, specific reagent classes were clustered. For instance, there are four clusters in 87-Fig. 5A representing the four ligands used in the reaction array which couples **30** and **14** to form **31**. These clusters are split into two subclusters, each representing one of the two reductants used. These clusters are then further split into three column-shaped clusters that represent the three catalysts and each of these columns consist of four points each corresponding to one of the four additives in the reaction array. Four clusters were formed in 87-Fig. 5B which shows 12 electrophile acids and 8 nucleophile amines, producing a successful reaction that uses **32** and **5** to form **33**. Each of the 12 acids formed a cluster in the manifold. In 87-Fig. 5C eight ligands and 12 catalysts were used to make **36** from **34** and **35**. The t-SNE clustered each reaction by ligand, resulting in eight clusters. In substrate scope ultraHTE experiments shown in 87-Fig. 5D, the t-SNE is clustered by the eight diazoniums used in the screen. In 87-Fig. 5D the reaction where **39** was formed from **37** and **38** is highlighted in the manifold. Finally, in the ultraHTE direct-to-biology assay shown in 87-Fig. 5E, a cluster is formed for each of the 80 amines used in the reaction array. The reaction which amide inhibitor **42** is generated from the coupling of **40** and **41** is identified in the t-SNE. This visual format is compared against pivot table heatmaps in Figure S10.

Reaction fingerprints can be used to rapidly identify the best conditions for a given substrate pair and transformations. As part of our studies, we have been exploring mild benzoic acid decarboxylation methods. We discovered simple reaction conditions for a decarboxylative–deaminative amine–acid *sp*^{2}–*sp*^{2} C–C coupling reaction between quinidinyl acids and amine diazoniums. This transformation proceeds when both substrates are added to solvent and irradiated by light, based on a recent report on irradiating diazoniums in methanol to achieve various chemistries.[191] Ultimately 424 experiments were performed over a series of 24 and 96-well reaction arrays (88-Fig. 6A).

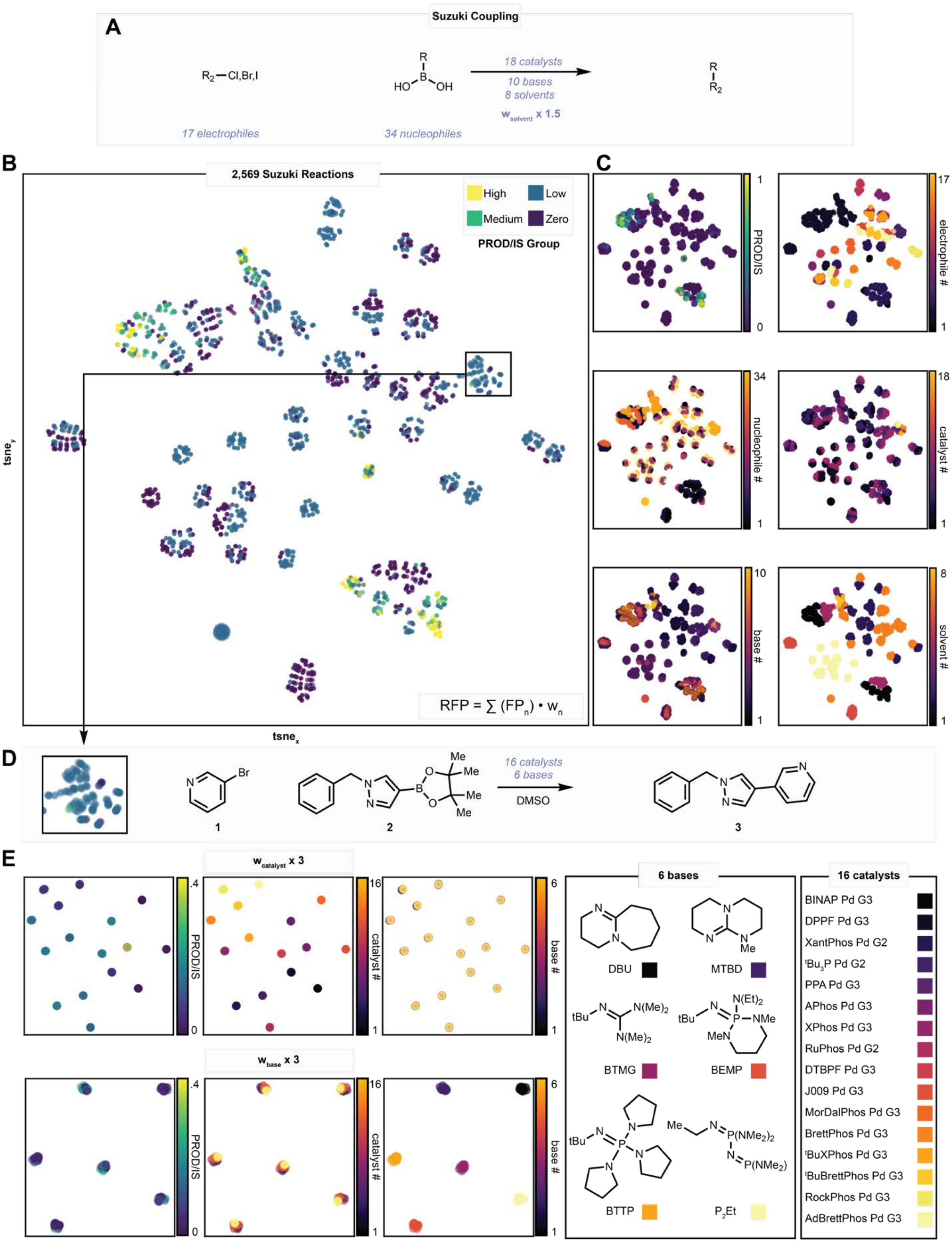
Isoquinolines **43**, **46**, **47**, and **48**, and diazonium salts **44**, **49**, **50**, **51**, **52**, and **53** produced appreciable yields (measured by product/internal standard integrations) in methanol, DMSO, and a 1:1 mixture of methanol and DMSO (88-Fig. 6B) to form biaryl products such as **45**. An array performed in the campaign is shown in 88-Fig. 6C, where the four acids and six diazonium salts in 88-Fig. 6B were irradiated in methanol under UV light for 18 hours. For the reaction between **43** and **49** to form **54**, we are able to rapidly identify from the HTE results and reaction fingerprint arrays (88-Fig. 6D) that the reaction works best in blue light and with 50:50 DMSO:methanol, as we can directly compare the PROD/IS result against the reaction performed in other regimes.

Conclusion

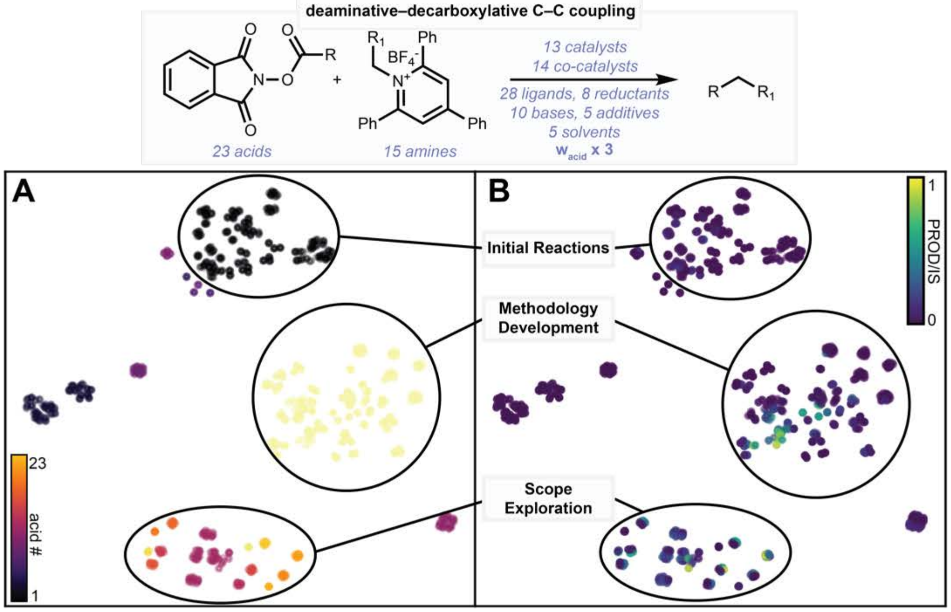
Weighted reaction fingerprinting is an easy-to-perform method for the analysis of massive reaction datasets. The algorithm allows chemists to rapidly navigate through large collections of reaction data. Reaction embeddings can be easily optimized and modified by changing the reagent weights.

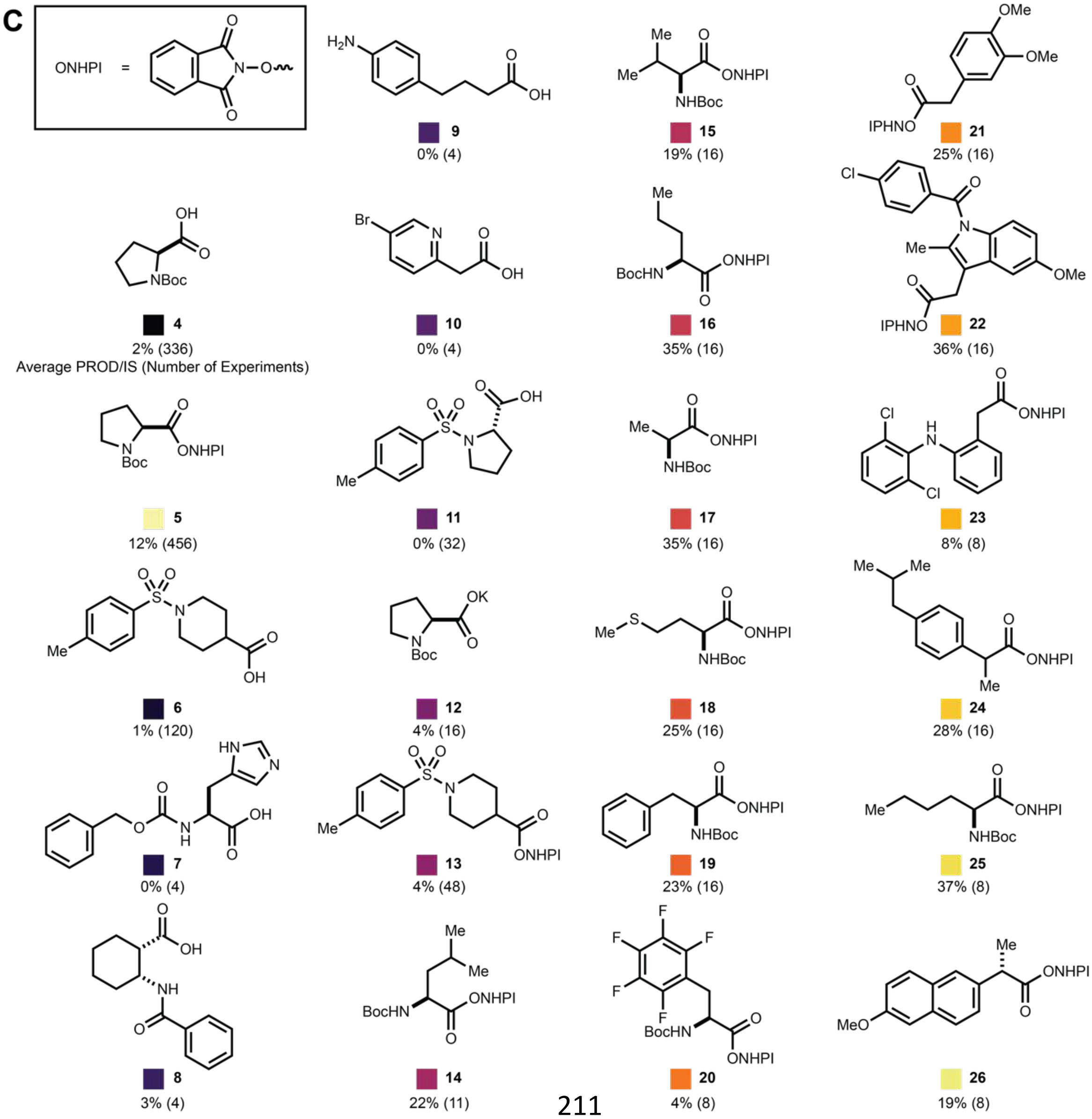


83-Fig. 1 | A) Anatomy of a reaction. Electrophiles and nucleophiles are arbitrarily defined as the reacting partners that form the product. B) The reaction fingerprint can be calculated by taking the product between a matrix of component fingerprints and a vector of weight arrays. Any fingerprint or feature vector can be utilized as long as the vector lengths for each reaction components are equal. Reaction component weights (w_n) are initialized at 1 and can be set to any value by the user. A component * weight product is added to the sum for each reaction component *n*.

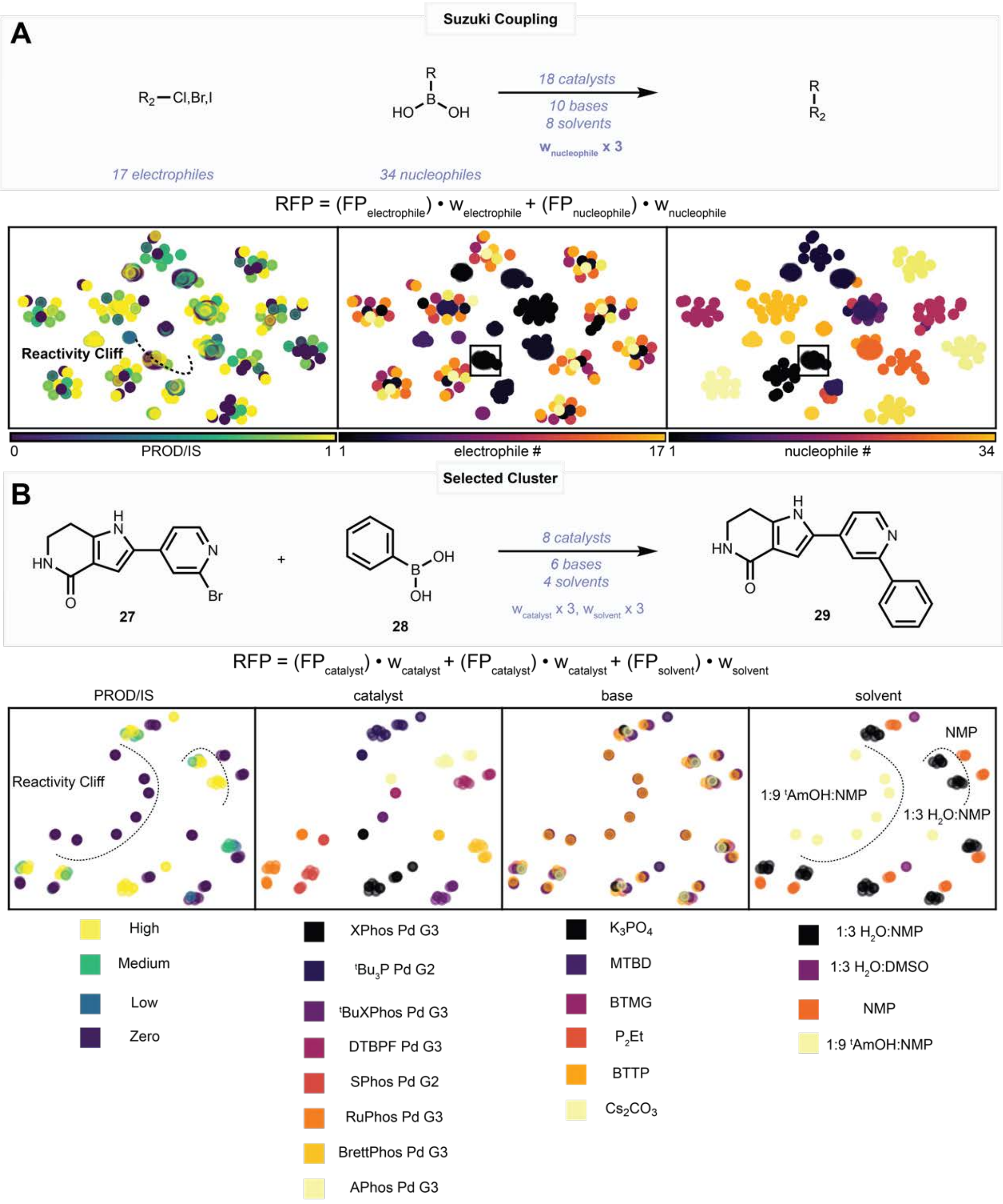

A
Suzuki Coupling
R_2-Cl,Br,I
17 electrophiles
34 nucleophiles
18 catalysts
10 bases
8 solvents
$w_{solvent}$ x 1.5
B
2,569 Suzuki Reactions
High
Low
Medium
Zero
PROD/IS Group
$RFP = \sum (FP_n) \cdot w_n$
$tsne_x$
$tsne_y$
C
PROD/IS
electrophile #
nucleophile #
catalyst #
base #
solvent #
D
1
2
3
16 catalysts
6 bases
DMSO
E
$w_{catalyst}$ x 3
w_{base} x 3
PROD/IS
catalyst #
base #
6 bases
DBU
MTBD
BTMG
BEMP
BTTP
P_2Et
16 catalysts
BINAP Pd G3
DPPF Pd G3
XantPhos Pd G2
$^tBu_3P$ Pd G2
PPA Pd G3
APhos Pd G3
XPhos Pd G3
RuPhos Pd G2
DTBPF Pd G3
J009 Pd G3
MorDalPhos Pd G3
BrettPhos Pd G3
$^tBuXPhos$ Pd G3
$^tBuBrettPhos$ Pd G3
RockPhos Pd G3
AdBrettPhos Pd G3

84-Fig. 2 | 2,786 Suzuki reactions performed in a high-throughput format plotted using a t-SNE trained on *reaction fingerprints*, each of which formed through the sum or concatenation of a reactant fingerprint for all components involved a particular reaction. (A) The template of the reaction performed across all experiments. (B) The 2,048-bit radius four Morgan Fingerprint of each reaction's electrophile, nucleophile, base, catalyst, and solvent were summed, plotted, and colored by output group. (C) The same manifold colored by exact product/internal standard values and reagent components (electrophile, nucleophile, catalyst, base, solvent.) (D) A specific reaction cluster containing all reactions between substrates **1** and **2**. (E) t-SNE of the reactions in the selected cluster when the catalyst fingerprint is multiplied by three (top) and when the base fingerprint is multiplied by three (bottom). The three manifolds for each result are colored by product/internal standard values, catalyst, and base.

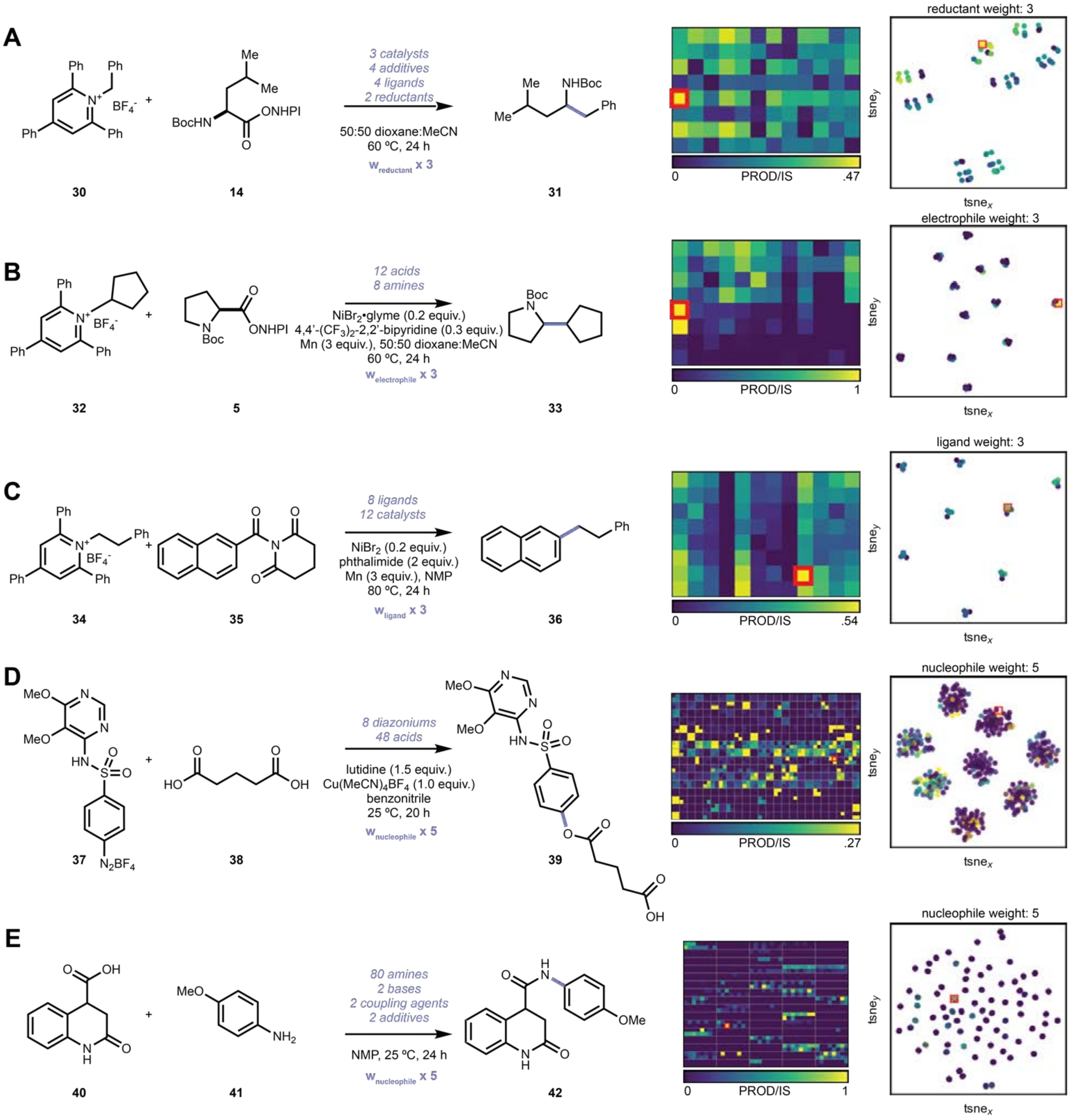

deaminative–decarboxylative C–C coupling
23 acids
15 amines
13 catalysts
14 co-catalysts
28 ligands, 8 reductants
10 bases, 5 additives
5 solvents
w$_{acid}$ x 3
A
B
Initial Reactions
Methodology Development
Scope Exploration
acid #
23
1
PROD/IS
1
0


C
ONHPI =
4
2% (336)
Average PROD/IS (Number of Experiments)
5
12% (456)
6
1% (120)
7
0% (4)
8
3% (4)
9
0% (4)
10
0% (4)
11
0% (32)
12
4% (16)
13
4% (48)
14
22% (11)
15
19% (16)
16
35% (16)
17
35% (16)
18
25% (16)
19
23% (16)
20
4% (8)
21
25% (16)
22
36% (16)
23
8% (8)
24
28% (16)
25
37% (8)
26
19% (8)

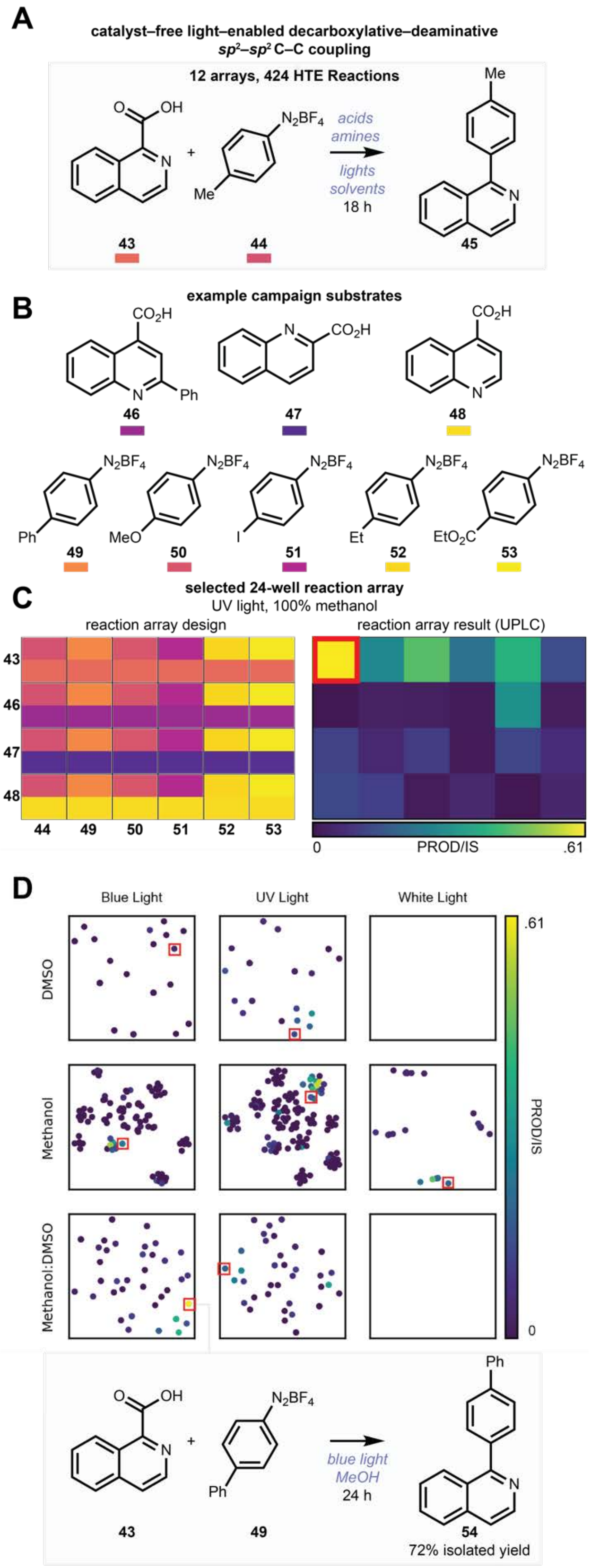
85-Fig. 3 | 1,296 reactions performed in an HTE format towards the development of a novel *sp*^{3}–*sp*^{3} deaminative-decarboxylative carbon–carbon cross coupling (ref. 13). The manifold splits into three regimes, each representing a different stage of the discovery campaign. (A) The manifold points are colored by the acid used in the experiment. (B) The manifold points are colored by the recorded PROD/IS value of the reaction. (C) All acids used in the experiment, with average PROD/IS values and the number of experiments they were used in.



86-Fig. 4 | Reactivity cliffs are identified when focusing on specific substrate pairs. (A) Compounds **27** and **28** cluster together with high and poor performing reactions. (B) Repeating the analysis on the cluster that form **29** from **27** and **28** reveals reagents that cause this reactivity to flip on. Reactivity cliffs are readily identified between solvent regimes. The 1:3 water:NMP regime sits between two reactivity cliffs, separating it from the two failing solvent systems 1:9 ^{t}AmOH:NMP and pure NMP.



87-Fig. 5 | Reaction array manifolds can be procedurally generated from the output files produced by phactor. (A)-(F) reaction array results as reported in ref. 8 and their corresponding manifolds colored by output value. Perplexity and weights were modified as described in the Supporting Information to optimize the latent space for visualization.


A
catalyst–free light–enabled decarboxylative–deaminative
sp²–sp² C–C coupling
12 arrays, 424 HTE Reactions
O OH
N
+
N2BF4
Me
acids
amines
lights
solvents
18 h
Me
N
43
44
45
B
example campaign substrates
CO2H
N Ph
46
N CO2H
47
CO2H
N
48
N2BF4
Ph
49
N2BF4
MeO
50
N2BF4
I
51
N2BF4
Et
52
N2BF4
EtO2C
53
C
selected 24-well reaction array
UV light, 100% methanol
reaction array design
reaction array result (UPLC)
43
46
47
48
44
49
50
51
52
53
0
PROD/IS
.61
D
Blue Light
UV Light
White Light
DMSO
Methanol
Methanol:DMSO
.61
PROD/IS
0
O OH
N
+
N2BF4
Ph
blue light
MeOH
24 h
Ph
N
43
49
54
72% isolated yield

88-Fig. 6 | (a) Standard conditions for catalyst-free light-enabled decarboxylative–deaminative sp^2–sp^2 C–C coupling. (b) Sampling of substrates explored. (c) Screen design and execution. (d) Trellised reaction array fingerprint of experimental campaign. Each box contains all reactions performed for the given light and solvent regime. Reactions that perform well in one regime but not others are quickly identified.

Reaction Array Analysis and Design via Reagent Networks

In this work, a circular network design was developed to visualize and evaluate the progress of reaction discovery campaigns. Reagent datasets for each component in a reaction were embedded into manifolds and spaced evenly around a circle. Links or chords connecting points from each manifold indicate their co-use in a reaction. When visualizing reaction data with the networks in a sequential format, a clear logic is developed that enables the invention and optimization of reactions. Initial experiments are generally mined from the literature or involve screening reagent inventories. When reactivity ceilings are hit, additional components of changes in reaction regimes are necessary for further optimization. Using this logic, Andrew McGrath optimized the one-pot etherification of esters.

Abstract

Reactions can be described in a standardized template of reagents, which facilitates the multiparameter analysis and design of reaction arrays in high throughput experimentation campaigns. Reagent networks can be utilized to evaluate the effectiveness of reaction arrays and to identify trends leading to desired reactivity. Herein, several high throughput experimentation campaigns from previously published methodologies are analyzed with reagent networks, from which a generalizable logic for reaction optimization was developed. This logic was then utilized to design reaction arrays towards the optimization of an etherification reaction from ester substrates.

Introduction

High-throughput experimentation (HTE) is a modern tactic in synthetic chemistry to invent and optimize novel reactivities and provides a way to efficiently search reaction space. Reaction array design is traditionally performed by interrogating literature as well as chemist intuition and expertise. However, the emergence of data-driven modeling has provided the modern chemist with tools to execute reaction arrays in a more efficient and effective manner.[248] Several studies have shown the utility of HTE in data-driven model development to predict reaction performance.[41, 125, 226, 244, 249, 250] These publications reveal marginal increases to reaction performance, indicating that machine learning models are not able to entirely capture the chemical reactivity from the provided reaction parameters. Indeed, much of the reported datasets are inconsistent, and are biased for working or high-yielding reactions. Recent work has attempted to standardize the reporting of HTE data, which streamlines analytics and ensures all aspects of the experimental procedure are captured.

We have recently disclosed a standardized template for the typical chemical reaction (89-Fig. 1A) and several reports of optimized methodology achieved through HTE with the HTE management software phactor. The machine-readable output from the HTE campaigns were collated and analyzed through network visualization (89-Fig. 1B). In this style of visualization, each vertex corresponds to a particular reagent class from the reaction template, and points within the vertices are specific reagents of that class to a point in each other vertex. The manifolds within the vertices are principal component embeddings of featurized molecules. Depending on the reagent class, a different featurization can be utilized for the embedding. A reaction is encoded by connecting the points in each vertex. Colored points represent reagents that have been tested in the HTE campaign and correspond to the mean product over internal standard integration for that reagent across all experiments. Similarly, links between points are colored by mean product over internal standard integrations for all experiments containing the reagents linked by the chord. Black points in the manifold represent compounds within the reagent class that have yet to be tested.

We show in this analysis that critical insights derived from the literature can inform initial plate designs as well as lead to key experiments that provide the highest boosts to reactivities. By viewing the networks in a sequential fashion, modifications to the reaction conditions that lead to large increases in product formation can be rapidly identified. Herein, we showcase the network analysis of several campaigns, and point to literature reports that led to reaction array designs that improved target reactivity, as evidenced by the change in reaction network visualizations once the literature information was incorporated into the campaign. Finally, we utilize the methodology optimization logic distilled from this analysis developed on previous chemistries to guide a novel optimization of an ester to ether reduction methodology.

In the first analysis, the optimization of an alkyl-aryl deaminative decarboxylative carbon-carbon coupling is tracked with sequential reaction networks. (90-Fig. 2A) Notably, the reaction networks consider the cumulative effect of the campaign, and colors represent the mean yields of reagents and reagent pairs across all experiments performed. The first reaction array (**A**) tested eight ligands, three solvents, and four catalysts in a 96-well plate. The design of this plate was based on reagents found in the 2020 *Organic Letters* report on nickel-catalyzed deaminative acylation between activated amines and amides by the Matsuo group.[251] The corresponding reaction network show faint lines, indictive of the initial hit of the desired C-C bond formation activity. The brightest network represents the reaction between activated acid **1** and activated amine **2** to form product **3** under nickel catalysis. A contemporary report in *Science* informed the additive screen **B**, where it was found that phthalimide drastically improved the yield of the reaction for many substrates, where in the meantime the nickel catalyst was slightly optimized to Nickel (II) Bromide. The addition of phthalimide is strongly indicated in the network, with bright connections emitting from the additive vertex. Finally, plate **C** was informed by the uncovering of a

1966 *Journal of the American Chemical Society* study by Halpern and Kemp on the role of Ru(II)Cl_2 in the decarbonylation of formic acid.[252] A screen of 24 Lewis acids in our inventory found that Ru(III)Cl_3 provides a fantastic boost to desired reactivity while decreasing formation of side products across many substrates. The final reaction conditions were able to produce a yield of 94% between **1** and activated acid **4** to generate **5**. Indeed, the point representing Ru(III)Cl_3 is colored brightly and emits bright chords to the other vertices in the network.

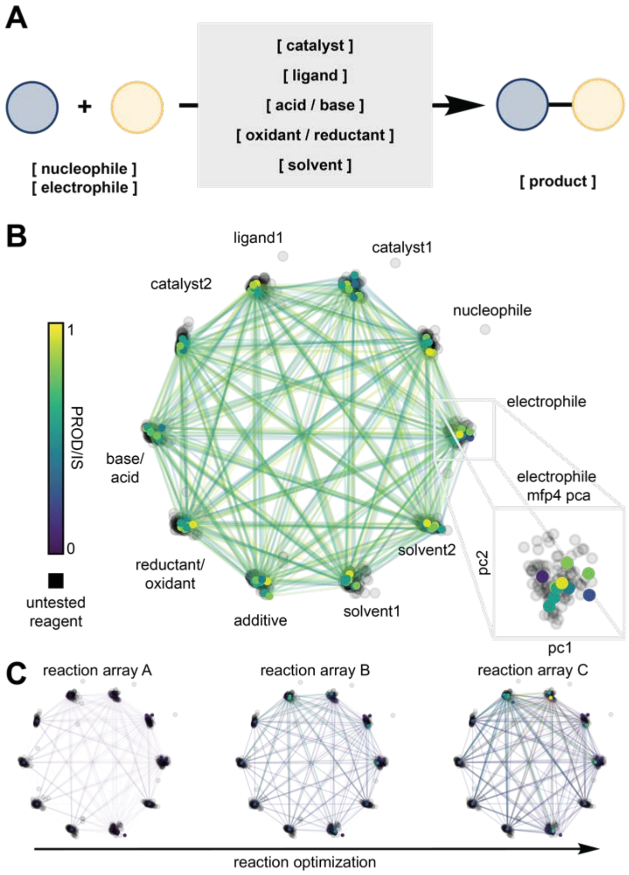
In 90-Fig. 2B, an analogous analysis is performed on the optimization of an alkyl–alkyl deaminative–decarboxylative carbon–carbon coupling. The discovery campaign was initially attempted through photoredox conditions found in the literature to couple activated acid **6** with activated amine **7** to form **8** through iridium catalyst under blue light.[253-257] In this initial screen (90-Fig. 2B – screen **A**), the best hit resulted in a PROD/IS integration of 15%. As a limit to the reactivity was realized, efforts moved couple the substrate pair with nickel catalysis in reductive conditions found in the literature (90-Fig. 2B – screen **B**).[258-262] By this point in the campaign, the conditions were far more effective in creating the desired product, as evidenced by the results of the plate and the improved connectivity in the network diagram. In this case, the change in reactivity regime is identified through the use of lithium chloride as additive, and a bright focal point can be identified in the additive manifold, as well as a general improvement of connections throughout the entire network. By switching to conditions that utilize a zinc reductant instead of photocatalysts, the desired product was obtained in 77% PROD/IS integration. The final boost in reactivity was identified through the optimization of the ligand and through the replacement of zinc with manganese as the reductant. The idea of using manganese as reductant was once again borne from literature data.[251, 263-266] With di-trifluoromethylbipyridine as ligand and the use of manganese, the optimal reaction conditions were found, resulting in reactivity amenable to activated primary amines such as **7** and activated secondary amines such as **9**. With these conditions a variety of substrates were used to form diverse products such as **10**, the extent of which shown in the colored connections between the amine and acid manifolds in the network. Additionally, thick and brighter lines are emitted from the ligand and reductant manifolds, reflecting the improvement in reactivity due to the use of these reagents.

The analysis of the previous campaigns can be reduced to a logic for HTE optimizations. Initial plate designs can be made using the ‘shotgun’ approach, where reagents are tested in a somewhat random or blind manner or be informed from literature on similar reactivities. After an initial hit is found, reagents similar to those found in the initial hit are tested, leading to minor improvements in yield and selectivity as local chemical space is explored. In periods of stagnation, additional factors or components can be added or removed from the conditions in hope of a serendipitous boost in yield, or new or existing literature can be incorporated in the design of the next reaction array. We can use reaction networks to clearly map the areas of reaction space that has been tested (and to what effect), as well as areas of unexplored reaction space. With statistical algorithms, the most promising networks can be tested in plate format.

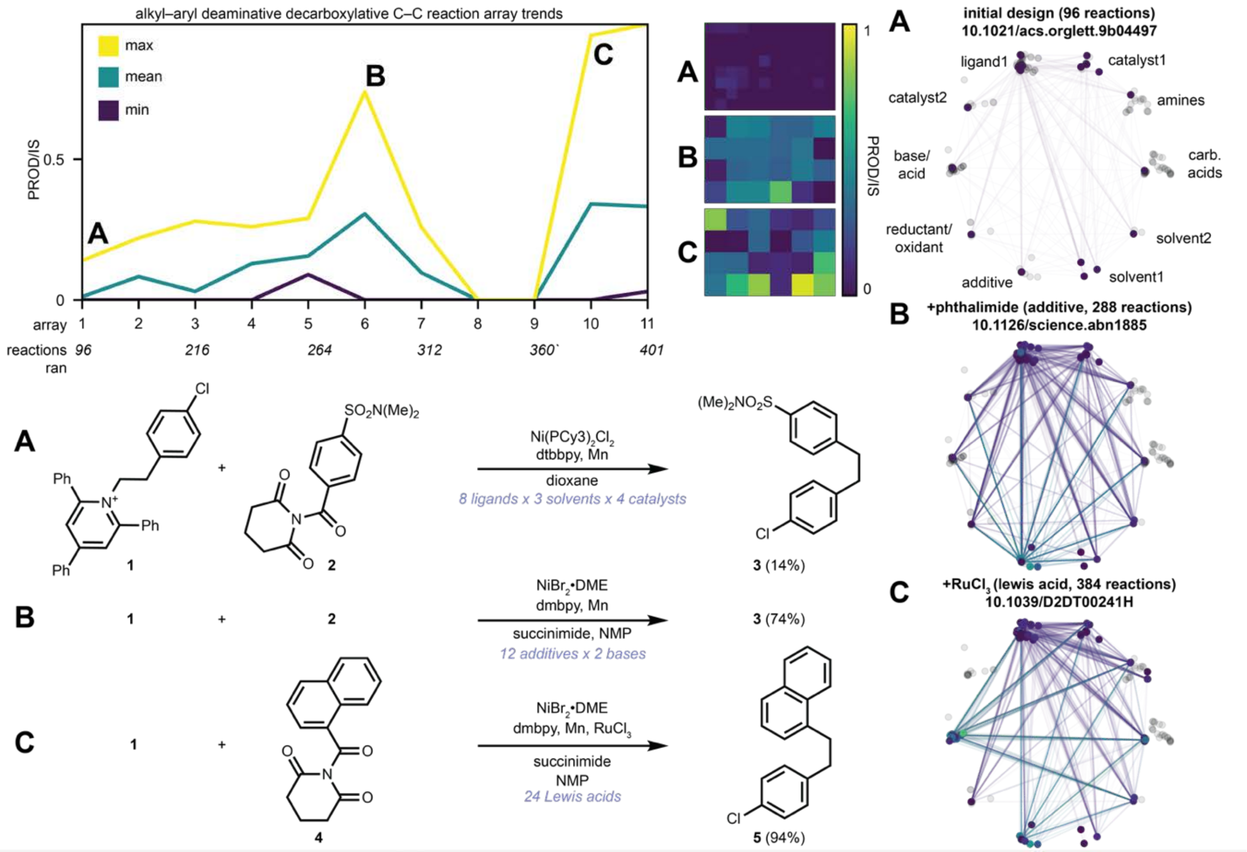
We showcase how this logic was used in the optimization of an ester to ether reaction (91-Fig. 3A). Initial studies were based off of two papers indicating the use of silanes and Lewis acids in the reduction.[267, 268] In 91-Fig. 3B, we identify that the silanes, boranes, and Lewis acids can be used to perform this reduction in an initial screen, albeit to low yields via PROD/IS integration. The large cloud of black points in silane and borane space reveal that there are many untested reagents that could improve the yield further. In 91-Fig. 3C, we looked to test the reagents from the untested pool using trends from the previous assay and interaction effects between the silanes and boranes, and the initial hit was expanded in a 96 well reaction array, testing a larger extent of metals, boranes, and silanes. Again, diphenylsilane proved to be the most effective reductant, and gallium (III) iodide was found to be the best Lewis acid tested in the assays. At this point, no borane nor silane was found to significantly outperform another, and a reactivity ceiling was hit. The reaction network clearly indicates the strong performance of the gallium reagent, whereas bright chords fail to emit from silane or borane space. Nonetheless, it was clear that additional screening of borane or silanes would not improve reactivity much further to an acceptable range. Using the logic developed through previous discovery campaigns, it was decided to expand the experimental space by adding an additional component to the reaction. In 91- Fig. 3D. the final optimization assay is shown, where gallium (III) iodide and diphenylsilane are standard conditions used in each reaction, and boranes and silane additives[269, 270] are screened. To our delight, the inclusion of an additional silane additive improved the reactivity drastically, as indicated by the many bright chords that are in the corresponding reaction network.

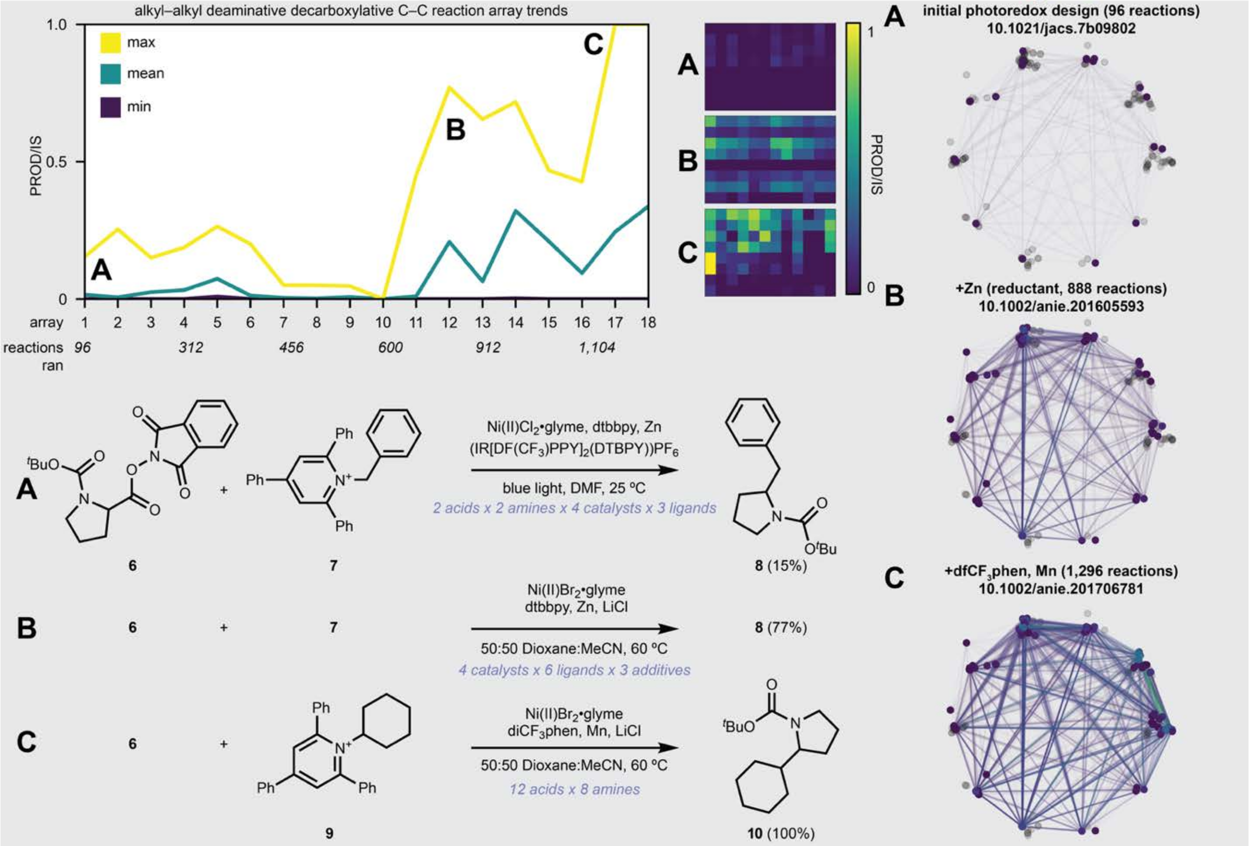
Conclusion

We present reagent networks as a method to analyze series of sequential reactions. Key reagents are readily identified through the formation of chords connecting embedded reagents from various manifolds representing classes of reaction conditions. Reactions can be optimized to high yields by following a logic developed from the analysis of reaction invention campaigns. Yield ceilings can be broken by incorporating additional reagents into the reaction mixture or changing reaction condition regimes after sufficient testing reveals limits to initial reactivity.

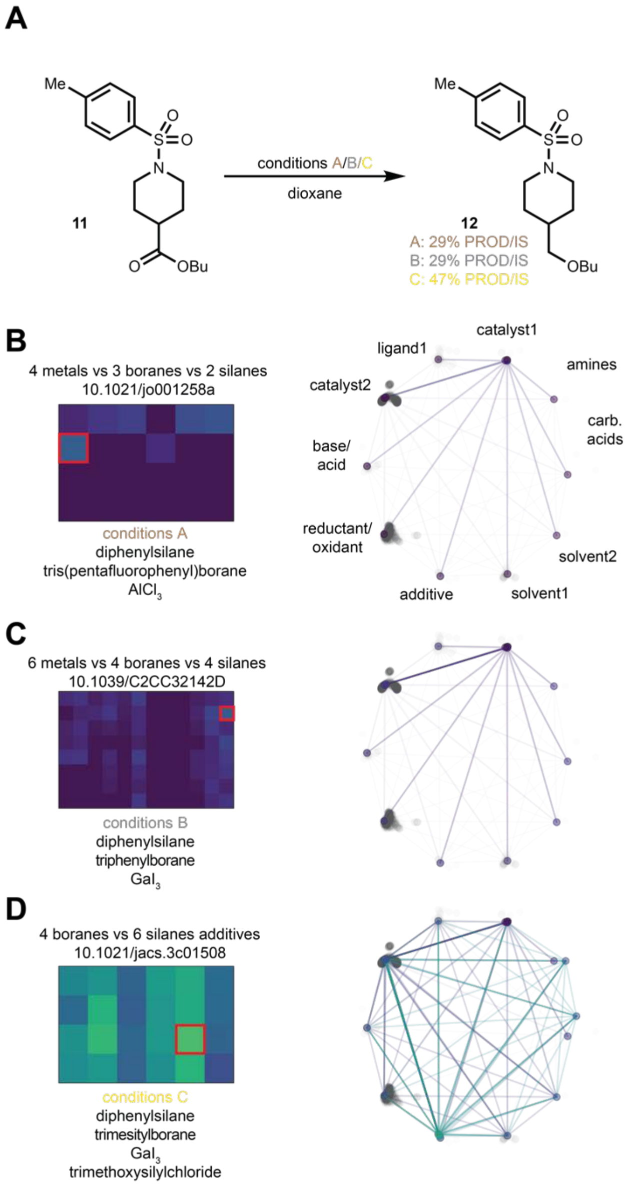


89-Fig. 1 | A) A standardized template for chemical reactions. B) A reagent network containing 817 high throughput reactions performed in an optimization campaign of an alkyl-aryl deaminative decarboxylative carbon-carbon coupling method. Each vertex corresponds to a different class of reagents used in the reaction. These vertices are reduced manifolds of reagents corresponding to that class and can be formed from different featurizations depending on the class. Points are colored by mean yields of the reagent, and links between two points are colored by mean yield of reactions containing both reagents, indicating the effectiveness of their co-use. Black points indicate reagents not utilized in the campaign and serve as potential design space.


alkyl–aryl deaminative decarboxylative C–C reaction array trends
max
mean
min
PROD/IS
0
0.5
A
B
C
array 1 2 3 4 5 6 7 8 9 10 11
reactions ran 96 216 264 312 360` 401
1
0
A
initial design (96 reactions)
10.1021/acs.orglett.9b04497
ligand1
catalyst1
catalyst2
amines
base/ acid
carb. acids
reductant/ oxidant
solvent2
additive
solvent1
B
+phthalimide (additive, 288 reactions)
10.1126/science.abn1885
C
+RuCl$_3$ (lewis acid, 384 reactions)
10.1039/D2DT00241H
Ni(PCy3)$_2$Cl$_2$
dtbbpy, Mn
dioxane
8 ligands x 3 solvents x 4 catalysts
1
2
3 (14%)
NiBr$_2$•DME
dmbpy, Mn
succinimide, NMP
12 additives x 2 bases
3 (74%)
NiBr$_2$•DME
dmbpy, Mn, RuCl$_3$
succinimide
NMP
24 Lewis acids
4
5 (94%)


alkyl–alkyl deaminative decarboxylative C–C reaction array trends
max
mean
min
PROD/IS
0
0.5
1.0
A
B
C
array 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
reactions ran 96 312 456 600 912 1,104
1
0
A
initial photoredox design (96 reactions)
10.1021/jacs.7b09802
B
+Zn (reductant, 888 reactions)
10.1002/anie.201605593
C
+dfCF$_3$phen, Mn (1,296 reactions)
10.1002/anie.201706781
Ni(II)Cl$_2$•glyme, dtbbpy, Zn
(IR[DF(CF$_3$)PPY]$_2$(DTBPY))PF$_6$
blue light, DMF, 25 ºC
2 acids x 2 amines x 4 catalysts x 3 ligands
6
7
8 (15%)
Ni(II)Br$_2$•glyme
dtbbpy, Zn, LiCl
50:50 Dioxane:MeCN, 60 ºC
4 catalysts x 6 ligands x 3 additives
8 (77%)
Ni(II)Br$_2$•glyme
diCF$_3$phen, Mn, LiCl
50:50 Dioxane:MeCN, 60 ºC
12 acids x 8 amines
9
10 (100%)

90-Fig. 2 | A) Graphical and reagent network analysis of an alkyl-aryl deaminative decarboxylative carbon-carbon coupling methodology optimization campaign. B) The analysis repeated on an alkyl-alkyl deaminative decarboxylative carbon-carbon coupling methodology optimization campaign.



91-Fig. 3. A) Best reaction from initial reaction array and reagent network analysis of ester to ether methodology campaign. The initial reaction array design was derived from a literature report. B) The second reaction array in the campaign. Boxplot trends reveal gallium iodide as the most robust Lewis acid for the transformation. C) Expected improvement manifolds calculated from the featurized silane and borane chemical spaces. New silanes and boranes were automatically selected to be tested in the follow up experiment.

Chapter V. Molecular Compression

Computational molecular representation plays an integral role in streamlining the selection of efficacious reagents for chemical reactivity and finetuning the structural integrity of drugs. Key factors such as chemical features that bear relevance to the optimization issue at hand need to be encoded. This aids in crafting a precise and ideal depiction of the chemical space that's being traversed.[24] Essentially, a molecule's chemical attributes are intrinsically interconnected with its chemical structure.[1] This necessitates that computational representations take into account both the physical shape of a molecule, the graphical connections it comprises, and how these aspects interact with the properties the molecule presents. String representations like SMILES (Simplified Molecular Input Line Entry System)[241] or SELFIES (Self-Referencing Embedded Strings)[271] offer canonical textual formats to depict molecules. However, the potential to efficiently analyze sequential or textual information was only recently achieved by computation. Molecular fingerprints, another computational method, encapsulates a molecule by converting it into a string of bits, each bit signifying the absence or presence of a specific molecule feature.[243, 245, 272]

The key to successful molecular compression is twofold. First, the representation must be computer-friendly or easily readable by a machine. Second, the original structure that was compressed needs to be accurately decompressed or reconstructed from the encoded representation. While SMILES and SELFIES can be reformed into a graphical rendition of the molecule, the majority of fingerprinting techniques result in a loss of vital information. This subsequently hinders the accurate reconstruction of the original molecular structure. In light of this challenge, this chapter introduces an innovative approach that looks at music as a potential data-rich platform onto which molecular information can be encoded.

Music, inherently sequential and temporal, mirrors the atomic structure of molecules, which follows a similar canonical sequencing. Therefore, it can serve as an alternative and perhaps a more effective representation. Additional molecular properties can be encoded into the musical domain, which allows for the embedding of more information gathered or discerned from the molecule. Both local and global properties of the molecule can be incorporated into the music, through features like timbre or note length that are reflective of the molecule's characteristics. For instance, note-wise properties could mirror local molecular attributes, while score-wise features could represent global molecular properties. Thus, the music offers a potentially powerful and more comprehensive way to represent and understand the complexities of molecular structures.

Molecular sonification for molecule to music information transfer

This creative work brought the merger of molecules and music using an encoding and decoding algorithm. The one-to-one algorithm used string representations of molecules to create musical scores based on the atomic structure and properties of the molecule. These musical scores were then shown to have molecular relevance. A seq-to-seq network was trained on the molecular music, and interpolation of the latent space created new music that generated novel molecules when decoded. When surveying a class of students, it was found that the majority of students were able to discern the similarity of molecules when listening to their musical encodings without any visual information.

Abstract

Organic chemical structures encode information about a molecule's atom and bond arrangement. The most established way to encode a molecular structure is through line drawing, although other representations based on graphs, strings, one-hot encoded labels, or fingerprint arrays are critical to the computational study of molecules. Here we show that music is a highly dimensional information storage medium that can be used to encode molecular structure. The resultant method allows a molecular structure to be heard as a musical composition, where the key of the music is based on the molecular properties and the melody is based on the atom and bond arrangement. This allows for a molecular generation approach that leverages modern artificial intelligence tactics for music generation.

Introduction

The representation of chemical structures is critical to the study and invention of functional molecules. Organic molecules are classically described as line drawings,[273] where all atoms and their corresponding bonds are drawn on paper or on a computer. Other simple molecular representations or identifiers include molecular formulae, IUPAC names or CAS numbers, which require little memory and are machine readable, but carry minimal information. Molecules can also be represented as graphs, with atoms as nodes and bonds as edges. By encoding atomic coordinates and connectivities line by line, the topology of molecules can be embedded as a graph on a computer for rendering, editing and analysis. The transmission of molecular information into machine-readable formats has invited new molecular structure representations, such as SMILES,[274] SMARTS,[275] InChI keys,[276] DeepSMILES,[277] and SELFIES.[271] These representations are cheap to store in memory and provide valuable structural information for rapid lookup and comparison. While these aforementioned representations have been useful for inputting molecules into computers, and encoding structural and stereochemical information, they are one-dimensional string representations that are more difficult for human users to interpret and interact with than the classic line drawing representation of molecules. To adapt them for machine learning and data science algorithms, SMILES and other strings are typically converted to vector representations *via* molecular fingerprints such as Extended Connectivity Fingerprints (ECFP),[243] Morgan Fingerprints,[242] atom-pair fingerprints,[278] and others. This dimensionality expansion is a core tactic in the analysis of virtual chemical libraries or predictions of molecular properties. Other high-dimensionality fingerprint representations, such as physics-based descriptors[279] or physicochemical descriptors,[142, 280] are also common. While computers can easily parse molecular information from these representations, interactivity with human users is difficult with the fingerprint-based information media. In addition, once converted to such a fingerprint, the molecule is typically no longer uniquely revertible to its atom-bond representation.[272]

Music is a high-dimensional information storage medium that maximizes both human and computer interactivity, interpretability, and creativity. We considered that music could be used for storage of molecular information. While many aspects of a molecule are readily visible to the trained eye, a soundbite may be able transmit more information about the molecule into the mind. The encoding of molecules as music is particularly intriguing since the multiple dimensions of music can allow encoding of many molecular properties.[281] Music is also highly interactive both for humans and for computers. Musicians can control many parameters that can embed information about a molecule, such as tempo, rhythm, notes, key, instrument, effects, *etc.* If molecules could be encoded as music, opportunities would emerge for visual-to-audio sensory substitution, for instance providing blind chemists new ways to interact with molecules.[282] Contemporary chemistry and drug discovery leverage artificial intelligence[283] (AI) and there has meanwhile been an explosion of AI methods in the study and creation of music,[284] so we were excited by the prospects of merging modern chemistry machine learning (ML) techniques with recent ML techniques for music. Our initial impetus was to explore how music could be used as a creative medium to generate new molecules, but in the course of our studies we have learned that molecules likewise can provide an inspiration and creative outlet for the generation of new music.

Sonification is the encoding of non-musical information as music and provides a means to encode information in many musical dimensions, while simultaneously providing a new means of interactivity.[285] A variety of information sources have been sonified, such as visual art,[286] the architecture of spider webs,[287] infrared spectra of chemicals,[288] amino acid sequences,[289] air pollution,[290] fire,[291] and many more.[292-299] The SELFIES representation provided a viable input for molecular sonification, both for the encoding of molecules into a melody and the construction of new molecules *via* performance on a musical instrument such as the piano. We developed a workflow for transferring molecules into music, and *vice versa*, which we call Sonic Architecture for Molecule Production and Live-Input Encoding Software, or SAMPLES (92-Fig. 1).

Methods and Workflow

Encoding

To create a melody based on a molecular structure, the key and the sequence of notes are derived from its physicochemical properties and its SELFIES sequence, respectively. To determine the key, the physicochemical properties of a molecule—such as $\log P$, molecular weight, and number of hydrogen bond donors and acceptors—are summed, and the final number is linearly projected into the integer space between 1 and 12 from the minimum and maximum values found in the entire dataset (in this case, DrugBank), with each bin corresponding to a particular key. In our encoding scheme, the projection is largely dictated by the molecular weight due to its magnitude with respect to the other features. Due to low deviation, low mean, and outliers of high molecular weight, most molecules from the DrugBank dataset fall into the first bin. In our scheme, the keys represented by the bins are ordered by the popularity of keys found on Spotify, with the most popular bin being the key of G (see Fig. S1). As such, most compounds in DrugBank are encoded into G and larger molecules are encoded into the less popular keys. The sequence of notes is determined from a one-to-one mapping between the SELFIES token of the molecule and multi-octave steps in the major scale (see ESI for the specific mappings and key distribution used in this study). By adding the MIDI value of the melody's key to the MIDI shifts that correspond to notes of the major scale (derived from the SELFIES tokens of the molecule), the final melody is produced. In our case, the MIDI shifts correspond to SELFIES tokens as ranked by their popularity in DrugBank. As such, lower MIDI shifts away from the encoded key of the molecule indicate the more popular the SELFIES token. Every fourth note, starting from the first note, was converted to a major chord to increase the texture of the music. We envision polyphony as a potential avenue to encode atomic features. This algorithm could be extended by including more descriptors in the encoding and assigning distinct keys to clusters. As well, mapping more elements or structural subgroups to MIDI shifts or other musical parameters can enable the encoding of proteins, inorganic molecules, and other chemical phenomena.

Decoding

The MIDI shifts are reverse calculated for each key and converted into a molecular structure. As such, multiple structures are generated (one for each key) for the same MIDI sequence. Each structure is then hashed into a key using the original key encoding algorithm. If the hashed key matches the key used in the reverse calculation, the molecular structure is decoded. It is guaranteed that at least one decoding key will match a hashed key for any MIDI generated from SAMPLES.

A demonstration of the SAMPLES encoding function is shown in 93-Fig. 2. Ammonia (**1**) appears as a single note while benzene (**2**) generates a slightly more complex musical composition. The unity of these two molecules produces aniline (**3**) whose musical sequence highly resembles the concatenation of the two musical sequences of **1** and **2**.

Expansion of **3** into indole (**4**) creates a slightly more complex melody owing to both the increased molecular size and the additional information content required to describe a ring fusion between the 5- and 6-membered rings. In the reverse direction, songs are readily translated to molecules, such as **5**, which is produced from the song “Twinkle, twinkle little star” when played using D flat as the reference note (the decoding key).

SAMPLES is readily scaled to more complex and drug-like molecules (94-Fig. 3). Tolmetin (**6**) and ketolorac (**7**) create a rich and textured musical composition. Meanwhile, tabersonine (**8**) and vindoline (**9**) provide complex melodies. Scaling to large complex molecules, such as taxol, oxytocin, or vincristine (see ESI) required no modifications and generated nuanced euphonic melodies.

Case studies

To showcase the utility of this novel algorithm, four experimental case studies are presented. Using our approach, molecular properties can be heard. For instance, the songs generated from molecules that pass the Lipinski rules[300] can be auditorily distinguished from those that fail the Lipinski rules based on the musical key. This is largely because the molecule's aggregate physicochemical properties were hashed to the musical key, with the most popular physicochemical property fingerprints from the pharmaceutical database DrugBank hashing to the most popular song keys from the music database Spotify.[301] The concept of molecular similarity is of high importance to molecular invention, such as in selecting molecules with comparable functional properties for drug discovery. We were curious to explore if SAMPLES generated from molecules with high Tanimoto similarity[302] (fingerprint based) would sound similar, appreciating that both molecular similarity and musical similarity are difficult to define.[303] Molecules represented in the t-SNE in 95-Fig. 4 are more similar if closer to each other on the plot. Indeed, we deemed the SAMPLES of codeine (**10**) and morphine (**11**) to sound similar to each other while the SAMPLES of sulfamethoxazole (**12**) and sulfadoxin (**13**) likewise sound similar, while the pair of **10** and **11** sounded distinct from the pair of **12** and **13** (95-Fig. 4).

Our second experiment investigates the generation of molecules *via* modification of the music domain. A key motivator for our research was the ability to generate new molecules through the interactivity of a piano keyboard, or other musical hardware or software. This was made possible in SAMPLES through the application of SELFIES, which enable editing of string bits while consistently producing valid molecular structures. Thus, starting from morphine (**11**), the musical score could be modified one note at a time (96-Fig. 5) to generate new chemical structures **14–16** bearing a clear relationship to **11** but with noticeably modified bond and atom architecture. Due to our encoding scheme, shifts further away from the melody's key result in SELFIES being modified into to atoms less commonly found in the DrugBank database. Random modification of SELFIES can result in drastic changes to molecular structure as seen in 96-Fig. 5, due to the non-atom encoding SELFIES tokens that dictate the size of features such as branches and rings. Note that SAMPLES may generate undefined stereocenters.

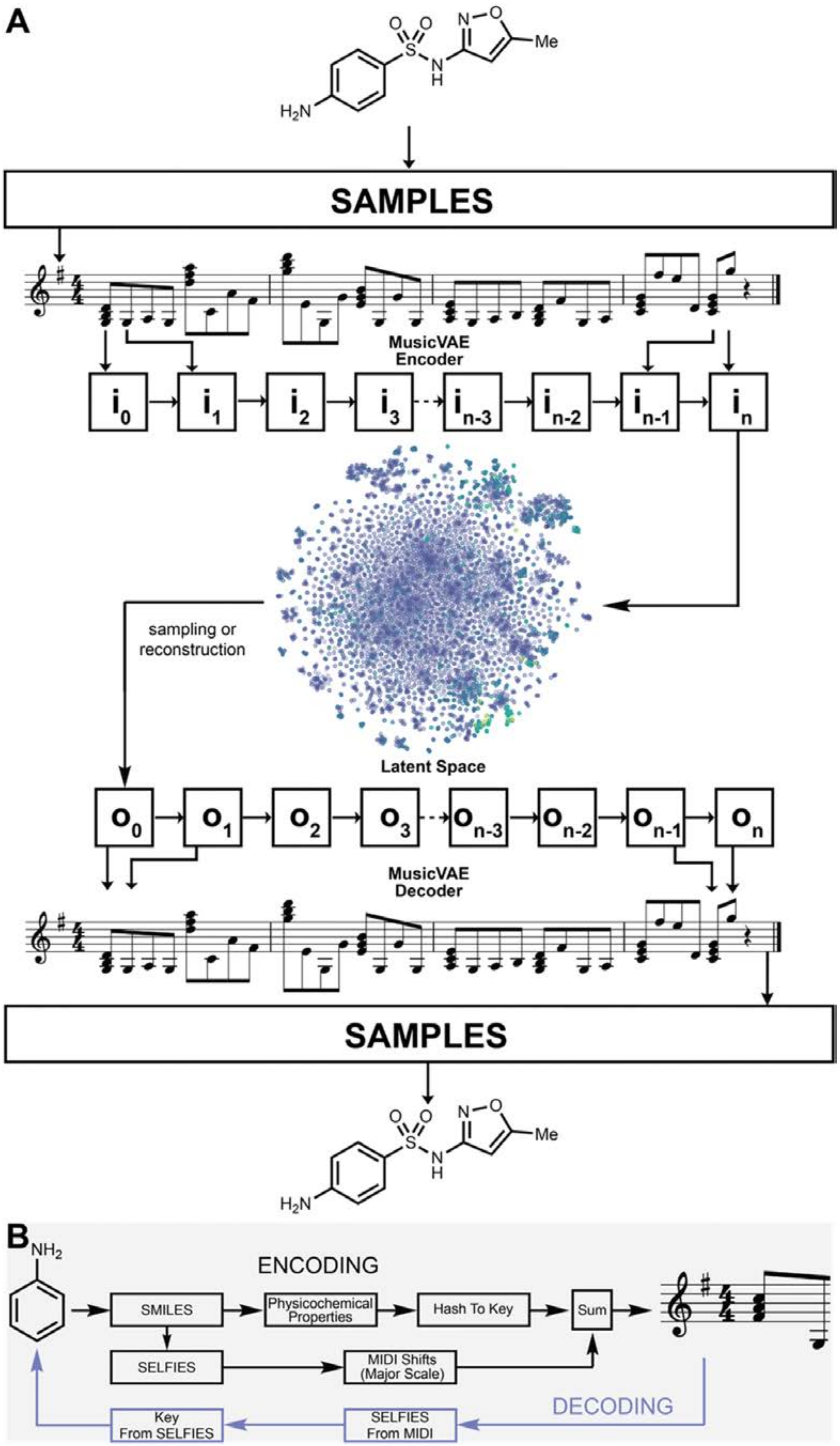
Having demonstrated the feasibility of molecular generation using SAMPLES, we explored the ability of modern machine learning methods developed for music generation as tools for molecule generation. In this third case study, we applied the melody mixing function of MusicVAE[304] using MIDI melodies derived from SAMPLES as inputs. Using MusicVAE, two melodies could be blended to generate an interpolated melody, and that new melody could be translated back to a molecular structure using SAMPLES, thus creating a new molecule that was a "blend" of the two input molecules (97-Fig. 6). We call this function CROSSFADE. The blending of musical compositions is an established practice, with considerable hardware and software to support the musical blending process. While algorithms that generate new molecules by blending the structures or properties of input molecules are known,[305] we are intrigued by the interactivity offered by CROSSFADE. As an example, glutamic acid (**17**) and acetylcholine (**18**) were CROSSFADEd to produce **19**, **20** and **21** CROSSFADE to **22** and similar results are obtained for **23–28**. A four-step interpolation is shown in the ESI.

As a final experiment, to take the editing of the molecules on the keyboard a step further, and to demonstrate the human-interactivity enabled by the SAMPLES algorithm, a human created a monophonic composition inspired by SAMPLES-generated music (98-Fig. 7), which was decoded to molecule **29**. It was necessary to exert some human bias into the musical composition, based on the composer's knowledge of chemistry and SAMPLES, since generating a molecule that is as carboniferous as most drugs and natural products requires bias towards the key's tonic note, in this case C, since that is mapped to the carbon atom. In other words, the song was written to ensure that the most played note corresponded to the carbon atom to reflect the nature of organic molecules. This required a basic understanding of organic chemistry and musical theory. While this implementation of SAMPLES focuses on major scale notes, the embedding algorithm can be easily modified to include mappings for minor scale notes or accidentals.

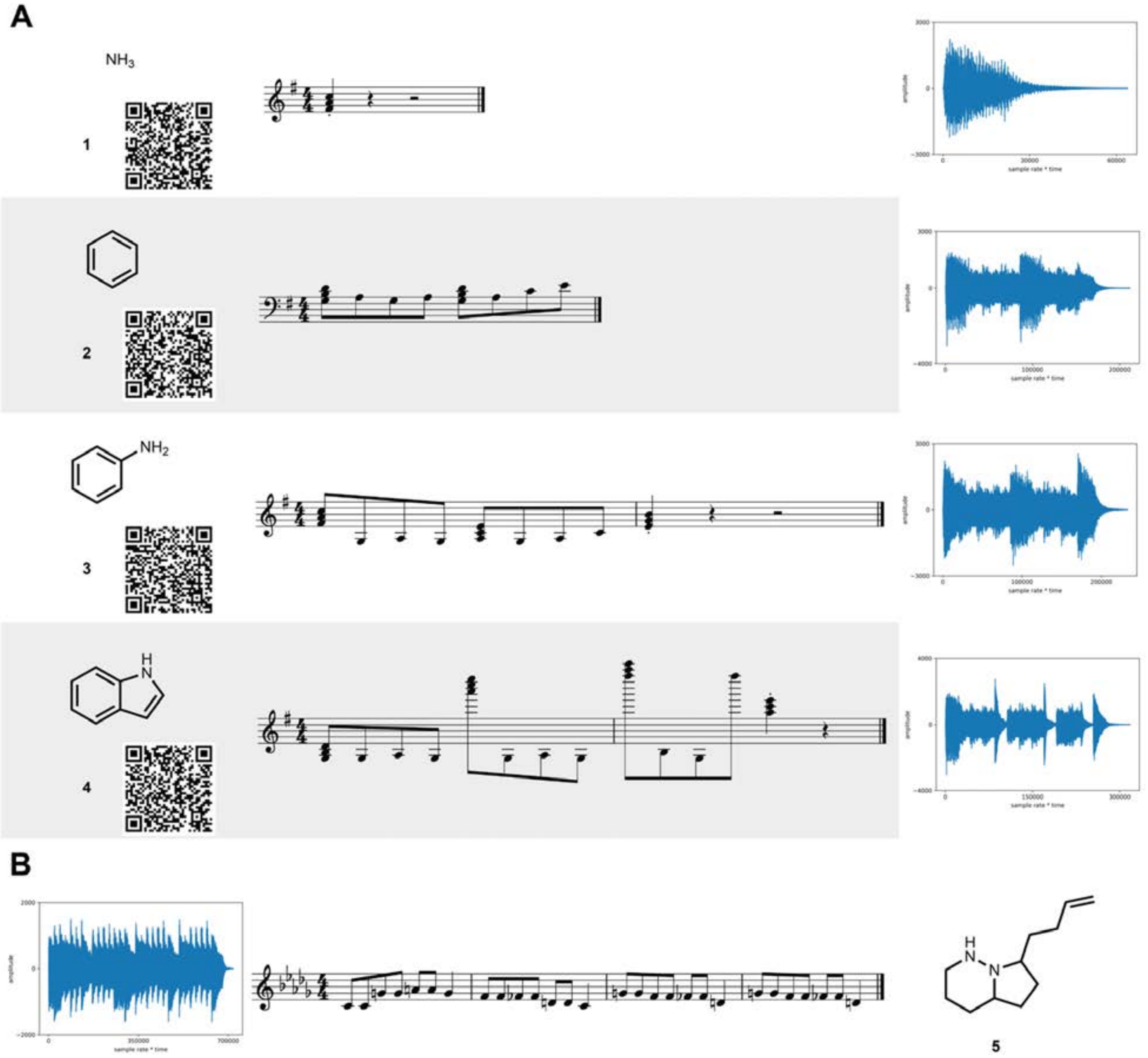
One goal of this embedding algorithm was to show that music can be used as a medium to encode molecular information and that similar molecules can produce similar sounding melodies. While this is largely dependent on the featurization and encoding of the molecules into music, our algorithm seemed to perform well for some druglike molecules as tested by the human ear. We surveyed 75 participants from a senior level undergraduate medicinal chemistry course with a blind multiple-choice poll. Students were given four questions, each of which asking the student to compare the melody of four encoded molecules to the melody of an encoded molecule given in the question. The student is prompted to select the melody that they believe to sound most similar to the given sample for each question without being shown any structures. In the first question, **10** was given to the student and the student was given a choice between the melodies of **11**, **30**, **31**, and **32**. In this case, despite the high Tanimoto similarity between codeine **10** and morphine **11**, most students did not recognize similarity between the two encodings, instead hearing **30** and **31** to sound more like **10**. However, in question two where **12** was compared to **13**, **33**, **34**, and **35**, most students correctly identified the melodies of sulfamethoxazole (**12**) and sulfadoxin (**13**) to be most similar. Students were also able to identify the most similar melodies and molecules as described by Tanimoto similarity in questions three and four, where melodies of **7** and **9** were correctly chosen to be similar to **6** and **8** over **36–38** and **39–41**, respectively. Puzzlingly, question one had the highest similarity between the test molecule and correct answer molecule compared to the other question pairs (99-Fig. 8).

Conclusion

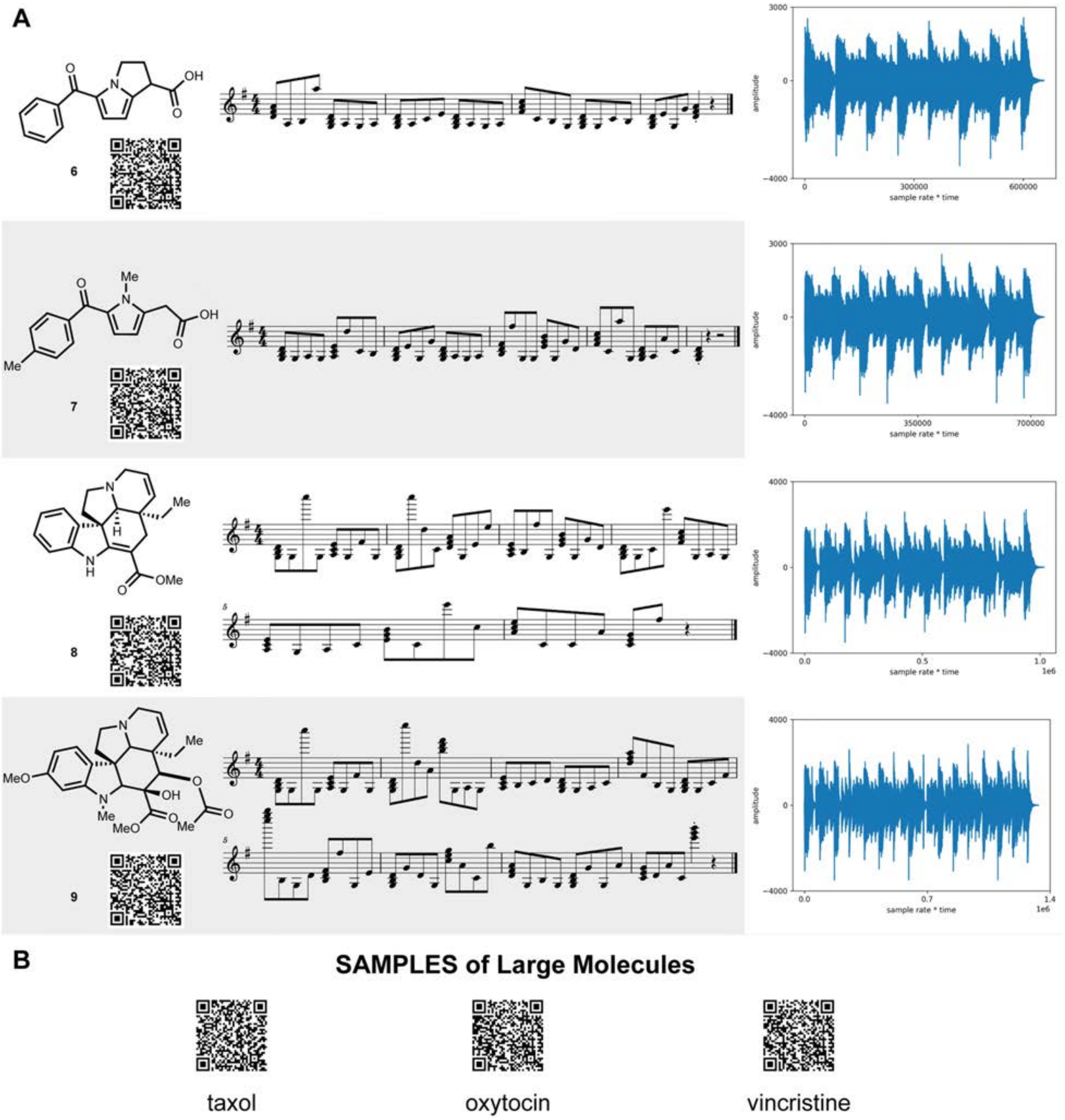
We report an alternative means of encoding organic molecules through music. The resultant melodies allow a human to interact with molecular structures through musical hardware and software *via* note editing, insertion and deletion, as well as produce molecular structures through original compositions. We note that far more molecular features can be encoded into a melody or even a short soundbite than can be visually represented by the 2D or 3D structure. Indeed, many richer possibilities can be enabled by encoding more features into musical elements such as other scales, chord compositions, rhythms, timings, timbres, *etc.* One transfer learning application for which the current study may be used is music generation. The motivation for machine learning for content generation is its generality, that is no formal grammar or rules must be specified for such a model to generate content.[284] Transforming molecules into music provides a rich collection of musical data that can be used to train music generation models, as seen with MusicVAE. Particularly, sequence to sequence (seq2seq) models, such as recurrent neural networks,[306] allow for the interconnection of domains containing data signals with variable lengths such as text, music, and machine-readable molecular representations based on structure. Seq2seq models can learn a fixed length embedding of variable length signals that can be used for classification tasks and direct mathematical comparison. For instance, word2vec[307] and GloVe[308] provide pretrained word embeddings that have been learned from massive text corpuses such as Wikipedia or Twitter. In a molecular context, variational autoencoders have been used to learn the distribution of molecular features, such as SELFIES tokens, to provide a continuous embedding of molecular space.[24] SAMPLES provides an avenue to directly connect molecules to content-generating machine learning models in the music domain. Computational exploration and interpolation within the melodies described herein is possible, generating new molecules that sound and look similar to existing molecules. This highlights the possibility of leveraging music-based artificial intelligence for molecular design. An online implementation of the encoding portion of the SAMPLES algorithm can be accessed at http://samples.cernaklab.com.


A
O
O
S
N
H
N
O
Me
H₂N
SAMPLES
MusicVAE
Encoder
i0
i1
i2
i3
in-3
in-2
in-1
in
sampling or
reconstruction
Latent Space
o0
o1
o2
o3
on-3
on-2
on-1
on
MusicVAE
Decoder
SAMPLES
O
O
S
N
H
N
O
Me
H₂N
B
NH₂
ENCODING
SMILES
Physicochemical
Properties
Hash To Key
Sum
SELFIES
MIDI Shifts
(Major Scale)
DECODING
Key
From SELFIES
SELFIES
From MIDI

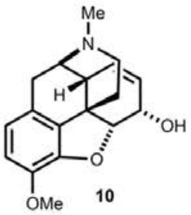
92-Fig. 1 | (A) Workflow for SAMPLES. Molecules are first assigned a musical key based on aggregate chemical properties, then converted into a sequence of notes based on SELFIES encoding. MusicVAE is trained on a collection of sonified molecules to formulate the chemical/musical latent space. The latent embedding of molecular music can then be sampled, for instance through the interpolation between two embedded molecules, decoded by the MusicVAE decoder, then converted back into a molecular structure by SAMPLES. (B) Detailed workflow schematic of SAMPLES. Molecules are binned into base keys based on their physicochemical properties. More druglike molecules correspond to melodies that are more popular as reflected in the distribution of songs listed in Spotify. Specific SELFIES tokens are mapped to MIDI shifts, that result in a final MIDI value when summed with the base key of the molecule, which is hashed from aggregate physicochemical properties. The MIDI shifts correspond to SELFIES tokens as ranked by their popularity in DrugBank. As such, lower MIDI shifts away from the encoded key of the molecule indicate the more popular the SELFIES token. Melodies can then be decoded back into molecules.



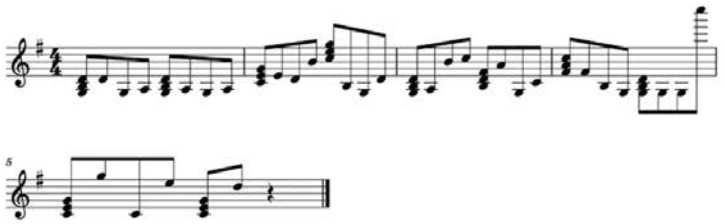
93-Fig. 2 | SAMPLES translates molecules into music. (A) The generation of increasingly complex molecules from **1** through **4** corresponds to increasing musical complexity. Each line shows the molecular structure, the corresponding musical score, and a waveform of the MIDI output. Audio recordings are available in the ESI and can be quickly retrieved by scanning the QR code with a mobile device. (B) In the reverse direction, the song "Twinkle, Twinkle Little Star" produces molecule **5**.

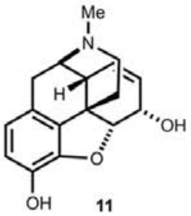


94-Fig. 3 | SAMPLES is amenable to encoding complex molecules. (A) The pair of similar molecules **6** and **7** have SAMPLES compositions that are distinct from another similar pair of molecules **8** and **9**. (B) QR codes linking to SAMPLES encodings of select large complex molecules.

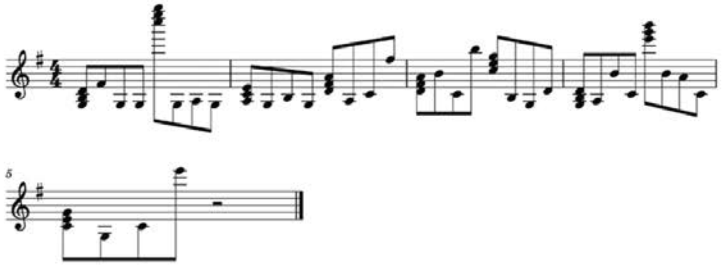

Me
N
H
OH
O
OMe

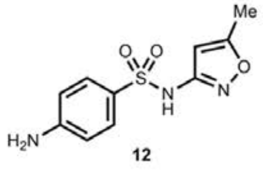
10


5

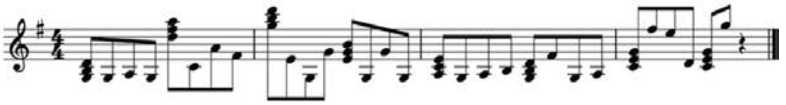

Me
N
H
OH
O
OH

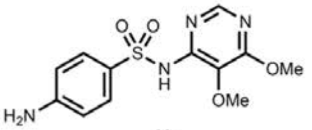
11


5

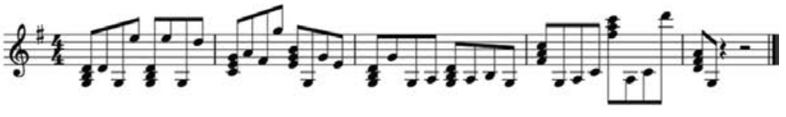

Me
O
O
S
N
H
N
O
H_2N

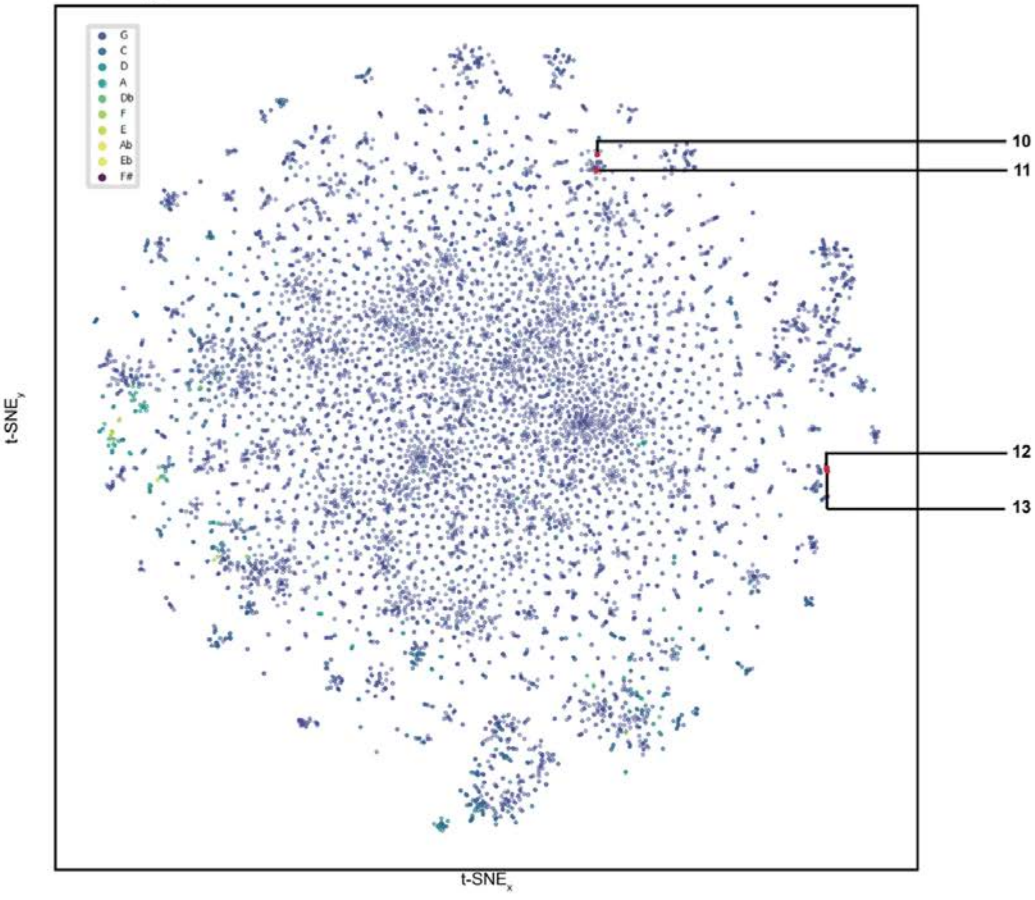
12



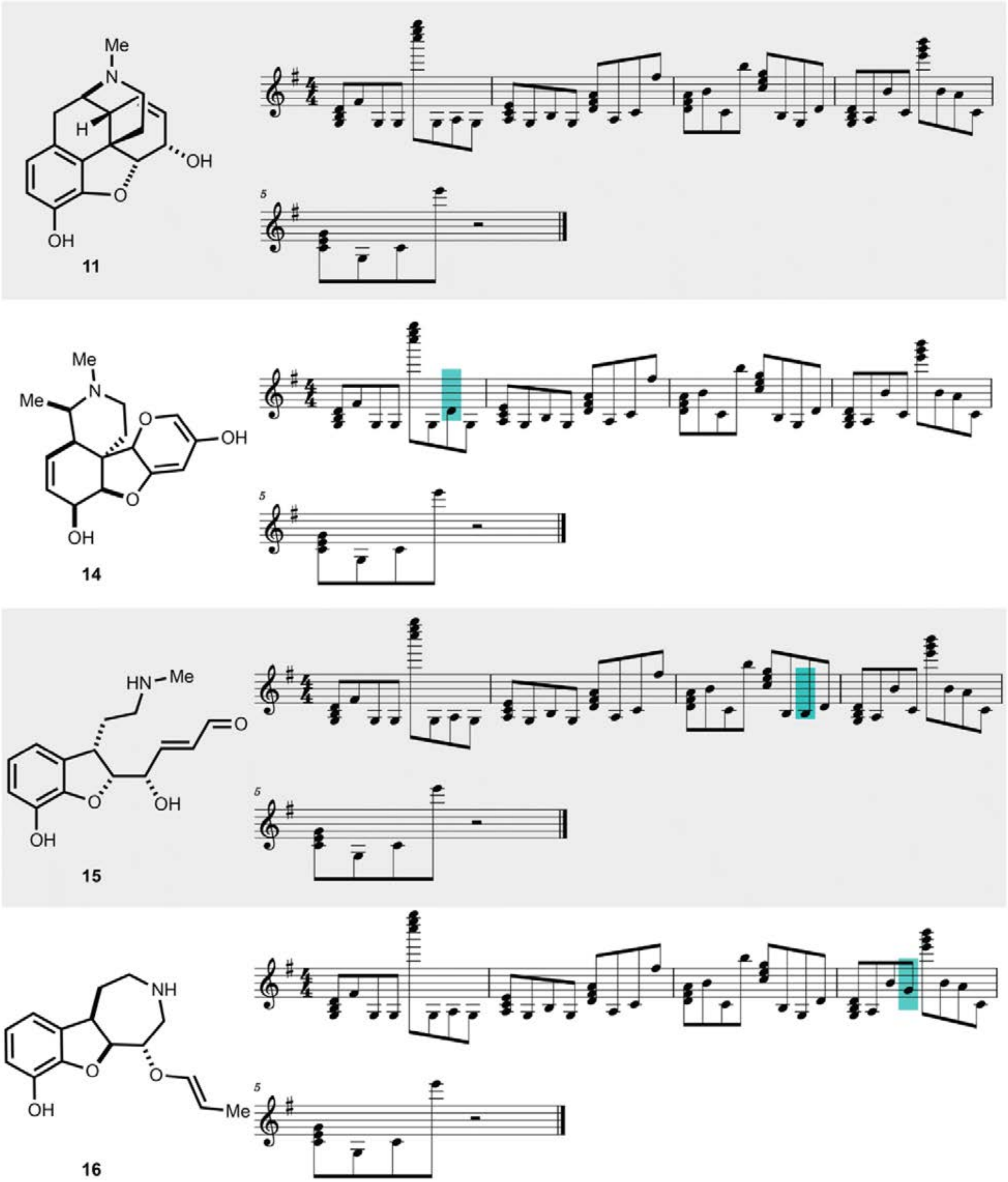

O
O
S
N
H
N
N
OMe
OMe
H_2N

13

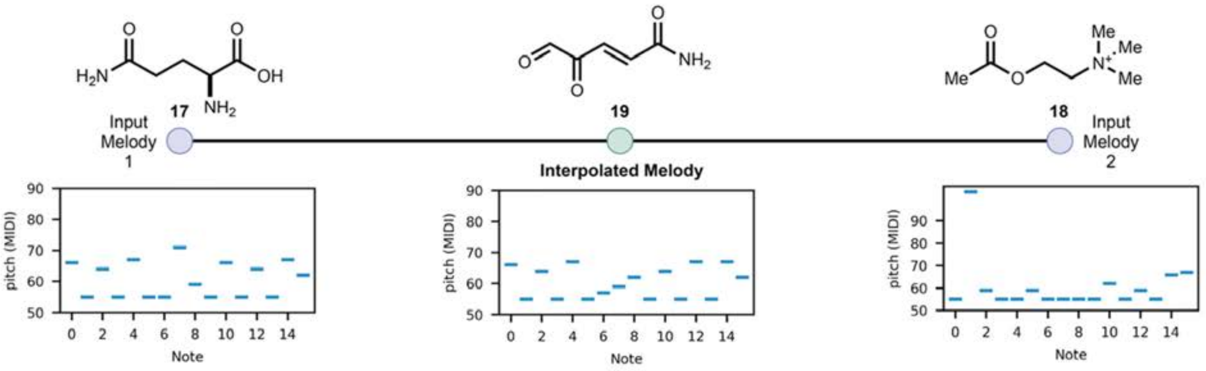


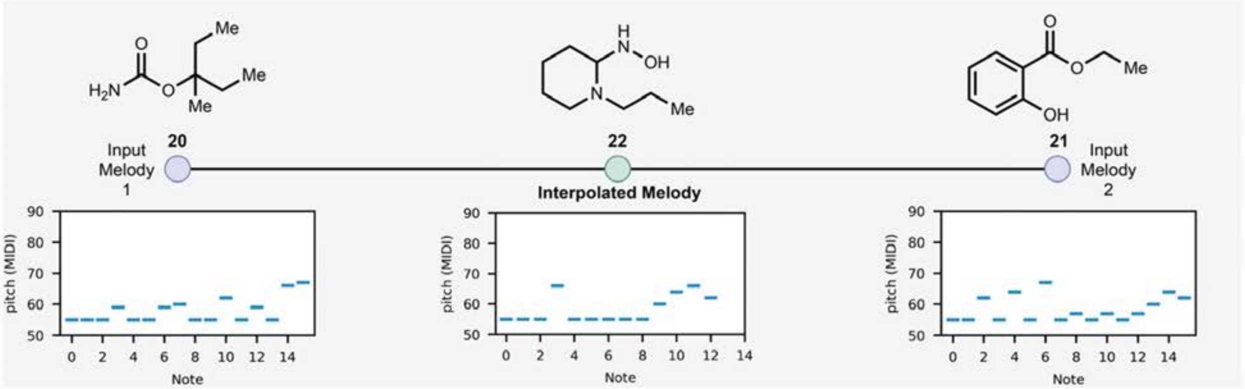

G
C
D
A
Db
F
E
Ab
Eb
F#
10
11
12
13
t-SNE$_y$
t-SNE$_x$

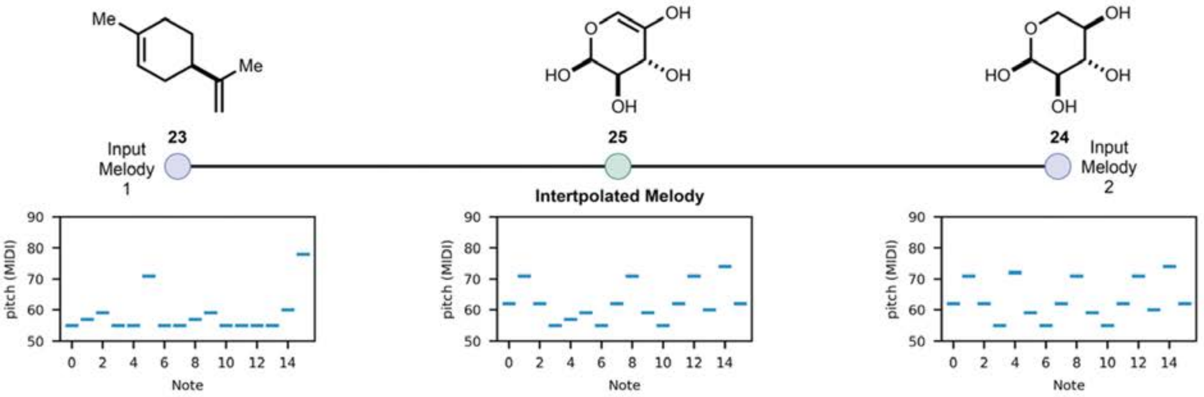
95-Fig. 4 | tSNE embedding of 11 159 drugs from DrugBank (2048 bit Morgan Fingerprints of radius 2), coloured by their SAMPLES musical key. Similar molecules **10** and **11** have SAMPLES outputs that are distinct from other similar molecules **12** and **13**. The music encoded by the score adjacent to the molecules can be listened to by following the QR code. It can be heard that these similar molecules have similar musical encodings when using our algorithm.

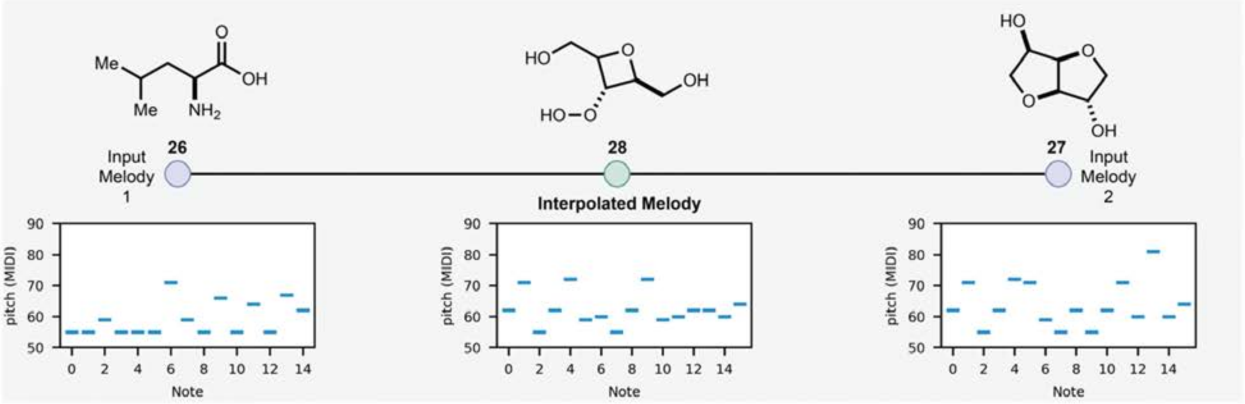


96-Fig. 5 | Molecular editing in SAMPLES generates distinct but related molecules. The manual editing of single notes in the SAMPLES of **11** leads to **14**, **15**, or **16**.

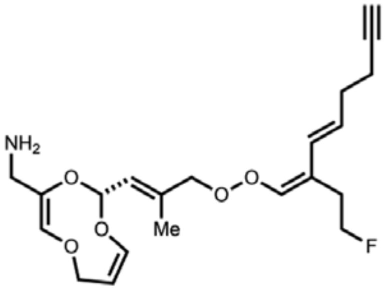

17
Input Melody 1
19
Interpolated Melody
18
Input Melody 2
pitch (MIDI)
Note


20
Input Melody 1
22
Interpolated Melody
21
Input Melody 2
pitch (MIDI)
Note

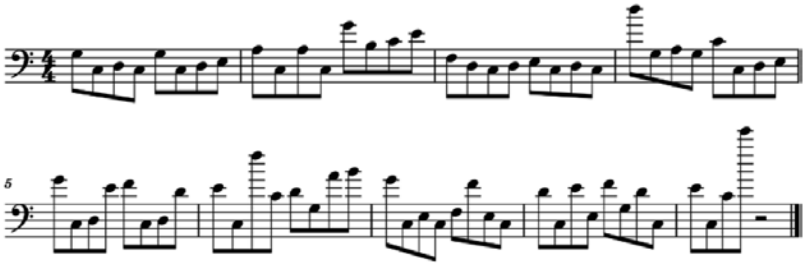

23
Input Melody 1
25
Intertpolated Melody
24
Input Melody 2
pitch (MIDI)
Note


26
Input Melody 1
28
Interpolated Melody
27
Input Melody 2
pitch (MIDI)
Note

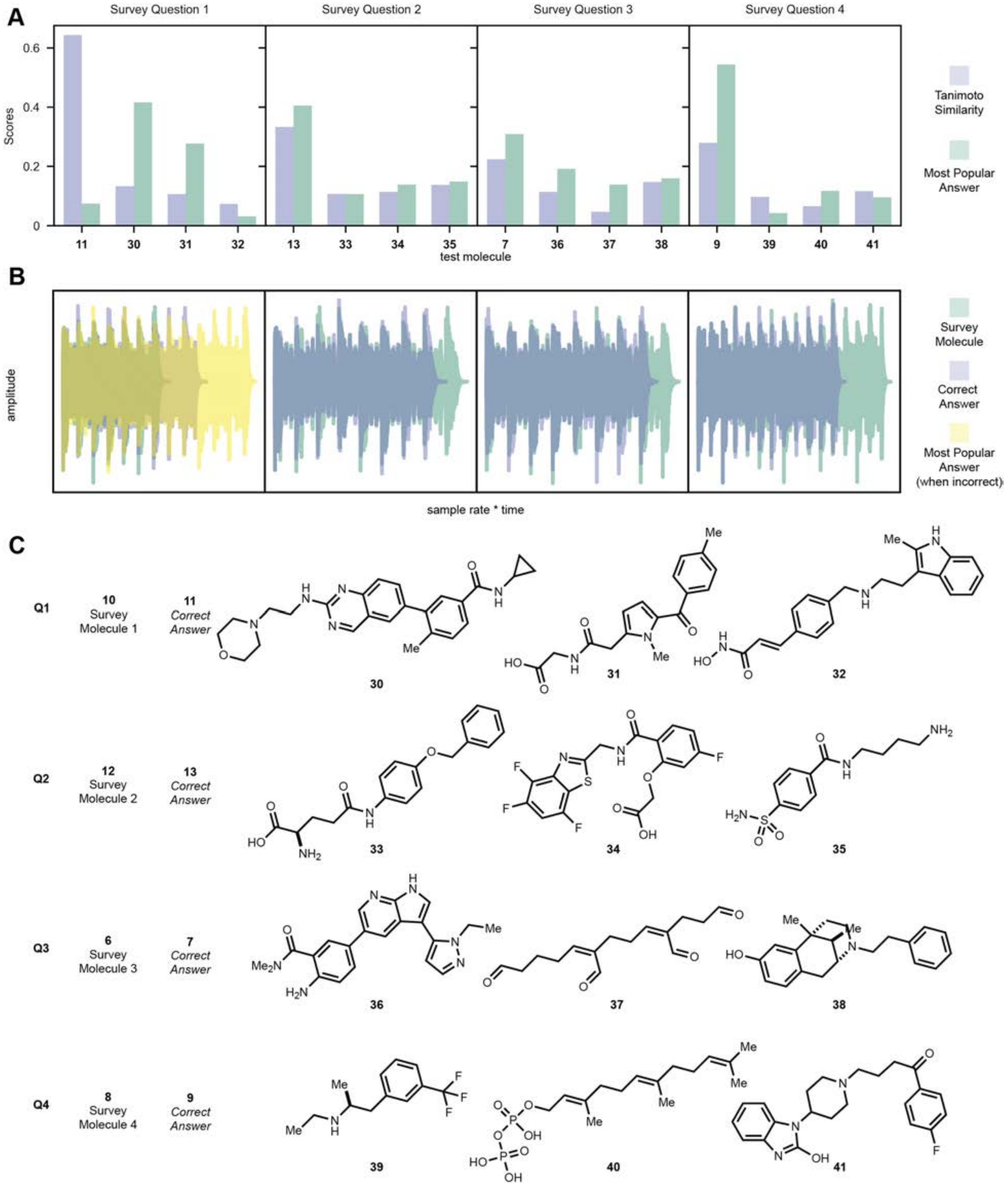
97-**Fig. 6 |** CROSSFADE merges SAMPLES with the melody mixing function of MusicVAE to create interpolated molecules based on two input molecules.



29



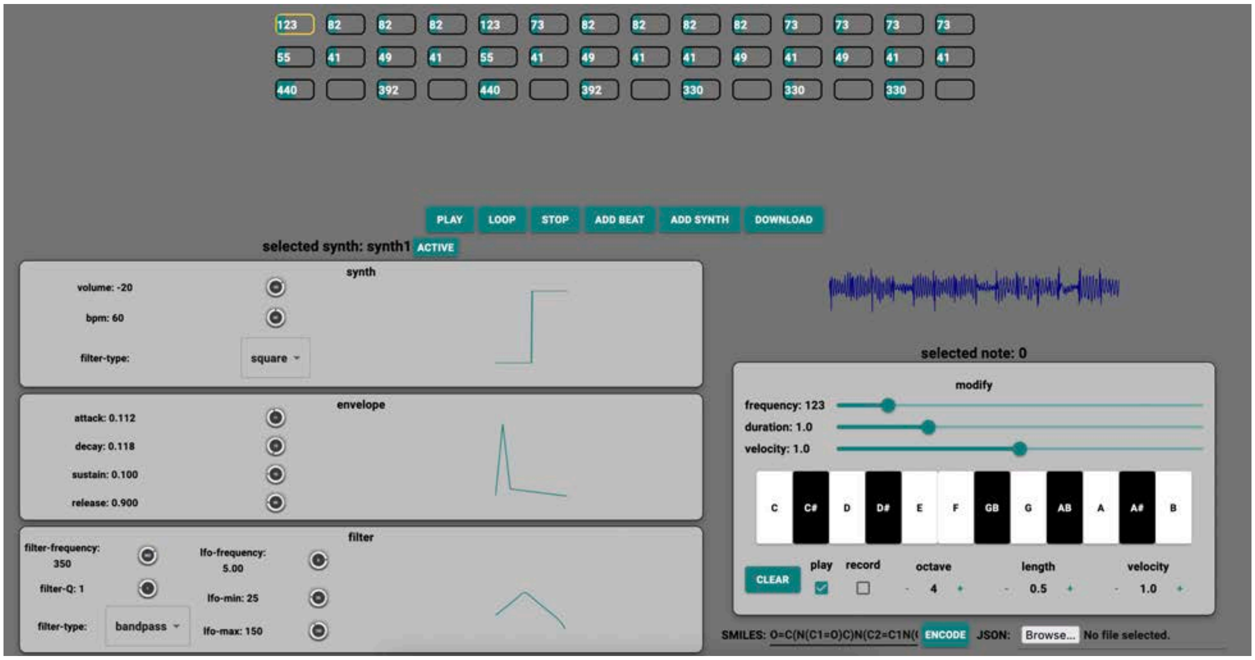
98-**Fig. 7 |** A human created music composition leading to **29**.



99-Fig. 8 | (A) Survey results from 75 participants. Each participant was given the SAMPLES encoded melody of four survey molecules. For each survey molecule, without knowledge of the name or structure of any molecule, each participant was asked to choose the most similar melody from a selection of four other SAMPLES encoded drugs. Survey responses are cross examined against the Tanimoto similarity between each test molecule and survey molecule for each question. (B) Superimposed waveforms, for each question, of the survey molecule, the molecule most structurally similar to survey molecule, and the most popular answer if the majority response in incorrect. (C) Structures of survey and test molecules for each question. One structure for each set of test molecules was chosen to have high similarity to the respective question's survey molecule to serve as the ‘correct answer’.

Molecular Featurization via Musical Additive Synthesis

In this project, we took the concept of sonification a step further by recreating an 8-bit additive synthesizer. Chemical properties are encoded into various settings of each synth. Molecular properties are encoded into features that affect the musical output globally, and atomic properties dictate the sequence of notes and corresponding musical features at that specific time such as note velocity and pitch. A unique data interface is provided allowing external programs to be written and generate their own encoding schemes.



123	82	82	82	123	73	82	82	82	82	73	73	73	73
55	41	49	41	55	41	49	41	41	49	41	49	41	41
440		392		440		392		330		330		330	

selected synth: synth1

volume: -20
bpm: 60
filter-type: square

attack: 0.112
decay: 0.118
sustain: 0.100
release: 0.900

filter-frequency: 350
filter-Q: 1
filter-type: bandpass
lfo-frequency: 5.00
lfo-min: 25
lfo-max: 150

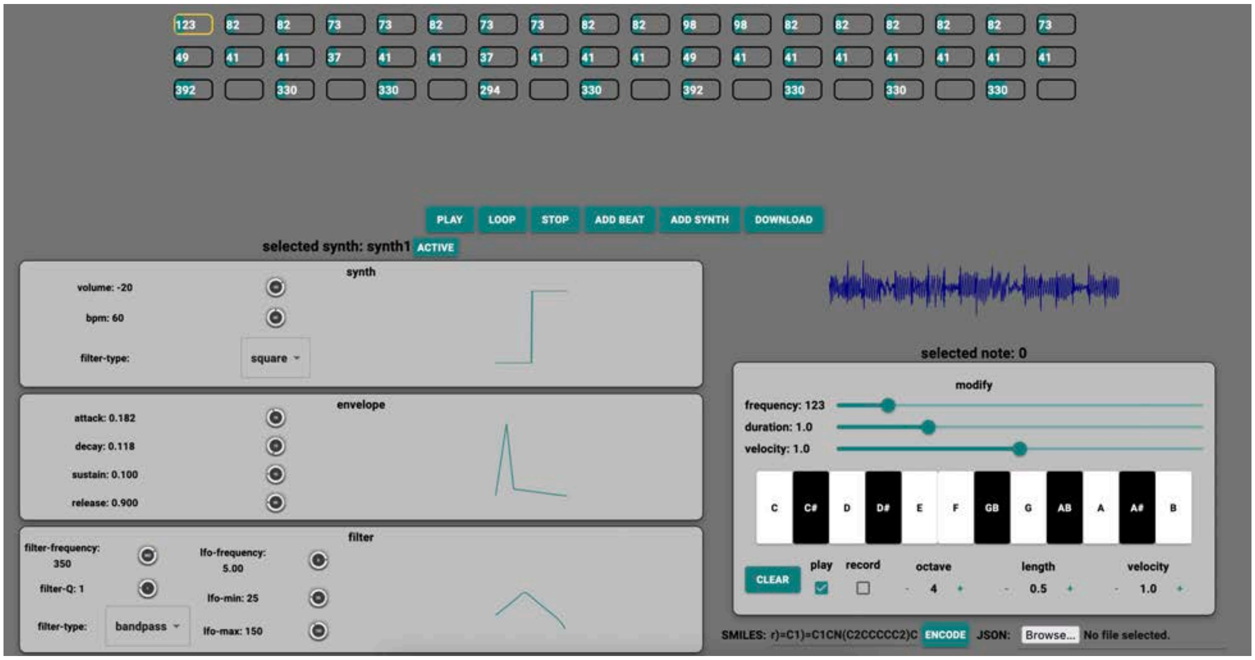
selected note: 0

frequency: 123
duration: 1.0
velocity: 1.0

play ☑ record ☐ octave 4 length 0.5 velocity 1.0

SMILES: O=C(N(C1=O)C)N(C2=C1N((JSON: No file selected.

100-Fig. 1 | The interface when caffeine is encoded. SMILES can be input in the bottom right to bootstrap a song. The number of notes corresponds to the size of the molecule, and additional notes and synths can be added at will. Notes can be changed using the interactive piano, and oscillators and filters can be modified on the left hand side.



123	82	82	73	73	82	73	73	82	82	98	98	82	82	82	82	82	73
49	41	41	37	41	41	37	41	41	41	49	41	41	41	41	41	41	41
392		330		330		294		330		392		330		330		330	

selected synth: synth1

volume: -20
bpm: 60
filter-type: square

attack: 0.182
decay: 0.118
sustain: 0.100
release: 0.900

filter-frequency: 350
filter-Q: 1
filter-type: bandpass
lfo-frequency: 5.00
lfo-min: 25
lfo-max: 150

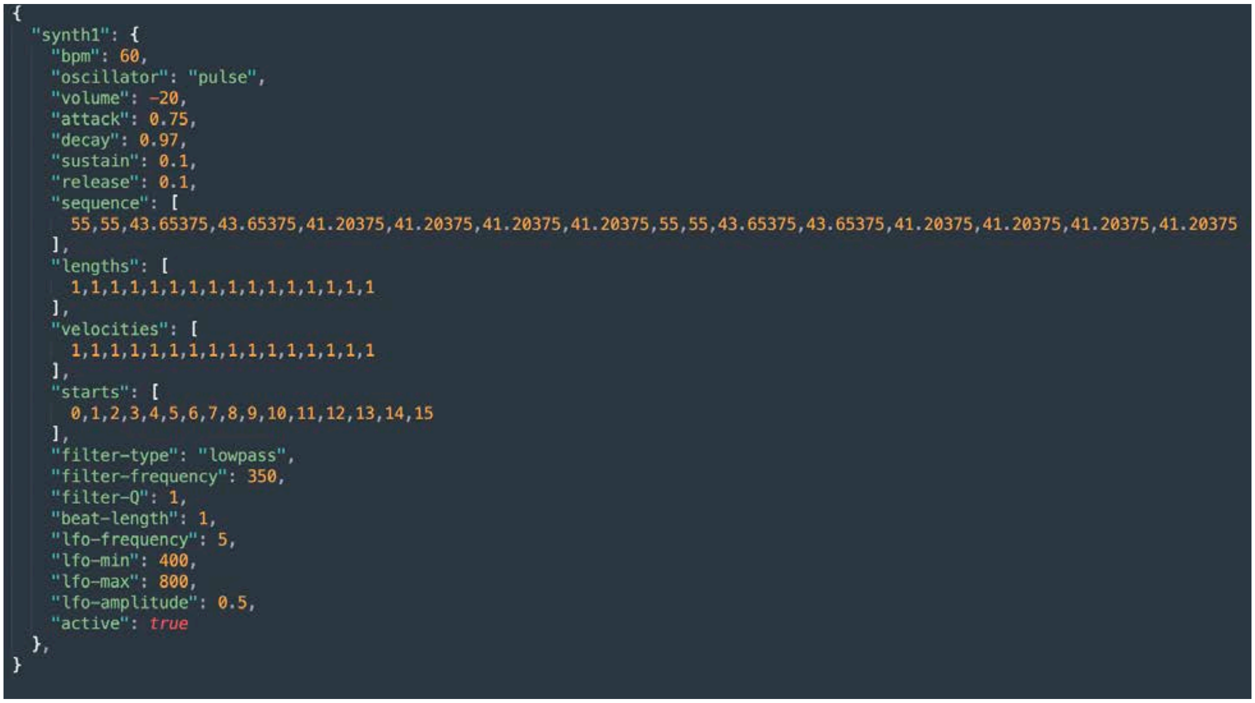
selected note: 0

frequency: 123
duration: 1.0
velocity: 1.0

play ☑ record ☐ octave - 4 + length - 0.5 + velocity - 1.0 +

SMILES: r)=C1)=C1CN(C2CCCCC2)C
JSON: No file selected.

101-Fig. 2 | The encoding of bromhexine produces a different musical output.



```
{
  "synth1": {
    "bpm": 60,
    "oscillator": "pulse",
    "volume": -20,
    "attack": 0.75,
    "decay": 0.97,
    "sustain": 0.1,
    "release": 0.1,
    "sequence": [
      55,55,43.65375,43.65375,41.20375,41.20375,41.20375,41.20375,55,55,43.65375,43.65375,41.20375,41.20375,41.20375,41.20375
    ],
    "lengths": [
      1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
    ],
    "velocities": [
      1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
    ],
    "starts": [
      0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
    ],
    "filter-type": "lowpass",
    "filter-frequency": 350,
    "filter-Q": 1,
    "beat-length": 1,
    "lfo-frequency": 5,
    "lfo-min": 400,
    "lfo-max": 800,
    "lfo-amplitude": 0.5,
    "active": true
  },
}
```

102-Fig. 3 | An abstractable data structure dictates the settings and notes for each synth. This data structure can be recreated in external scripts and dropped into the program to create the music.

Chapter VI. Informatics Tools and Software for Education and Automated Research Workflows

This chapter discusses informatics tools and education software that I have significantly contributed to or led during my PhD with relevance to the thesis of this dissertation. While the work described in this chapter and in this overall thesis is not comprehensive of my total efforts, the general outline and scientific progress discussed herein share a theme of advancement and education of medicinal chemistry in a systematic effort. It is my hope that any reader that reads to this point can recognize the dedication I have committed to my studies and towards general science.

Interactive Python Notebook Modules for Chemoinformatics in Medicinal Chemistry

In this project, I helped developed a lab section within the medicinal chemistry curriculum that taught undergraduate students chemoinformatics using python. A large effort was standardizing the software such that it was compatible with all computers and firmware that students were using. By the end of the class sessions, students had a firm understanding of basic chemoinformatic concepts in medicinal chemistry and the ability to execute some basic analyses given datasets.

Abstract

Data science is becoming a mainstay in research. Despite this, very few STEM graduates matriculate with even the most basic formal training in programming. This lesson plan was developed to introduce undergraduates studying chemistry or biology to chemoinformatics and data science in medicinal chemistry. The objective of this lesson plan is to introduce students to common techniques used in analyzing medicinal chemistry datasets, such as visualizing chemical space, filtering out non-Lipinski molecules, and principal component analysis. The content provided in this lesson plan is intended to serve as a tutorial-based reference for aspiring researchers to use in future research. The lesson plan is split into two three-to-four-hour class sessions, each with an introductory slide deck, Python notebook consisting of several modules, and lab report template. During this activity, students learned to parse medicinal chemistry datasets with Python, perform machine learning analyses, and develop interactive graphs. During each session, students completed the Python notebook protocol and fill out a lab report template after a short lecture. By the end of the lesson plan, students were able to generate and manipulate various plots of chemical space and reported to have increased confidence in their understanding of chemistry, Python, and data science.

Introduction

Chemoinformatics is the use of computational techniques to solve problems in chemistry. These in silico methods can be used to transform data into information and aid in the process of drug discovery. Recently, a rise in computational power and increased availability of developed tools have turned chemoinformatics into an invaluable tool for research. There has been recent interest in teaching young scientists how to work at the interface of physical science and data science.[309-314] In previous works, lesson plans have been developed incorporating programming into physical chemistry courses,[309, 311] a general chemistry course,[312] bioinformatics,[314] and an analytical chemistry course.[313] This paper extends the previous works through a lesson plan that introduces students to the basics of chemoinformatics in medicinal chemistry with the most popular scripting language, Python. Specifically, the exercises included in this activity are based on traditional techniques used by medicinal chemists to visualize and analyze chemical space.

Objectives

The purpose of this experiment is to introduce fundamental chemoinformatics using Python through medicinal chemistry-based exercises. The modules teach:

- how to load compiled medicinal chemistry datasets that is suitable for sharing and analysis
- how to visualize drug-like chemical space in a multitude of ways
- how to filter chemical datasets based on medicinal chemistry properties
- how to validate medicinal chemistry principles using data visualization
- how to simplify multidimensional physicochemical properties of drugs using principal component analyses (PCA)

Structure and Content

In this text, “lesson plan” refers to the entirety of this work and consists of two “class sessions”. Each “class session” consists of a brief lecture, to be administered at the beginning of the class session, a Python notebook consisting of several “modules” to be completed by students, and a lab report to be completed and submitted for a grade. As mentioned, this lesson plan is executed over two separate three-to-four-hour class sessions, supervised by one or two graduate student teaching assistants. Students participating in this activity are expected to have familiarity with basic concepts of chemistry, such as knowledge of molecules and druglike properties. Classes are typically between 20-40 students. Each class session consists of a brief slide deck and lecture introducing the very basics of Python and its capabilities, presented to the students, followed by an interactive Python notebook, composed of multiple modules.

These notebooks are written in Google Colaboratory (Colab)[315], an easily accessible online Python environment that executes code on the cloud for free, based on the popular Jupyter software.[316] Colab has several important data science packages preinstalled, including Pandas,[317] Numpy,[318] Scipy,[319] and Matplotlib[320] – Python add-ons that simplify the manipulation and visualization of data. Its primary advantages here are allowing a fast and simple way for new students to get started coding, as it is agnostic of computer, operating system and requires no technical setup. Code is separated into blocks called ‘cells’, which execute independently of each other. The two Colab notebooks are organized by modules, each with specific objectives and corresponding questions in the notebook’s respective lab report and walk the students through various exercises to meet the teaching objectives. Each notebook is to be completed alongside provided lab report templates consisting of module-specific questions and discussion items. Graduate student teaching assistants provide guidance and feedback to students during the class.

Class Session 1: Introduction to Colab, Python, and Chemoinformatics

In the first class session, an initial slide deck is presented by the graduate student teaching assistants to the class explaining the increasing popularity of scripting languages and their use in the industrial market and in academia. Several examples of data visualizations generated by Python are shared with the students (Fig. S1). Python as a scripting language is then formally introduced, as well as Colab; students are directly shown how to execute “print(“hello world”)” and are encouraged to log into Colab from their computer and run a line of code. Lists and dictionaries are introduced as two basic data structures. *If* statements and *for* loops are introduced through their utility in filtering a list of dictionaries. Finally, it is shown how the Python package Pandas[317] can be used to load tabulated datasets from CSVs or JSON files. Students then complete the first Python notebook and its corresponding lab report template after the brief lecture in the remaining class time with graduate student supervision.

The notebook given in the first class session introduces students to Google Colab, plotting in Python, and basic chemoinformatic concepts. The objectives of this notebook are to learn basic python coding and to quickly load and plot chemoinformatic data from spreadsheets or other data formats. The module also exemplifies how to customize plots generated in Python.

The first module walks through basic plotting in Python. The students receive a code template (103-Fig 1a, Fig S2), which generates a plot when executed in the notebook. Several lines have been deactivated, as indicated by lines containing green text and beginning with a “#”. Students are instructed to run the script, note what happens, then reactivate a line of code (by deleting the “#” symbol) and run the script again. This approach is intended to build familiarity with Colab and to exemplify how certain lines of code affects the script’s output.

As with the code template, most Python scripts begin with a list of package imports, lines of code containing the keyword “import”. Packages in Python are codebases that can be downloaded and utilized in new scripts. For instance, to avoid writing the code to render plots ourselves, Python is instructed to import the package “Matplotlib”, a well maintained and versatile graphing software. At this point forward in the code, functions (blocks of code that generate a specific response when executed) provided by the imported package can be utilized by referencing the desired function in a line of code. As seen in the first line of the code template, a package is imported through specific syntax: “import <package_name>” or “import <package_name> as <variable>”. In the first syntax, future references to the package must use the exact package name written in the import statement. In the second syntax, the “as” keyword instructs Python to allow the coder to refer to the package as “plt”, a “variable” that represents the package in code. Variables are abstract objects that are named by the user and can store data or references to previously written code in other files or installed packages. Variables are subsequently referred to in code by their given name.

In the following two lines, two lists are instantiated. Lists are variables that store data in a sequential order, similar to a row in a spreadsheet. The first list is named “asdf”, and stores the integers 1,2, and 3, in that order. The second list is named “qwer” and stores the integers 3,5, and 6 in sequential order. In the following line of code, the variables “fig” and “ax” are instantiated and set equal to the output of the function “plt.subplots()”. This “subplots” function is provided by the Python package “Matplotlib” (referred to in code as “plt”, as imported earlier) and provides a blank canvas to be later manipulated by the programmer. In Python, a function is asked for a response by “calling” it, which is done by appending parenthesis () to the function name – in this case, “plt.subplots”, where the period between “plt” and “subplots” indicates that “subplots” is a function defined in the package “plt” (short for “Matplotlib”). The response is captured in the variables “fig” and “ax”, as named by the user, and are later referred to in the code to draw and save a plot. The portion of the response that is stored in “fig” is described as “the top-level container for all the plot elements” in its documentation.[321] In essence, the variable “fig” now represents the digital data structure, stored in the computers random access memory, that will eventually render into a visual image. Similarly, “ax” represents an abstract object that is used in code to draw plots. The scatter function, found in the object represented by the variable “ax”, can be provided data to generate a scatter plot. As with before, the function is called by appending parenthesis to the name of the function, “ax.scatter()”.

Data is “passed” to the function as input through what is written between the parenthesis of the function call, otherwise known as the “parameters” of the function. In the case of the code template, 4 lines of code are written in-between the parentheses, each representing a parameter to the scatter function. Syntactically, parameters must be separated by commas, and new lines between parameters are optional but encouraged for legibility. The first parameter is passed in line six, where it is specified to the function that the x axis data is equal to the list of data that is represented by the variable “asdf”. In the following line, the y data is set to “qwer”. The final two lines of parameters are “commented out”, and thus deactivated. The scatter function requires both x axis data and y axis data to be provided. The remaining parameters are optional. Following a deactivated line of code after the scatter function, the final line of code is “plt.show()”, which instructs the notebook to render the plot. Thus, running the code as given generates the scatter plot shown in 103-Fig. 1B. By “uncommenting” the line “# s = 275”, students reactivate the size parameter of the scatter plot function. Rerunning the code now generates the plot of 103-Fig. 1C. Reactivating the color parameter creates the plot of 103-Fig. 1D. Students can then change the background color of the plot by using the “set_facecolor” function provided by “ax”. By uncommenting the line of code, a hexadecimal color code is passed as a parameter to the function, and results in the plot of 103-Fig. 1E. In the module, students are provided the documentation to the scatter function,[322] and are asked to add an additional parameter to change the shape of the points, generating a plot such as 103-Fig. 1F.

To expand on module one, the second module guides the students through creating a plot of a large dataset, which is provided as a datafile to the students. The datafile is a JavaScript Object Notation file (json file) and contains a list of dictionaries (similar to a list, but data is stored via key:value pairs, where the value is retrieved by passing the key to the dictionary). Each dictionary in the list contains information regarding a drug in DrugBank,[323] containing its SMILES as well as pre-calculated physicochemical properties. Students are instructed to upload their data file into Colab by dragging and dropping their file into the file system tab. (Fig. S3) With provided code, students are then instructed to import the data contained in the given file using the json package and to convert the data into a “DataFrame”, a data structure representation provided by the Pandas package, allowing for spreadsheet like manipulation of the dataset. (Fig. S4) Using a “for loop”, ten dictionaries from the json file are added to a list, and the “pd.DataFrame” function is used to store the data as a DataFrame. Students then print the contents of the DataFrame and inspect the contents of the json file. (104-Fig. 2, Fig. S5) Finally, another graphing template is provided, and the students must execute it to plot the drugs by their properties in an x-y scatter plot. In this template, a “for loop” is used to iterate through the DataFrame, and store the LogP, polar surface area, and quantitative estimate of druglikedness of each drug into respective lists that were initialized. Then, the scatter function is used to plot the drugs on a graph where the x axis represents the logP of the drug and the y axis represents the polar surface area of the drug. (Fig. S6) The list of quantitative druglikedness estimates is passed into the scatter function via the color parameter (c), coloring each point on the scatter plot by the drug’s druglikedness score.

Through this exercise, the Lipinski properties and their correlation to druglike molecules is explored, and students are exposed to a method of visualizing large chemical spaces.

In the third module, the student uses the code provided in module 2 to generate another plot that visualizes the uploaded DrugBank data but with a different color than before. In this module, additional physicochemical properties are introduced. Common parameters of druglikedness are included in the dataset such as LogP, polar surface area, number of aromatic rings, and hydrogen bond donors. The full list of included properties is listed in the supporting information. Concepts such as variables (user specified names that represent values stored in the computer's memory) and f-strings (a feature in Python that allows values to be embedded in text) are introduced here. The user can set the color variable to a column header from the DataFrame. Through function parameters, the color of the plot and the title of the output figure is automatically updated.

In the final module, students are asked to now modify the axes of the plots, which can be easily changed by switching the DataFrame header referred to in the code. The student also has the option to investigate in a third dimension by modifying the color of the plotted points. In the cumulation of this notebook, students are now comparing trends of physicochemical properties by modifying the x-axis, y-axis, and color attributes. Several plots generated by students are showcased in 105-Fig. 3.

This notebook is concluded with the following questions, which the students answer in the discussion of their lab report:

1. What are the Lipinski rules?
2. Write code for filtering by drugs that pass all the Lipinski rules
3. Suggest a research question that you could ask of the DrugBank dataset

These discussion questions evaluate student understanding of medicinal chemistry concepts such as druglike properties and their ability to manipulate datasets containing these properties into basic Python plotting code. It is expected students will utilize internet search engines to assist in the writing of filter code to answer the second question, to emphasize the importance of independent problem solving when encountering unfamiliar bugs or errors when coding. Filters can be written manually with *if* statements and *for* loops, as explained in the preliminary presentation, or using a function included in the Pandas package.

The second class session begins with another short lecture. In this lecture, the concept of machine-readable molecular representations is introduced through SMILES. Box plots are shown as a way to visualize molecular datasets via distributions of Lipinski[300] and other relevant medicinal chemistry properties. Finally, histograms and principal component analyses are introduced as other ways to analyze distributions of molecular data. The second notebook and its lab report are then completed after the lecture.

The notebook provided with this class session utilizes a dataset from Diamond XChem's COVID Moonshot project.[324] Inhibition data against the SARS-COV-2 main protease alongside precalculated physicochemical properties and SMILES for various inhibitors are included in a CSV provided to the students. The learning objectives of this notebook are to filter out unusable data, use data visualization to validate medicinal chemistry principles, perform basic statistical analyses, and simplify multidimensional data using principal component analysis.

In the first module, students are provided with the Moonshot CSV and is instructed to load and inspect the CSV using the Pandas package. (Fig. S7, Fig. S8) Some lines are missing data, and the concept of curating datasets is introduced. Using Pandas, entries without IC_{50} values are then filtered, and students are asked to record the number of remaining molecules in their report.

In the second module, the students are asked to validate GlaxoSmithKline's (GSK) Solubility Forecast Index[325] using the filtered dataset from the previous module. This module emphasizes the importance of certain physicochemical properties in compound solubility, an important concept in medicinal chemistry. (Fig. S9) The students are directed to use the template, plotting code to show how solubility is affected by the number of aromatic rings and LogP. (106-Fig. 4) In this case, the template script is missing lines of code, requiring students to fill the missing portions themselves. (Fig. S10) This requires students to correctly plot certain properties from the dataset, which is possible given an understanding of the components in the index and the script template, directly building on the concepts of the first lab exercise. By this point, students should be comfortable loading datasets into Python and creating and styling plots based on column headers and have growing familiarity with visualizing and exploring chemical space via relevant properties.

Also in this module, students are instructed to make box plots of various properties encoded in the datafile, providing an alternative visualization of the dataset. The code template is provided, and students are asked to analyze and modify the code to reshape the grid of boxplots. Students are encouraged to improve the plots aesthetically and to practice modifying the data that is plotted by changing the variables. Here the ability to plot multiple distributions into the same figure using Matplotlib's "subplots" function is introduced. As molecules can have many descriptors, multiplexed figures are essential in evaluating medicinal chemistry datasets.

In the third module, a principal component analysis (PCA, a technique that identifies patterns in multidimensional datasets by looking for the most important components) is performed on the dataset. The utility of a PCA is explained to the students within the lecture slides as a method to combine multiple features of a datapoint into a one or two axes, while the math is omitted. Students are instructed to create a matrix containing the Lipinski physicochemical properties for each entry in the dataset, resulting in a DataFrame with five columns and a row for each compound. Using the provided template code (Fig. S11), the students feed their matrix into scikit-learn's PCA decomposition algorithm, where the matrix is reduced to two dimensions. The resultant data is then plotted and colored by a property. Students are subsequently instructed to filter out non-Lipinski compliant molecules, rerun the reduction and compare the final graphs. (107-Fig. 5) To complete this objective, students must use what they have learned in the previous modules or in the lecture notes to write a "*for* loop" and use "*if* statements" to collect datapoints that pass the Lipinski filter rules. Alternatively, students are encouraged to search for Pandas documentation and to perform the data filtration using a DataFrame function. In this module students begin to familiarize with dimensionality reduction and the identification of features that lead to dataset variance.

In the final module, students use the filtered dataset and are introduced to a new package that allows for the creation of interactive plots. Using the template code, students use the package plotly to generate an interactive PCA that displays SMILES and other information for each plotted entry. Plotly works similarly to the Matplotlib package, but instead produces an interactive scatter plot, where datapoints can be hovered over with the cursor to inspect additional information. Students are asked to engage with the interactive scatter plot and record several of molecules from different clusters.

Participants

The participants in this study were students enrolled in a senior level undergraduate medicinal chemistry course. Anecdotally, through in-class surveys, we found that nearly all participants had little to no previous coding experience at the time of the study, and these modules were their first introduction to a hands-on coding activity. This study was developed and conducted over five years of students; in an in-person and remote format. In the first two years, only the first module was taught. Over 100 students have participated in the study.

Implementation

This lesson plan was given to undergraduate classes of pharmaceutical science majors in mid-semester of their senior year, after several units introducing basic medicinal chemistry concepts are completed. In our case, each class consisted of around 30 students and were supervised by two graduate student teaching assistants. As mentioned, the lesson plan is split into two three-to-four-hour class sessions. At the beginning of each class session, graduate student teaching assistants present a lecture and slide deck with basic coding and Python concepts before the students begin working on the notebook corresponding to the class session (notebook 1 for the first class, notebook 2 for the second class). The initial implementation of this activity was done without Colab, with tutorials provided to install Python on each student's personal computers. The diverse operating systems and security settings of individual required some installation oversight for a handful of students each year. With the introduction of Colab to the academic community, the onboarding process for the activity was greatly simplified as it became guaranteed any student could complete the activities agnostic of personal hardware or software. During the class sessions, graduate student teaching assistants are available for troubleshooting and questions. In our implementation, the most common problem students encountered were syntax issues in their code. While the templates and instructions provided were meant to minimize syntax errors, it was beneficial for the graduate student teaching assistants to have familiarity with common Python error messages to quickly identify the source of the bug and to direct the students to correct the invalid syntax. Students are permitted to work in groups to complete the exercises, and most students were able to independently complete all exercises without supervision or additional guidance from an instructor.

Assessment of Effectiveness

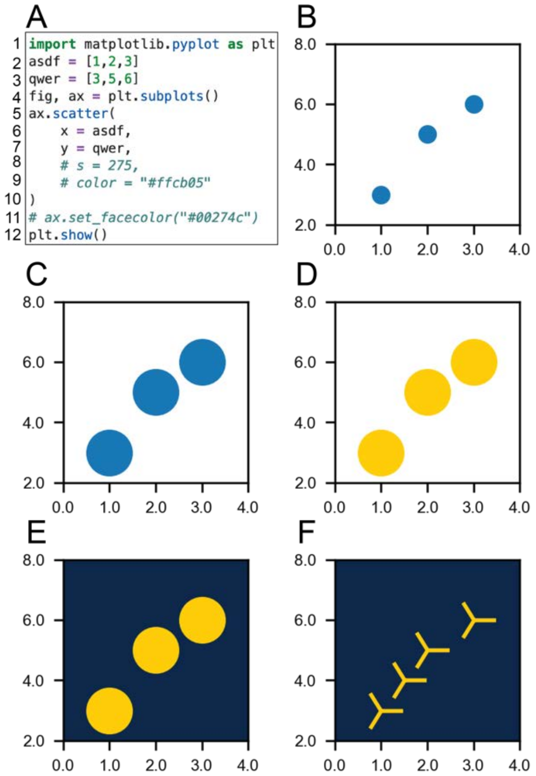
As of 2023, nearly all senior undergraduates who have enrolled in the class have reported having little to no experience in coding or programming prior to taking this lab exercise. Throughout the lesson plan, students submitted plots they have generated to the graduate student teaching assistants alongside their code to be graded for accuracy. By the end of the two classes, each student was able to successfully use the code to generate various plots of chemical space that were correctly labeled. Grades for submitted lab reports were consistently above 90%, with competent students able to produce plots from text prompts and no provided template. Furthermore, after the lab sessions, students from the last two years lab were asked to complete the following five-question survey:

1. This exercise improved my understanding of chemistry
2. This exercise improved my understanding of Python
3. This exercise improved my understanding of chemical space
4. This exercise improved my data science
5. I enjoyed this exercise

Based on the responses received and the high marks earned by the students on their lab reports, we conclude that the current state of the module is effective in improving the understanding of basic data science and informatics in medicinal chemistry for most students (108-Fig. 6).

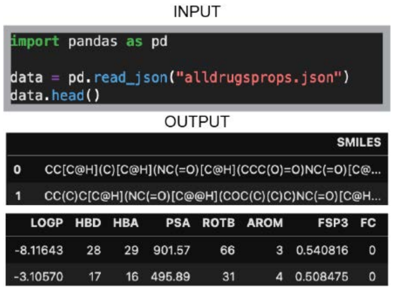
Summary

A lesson plan to teach undergraduates the basics of data science in medicinal chemistry was developed and validated over several semesters. Over the course of two lab sessions, students are introduced to Python, Google Colaboratory, and several Python packages. Students learn these tools through guided, interactive modules that begin at learning how to function Colab and ends with developing a program that reads abstractable datasets and generates user-interactive data analytics through Python. An assessment reveals the lesson plan seems to be effective in improving student's familiarity with modern chemoinformatic tools and concepts. We postulate that the lesson plan may be effective for younger students and can be completed by anyone with a computer and internet access.



```
 1 import matplotlib.pyplot as plt
 2 asdf = [1,2,3]
 3 qwer = [3,5,6]
 4 fig, ax = plt.subplots()
 5 ax.scatter(
 6     x = asdf,
 7     y = qwer,
 8     # s = 275,
 9     # color = "#ffcb05"
10 )
11 # ax.set_facecolor("#00274c")
12 plt.show()
```

103-Fig. 1 | (a) A self-contained code template generates several plots and is provided to students. Several lines are deactivated, as indicated by green text. Lines of code can be reactivated by removing the leading “#”, resulting the code to produce a different visual output when executed. (b) The plot generated by the code in (a) when run as is. The three plotted points are defined in the code template, connecting the code to its visual output. (c) The plot generated when line 8 of (a) is activated, resulting in larger points. (d) The plot generated when line 9 is also activated, resulting in different point colors. (e) The plot generated when line 11 is activated, creating the same scatter plot with a blue background. (f) A potential plot generated after students are instructed to include an additional parameter to the scatter function. In this case, the shape of the points are changed.



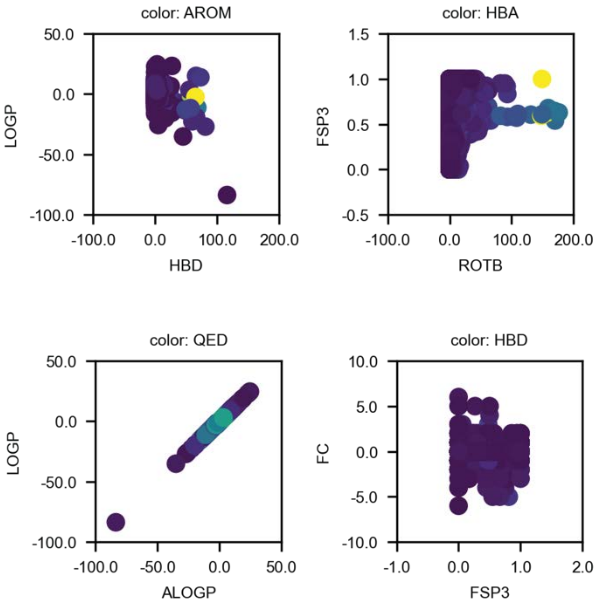
INPUT

```
import pandas as pd

data = pd.read_json("alldrugsprops.json")
data.head()
```

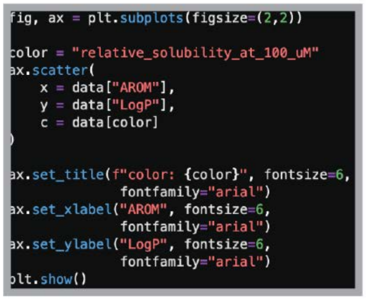
OUTPUT

```
                                                          SMILES
0   CC[C@H](C)[C@H](NC(=O)[C@H](CCC(O)=O)NC(=O)[C@...
1   CC(C)C[C@H](NC(=O)[C@@H](COC(C)(C)C)NC(=O)[C@H...

    LOGP  HBD  HBA     PSA  ROTB  AROM      FSP3  FC
-8.11643   28   29  901.57    66     3  0.540816   0
-3.10570   17   16  495.89    31     4  0.508475   0
```

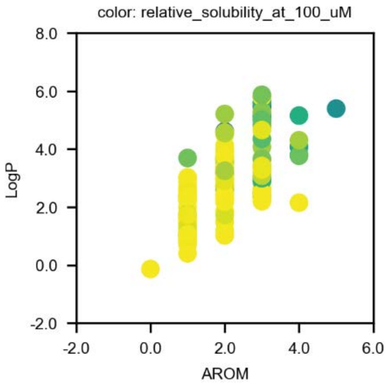
104-Fig. 2 | Students are provided with the code to import any tabular JSON file. The utility of the package Pandas is used in reading tables programmatically. The “alldrugsprops.json” file is provided to students, and provided with a template to import the data into Python, as seen in the input block. Students then inspect the contents of the file with Python, revealing a datafile of over 9,000 drugs and their properties downloaded from DrugBank.



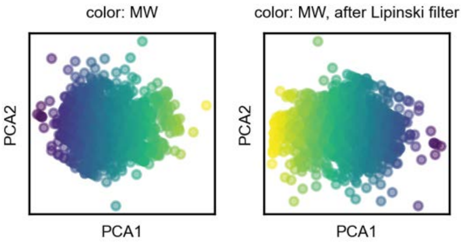
105-Fig. 3 | Four chemoinformatic experiments run by students during the first session. In each graph, over 9,000 drugs are graphed onto scatter plots as point, where their x and y locations correspond to calculated properties of the drug. Students learn how variables can be used to rapidly change the axes and colors of the plot, exposing them to a strategy to rapidly investigate chemical space and generate reports.



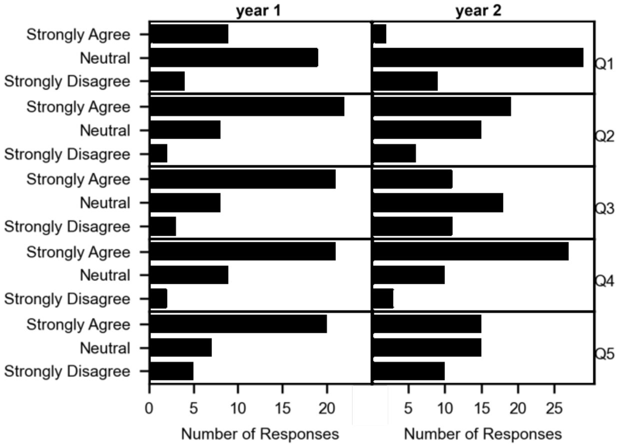
```
fig, ax = plt.subplots(figsize=(2,2))

color = "relative_solubility_at_100_uM"
ax.scatter(
    x = data["AROM"],
    y = data["LogP"],
    c = data[color]
)

ax.set_title(f"color: {color}", fontsize=6,
              fontfamily="arial")
ax.set_xlabel("AROM", fontsize=6,
              fontfamily="arial")
ax.set_ylabel("LogP", fontsize=6,
              fontfamily="arial")
plt.show()
```



106-Fig. 4 | (a) The template plotting code is incrementally improved until it can be used effectively to make manuscript-ready graphics. This code creates a scatter plot of the Moonshot compounds provided in the CSV, where the x axis represents a compound's number of aromatic rings, and the y axis represents its LogP. (b) By coloring the points by the compound's recorded solubility, a trend is revealed where compounds with fewer aromatic rings and lower LogP are more soluble. Validating GSK's Solubility Forecast Index is a simple experiment to allow students to build confidence in their ability to manipulate and analyze datasets.



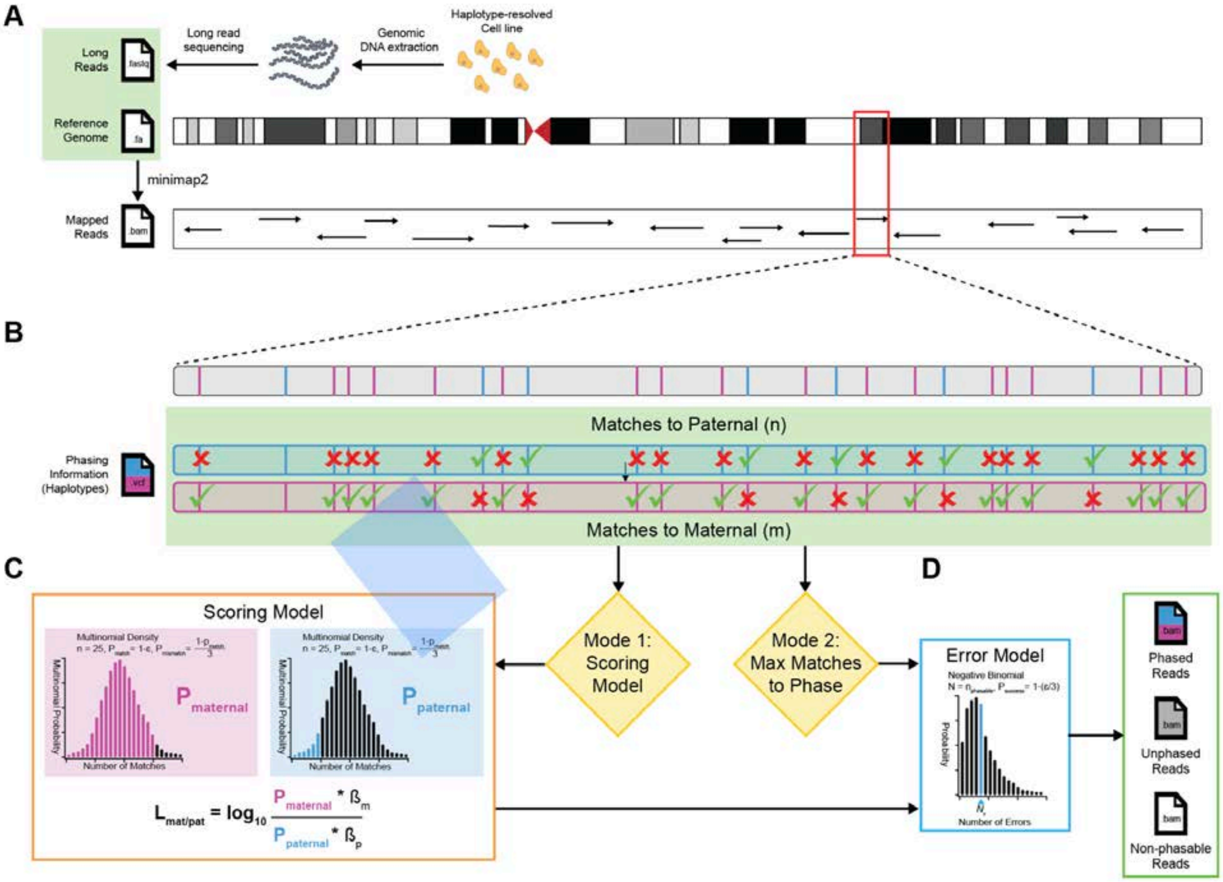
107-Fig. 5 | By the end of the modules, students have implemented a data filter and a principal component analysis on a dataset of SARS-COV-2 Main Protease inhibitors. (a) the PCA before the data filter is applied. Points represent DrugBank compounds and are colored by their molecular weight. By changing the color of the points, the correlation of the reduced axes can be mapped to specific features of the dataset. In this instance, molecular weight tracks with the first principal component, increasing as the x value increases. (b) the PCA after the Lipinski filter is applied. Outliers are removed by the filter, and the distribution of molecular weights in the first principal component becomes wider.



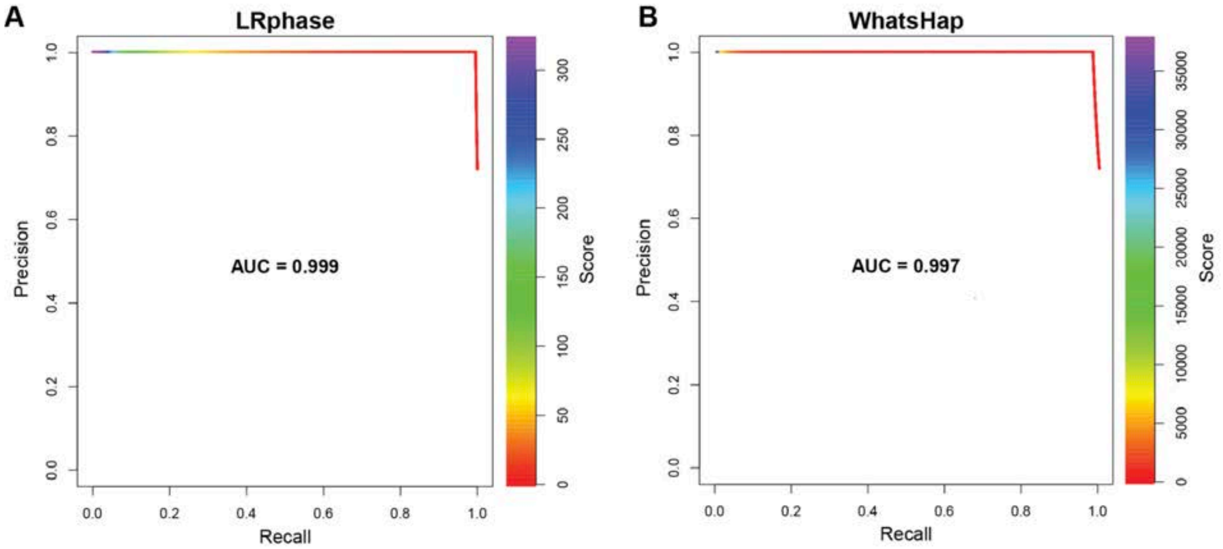
108-Fig. 6 | Student feedback to the lesson plan collected after the lab. Questions were intended to gauge the student's perspective on their learning experience. In the first year, students felt strongly that their understanding of Python, chemical space, and data science were improved. In the second year, students felt strongly that their understanding of data science improved but were neutral on other learning objectives. The difference between the two years may be accounted for by the level of involvement the graduate student teaching assistance had in developing the course. In the first year, the class was taught by the writer of the lesson plan; the second year was taught by assistants who were given the lesson plan to teach.

LRphase: an efficient algorithm for haplotagging long reads

For a class project, I wrote an algorithm that phased long-read DNA samples by statistically matching them to the maternal or paternal allele based on the number of associated single nucleotide polymorphisms. In collaboration with the Boyle lab in the bioinformatics department, the package was extended into a full research paper. This is a classical bioinformatic package that can be installed with pip and used in typical bioinformatic pipelines.



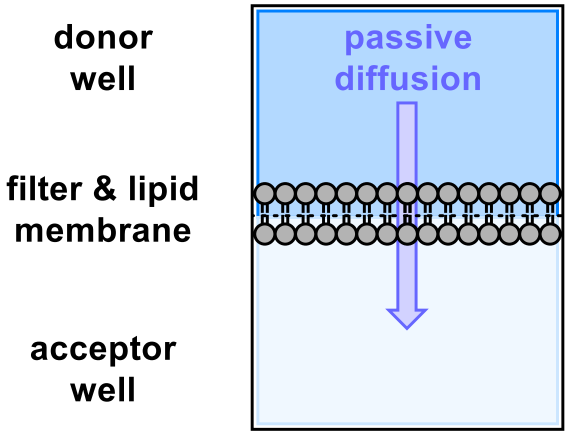
109-Fig. 1 | Overview of LRphase. A. Long sequencing reads are first prepared from genomic DNA fragments isolated from cells with available phased variant data for all parental haplotypes. Reads are mapped to the reference genome, either within LRphase with minimap2, or externally using any desired mapping/filtering workflow, with mapped reads supplied in BAM format. B. Haplotagging begins by intersecting individual mapped reads with known phased, heterozygous single-nucleotide variants (SNVs) supplied as a VCF file. The number of matches and mismatches are counted for both maternal and paternal phases and resulting counts are subsequently used to choose the best match to phase using one of the two scoring modes. C. Reads may be haplotagged using either the scoring model (Mode 1, orange box), or by simply assigning to the phase with the greatest number of matching alleles across all heterozygous sites (Mode 2). In Mode 1, match and mismatch counts are used to calculate log-likelihoods for matching to maternal and paternal phases. Likelihoods are computed as multinomial probabilities ($P_{maternal}$ and $P_{paternal}$ respectively), representing the aggregate probability of obtaining the observed number of matches and mismatches to either phase given the observed sequencing error rate, ϵ. $P_{maternal}$ and $P_{paternal}$ are adjusted by a Bayesian prior ($ß_m$ and $ß_p$) and log-likelihood ratios (LLRs) $L_{mat/pat}$ and $L_{mat/pat}$, are calculated to determine which phase has the most evidence for matching a given read. Reads are assigned to the phase corresponding to the greater of $L_{mat/pat}$ and $L_{mat/pat}$, are labeled “unphased” if there is a tie, or “nonphasable” if there are no intersecting heterozygous variants. D. After all reads are haplotagged, the false-discovery rate (FDR) is controlled by calculating the expected number of incorrect haplotags, $N\varepsilon$, as the mean of the negative-binomial distribution with N = the number of phaseable reads and $P_{success} = 1-(\epsilon/3)$. Phaseable reads are sorted by LLR and the $N\varepsilon$ * (1-FDR) lowest-scoring reads are relabeled as unphased. The remaining haplotagged reads are expected to include errors at a rate corresponding to the specified FDR. Finally, haplotagged results are written to output file(s) in BAM format, with the assigned haplotype stored in the “HP” tag, while custom tags are used to store all variables used in LLR calculations.



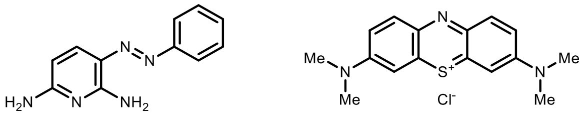
110-Fig. 2 | Haplotagging performance of LRphase and WhatsHap. Precision-Recall (PR) curves are used to assess the ability of LRphase (A) and WhatsHap (B) to identify the correct haplotag. Curves describe the fraction of correct and incorrect haplotag assignments across varying score thresholds. Individual points along each curve are color coded to illustrate score ranges at each combination of precision and recall. A. PR curve for LRphase haplotag assignments. B. PR curve for WhatsHap haplotag assignments.

An Automated and Colorful PAMPA Assay To Investigate Drug Permeability

In this work, we introduce students to a popular early-stage tool to determine the permeability of compounds called the parallel artificial membrane permeability assay (PAMPA). Students use PAMPA to determine the permeability of two brightly colored drug molecules, phenazopyridine and methylene blue. The students use the Opentrons OT2 robot, an inexpensive liquid handling robot that can be controlled using simple python code, to perform the assay.



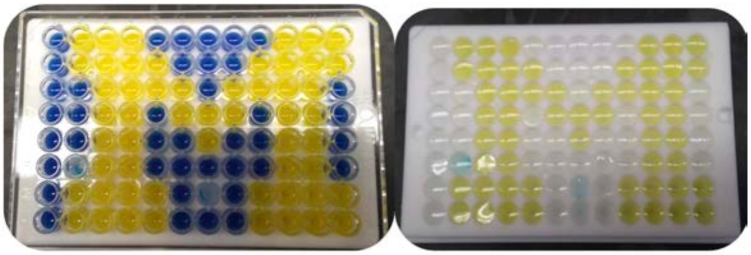
111-Fig. 1| A representative cartoon of how PAMPA works.



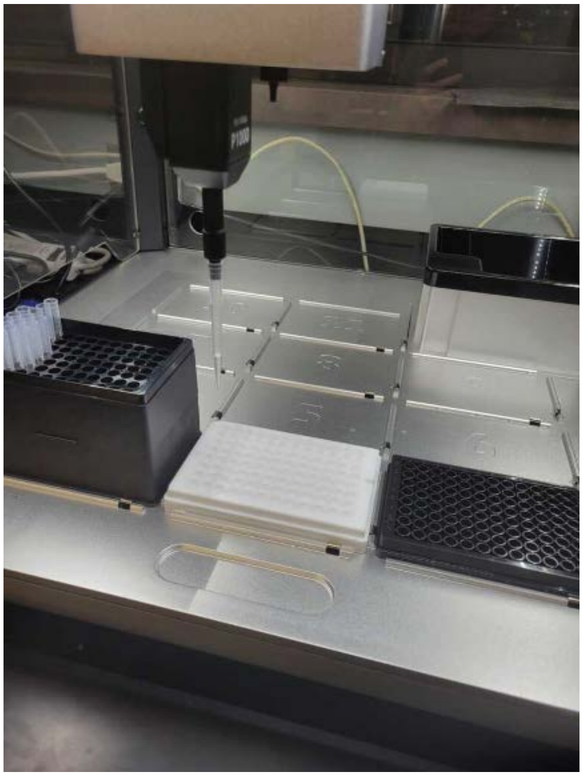
Phenazopyridine (1)

Methylene Blue (2)

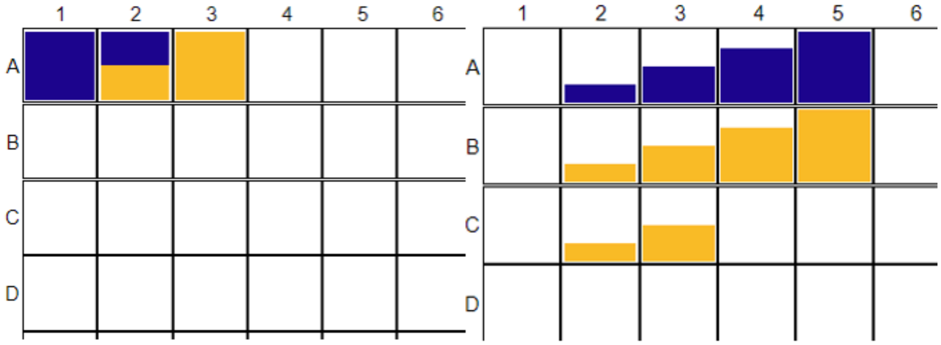
112-Fig. 2 | Structures of phenazopyridine and methylene blue.



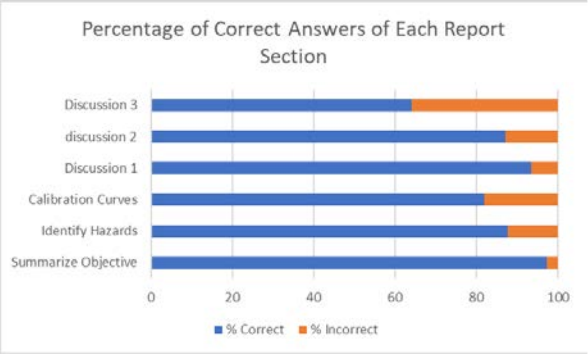
113-Fig. 3 | A PAMPA plate donor plate (left) and the acceptor plate (right) after a one-hour incubation. Note only phenazopyridine is permeable.



114-Fig. 4 | The setup of the Opentrons OT-2 layout.



115-Fig. 5 | Top: Opentrons transferring from PAMPA plate to plate reader plate. Right: PAMPA (white, source) and plate reader (black, destination) plates for Opentrons dosing. Blue represents methylene blue and yellow represents phenazopyridine. The height of the color represents relative concentration. The source plate shows what was placed into the donor plate. Row A in acceptor plate is methylene blue calibration curve, Row B is phenazopyridine calibration curve, and row C is what permeated through the donor plate.



116-Fig. 6 | Distribution of students correctly answering each part of the lab report. N=75

Chapter VII. Software Packages

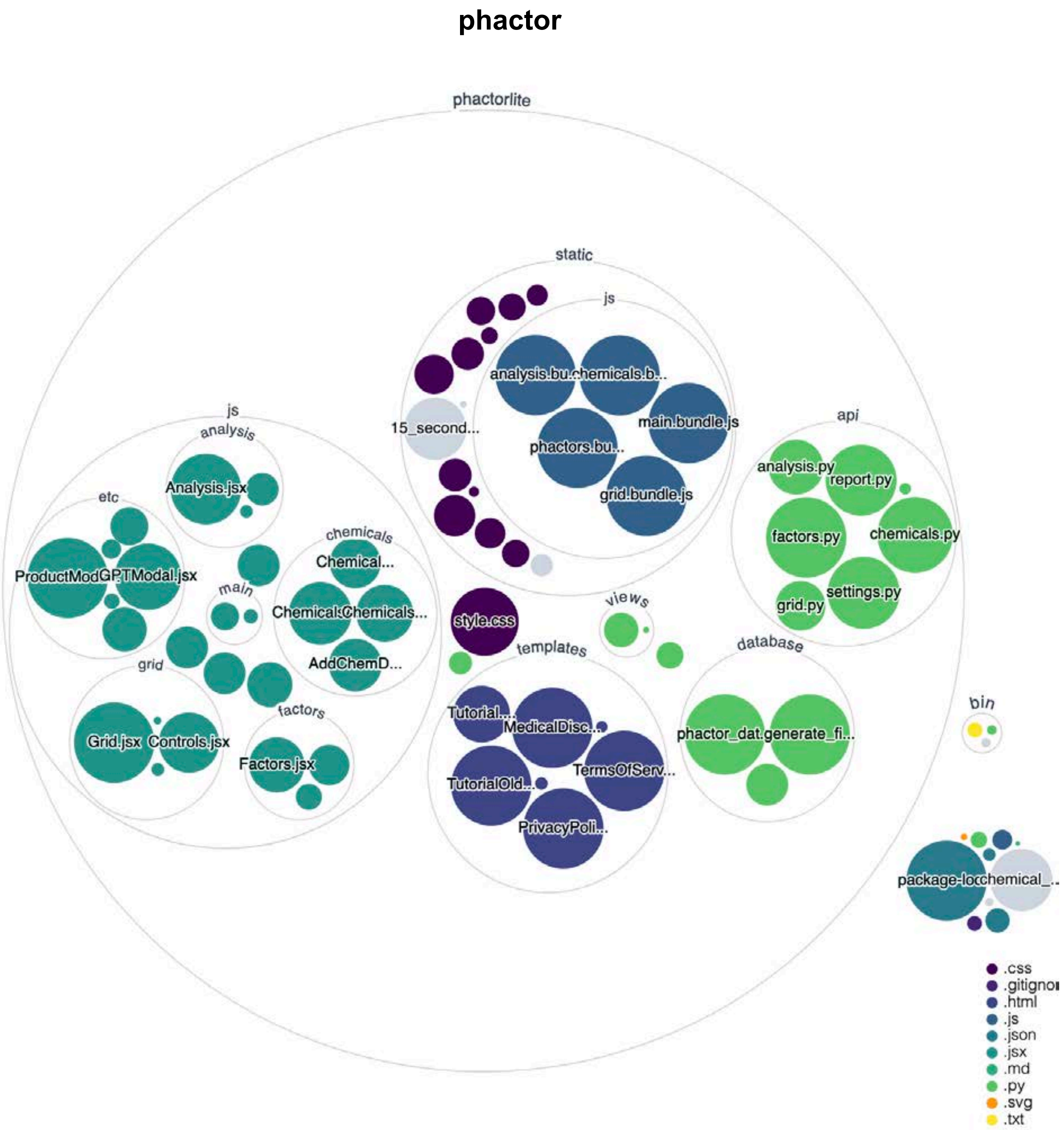
During my graduate studies, I composed an extensive array of codes and software packages. While this chapter does not encompass all the software I have crafted, it does include those that have reached a significant level of completion. My undergraduate training as a full-stack developer proved invaluable, allowing me to expand many of my research projects and codebases to the web. The browser served as a robust interface for interactive visualizations and workflows. All the software documented in this chapter was developed as Python/React packages, wherein the backend and API were built using Python, while React was used for developing the front-end graphics.

As I delved deeper into the development of these codebases, I gradually became adept at structuring data to enhance connectivity between the front and back ends. This led to a more efficient development process and a streamlined code infrastructure. The core philosophy I adopted involved maintaining all information within a single JSON object. This object could be seamlessly passed between the front and back end, and stored as memory on the front end in the browser. Such an approach eliminated the need for updating any database with each user action or the use of back-end session memory leading to cache memory build-up. Instead, long-term storage could be facilitated by saving the JSON object on demand. The reloading of this object would subsequently recreate the state of the software.

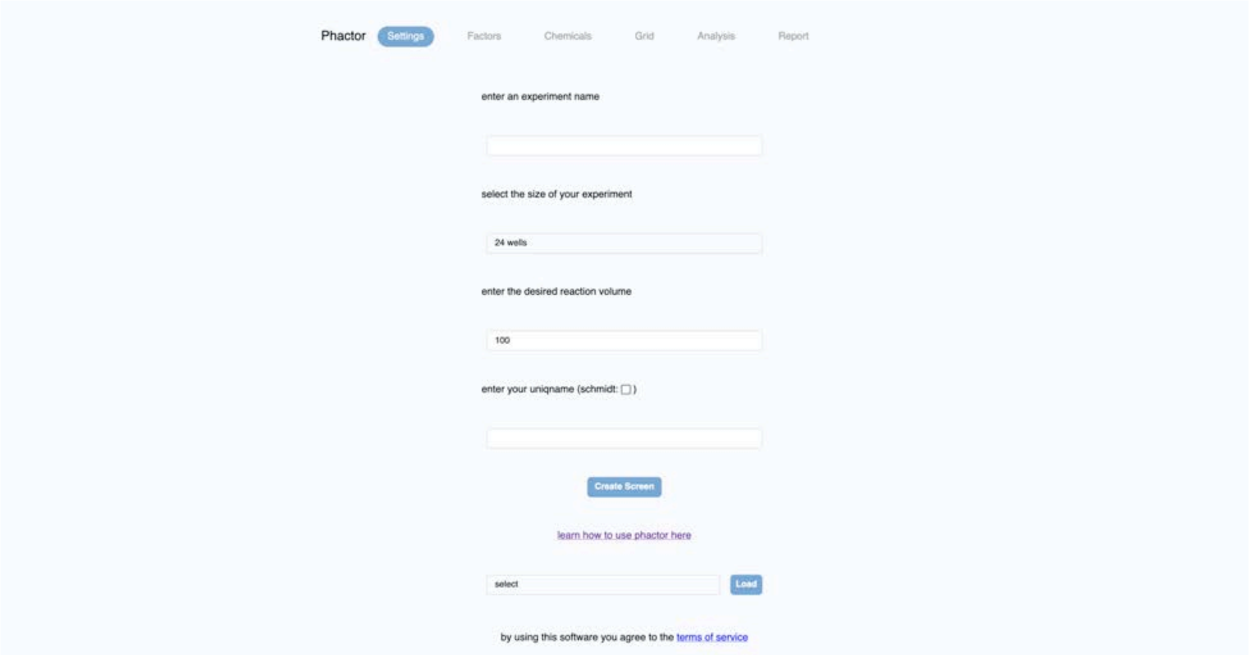
This methodology also proved beneficial for larger infrastructures, where independent software could generate the data object to be input into another software. Consequently, communication between two distinct software could be abstracted into a single data format. This eliminated the need for a developer to understand the details or structure of the other software. A practical demonstration of this philosophy can be seen in the database software I have developed for managing the lab's reagent and compound inventory. Through this web app, the required data structure can be directly integrated into phactor's data structure, facilitating the creation of screen designs without necessitating an understanding of phactor's internals.

The software described in this chapter are:

1. Phactor – the high throughput experimentation manager
2. Phactor inventory – a database webapp managing the lab’s reagent and compound collection
3. SAMPLES – the molecule to music sonification webapp
4. Retrosynthetic visualizer – the webapp visualizing the routes analyzed in our COVID-19 work



117-Fig. 1 | A visualization of phactor's codebase.



enter an experiment name

select the size of your experiment: 24 wells

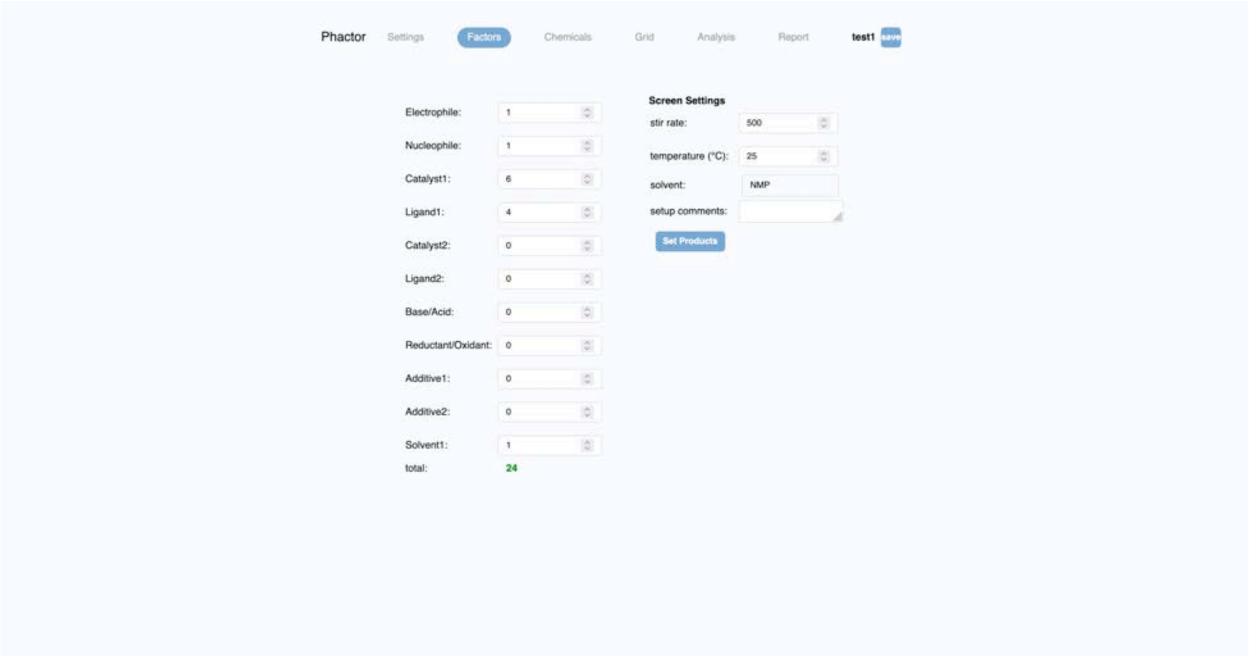
enter the desired reaction volume: 100

enter your uniqname (schmidt: ☐)

select

by using this software you agree to the terms of service

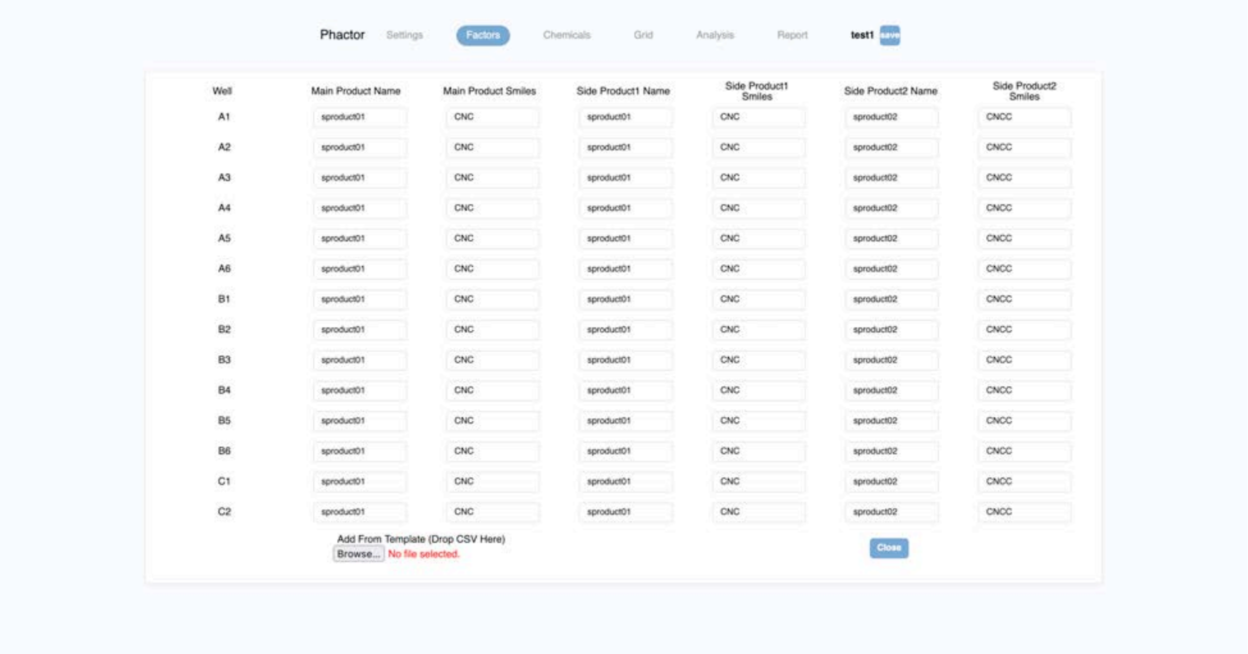
118-Fig. 2 | The settings page of phactor.



Electrophile: 1
Nucleophile: 1
Catalyst1: 6
Ligand1: 4
Catalyst2: 0
Ligand2: 0
Base/Acid: 0
Reductant/Oxidant: 0
Additive1: 0
Additive2: 0
Solvent1: 1
total: 24

Screen Settings
stir rate: 500
temperature (°C): 25
solvent: NMP
setup comments:

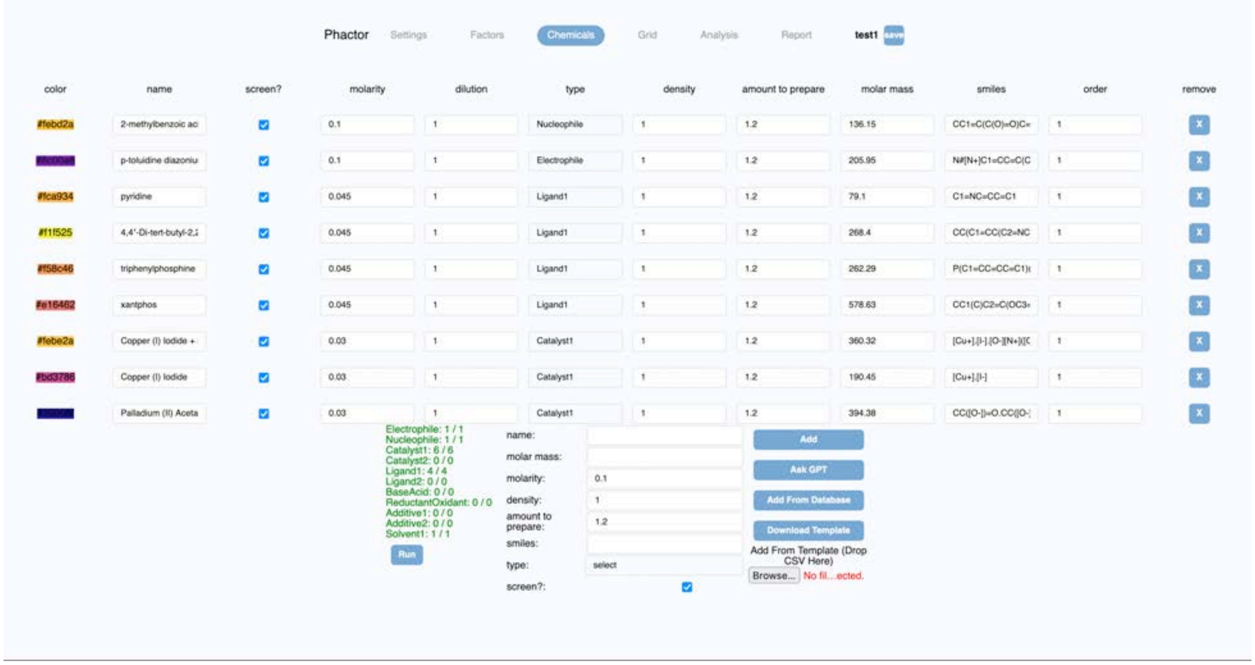
119-Fig. 3 | The factors page of phactor.



Well	Main Product Name	Main Product Smiles	Side Product1 Name	Side Product1 Smiles	Side Product2 Name	Side Product2 Smiles
A1	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
A2	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
A3	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
A4	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
A5	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
A6	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B1	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B2	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B3	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B4	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B5	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
B6	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
C1	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC
C2	sproduct01	CNC	sproduct01	CNC	sproduct02	CNCC

Add From Template (Drop CSV Here)
Browse... No file selected.

120-Fig. 4 | The product input pop-up page of phactor on the factors stage.



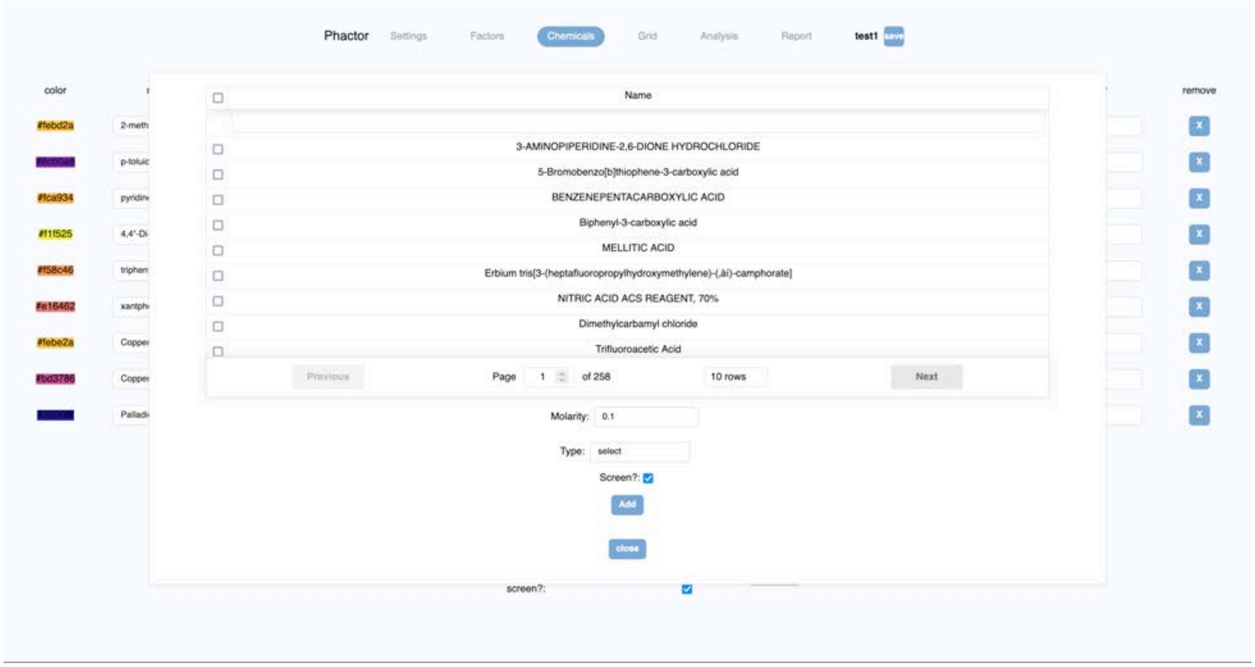
color	name	screen?	molarity	dilution	type	density	amount to prepare	molar mass	smiles	order	remove
#febd2a	2-methylbenzoic ac	☑	0.1	1	Nucleophile	1	1.2	136.15	CC1=C(C(O)=O)C=	1	X
[illegible]	p-toluidine diazoniu	☑	0.1	1	Electrophile	1	1.2	205.95	N#[N+]C1=CC=C(C	1	X
#fca934	pyridine	☑	0.045	1	Ligand1	1	1.2	79.1	C1=NC=CC=C1	1	X
#f1f525	4,4'-Di-tert-butyl-2,2	☑	0.045	1	Ligand1	1	1.2	268.4	CC(C1=CC(C2=NC	1	X
#f58c46	triphenylphosphine	☑	0.045	1	Ligand1	1	1.2	262.29	P(C1=CC=CC=C1)(	1	X
#e16462	xantphos	☑	0.045	1	Ligand1	1	1.2	578.63	CC1(C)C2=C(OC3=	1	X
#febe2a	Copper (I) Iodide +	☑	0.03	1	Catalyst1	1	1.2	360.32	[Cu+].[I-].[O-][N+]([C	1	X
#bd3786	Copper (I) Iodide	☑	0.03	1	Catalyst1	1	1.2	190.45	[Cu+].[I-]	1	X
[illegible]	Palladium (II) Aceta	☑	0.03	1	Catalyst1	1	1.2	394.38	CC([O-])=O.CC([O-]	1	X

Electrophile: 1 / 1
Nucleophile: 1 / 1
Catalyst1: 6 / 6
Catalyst2: 0 / 0
Ligand1: 4 / 4
Ligand2: 0 / 0
BaseAcid: 0 / 0
ReductantOxidant: 0 / 0
Additive1: 0 / 0
Additive2: 0 / 0
Solvent1: 1 / 1

name:
molar mass:
molarity: 0.1
density: 1
amount to prepare: 1.2
smiles:
type: select
screen?: ☑

Add From Template (Drop CSV Here)
Browse... No fil...ected.

121-Fig. 5 | The chemicals stage of phactor with compounds input.



☐	Name
☐	3-AMINOPIPERIDINE-2,6-DIONE HYDROCHLORIDE
☐	5-Bromobenzo[b]thiophene-3-carboxylic acid
☐	BENZENEPENTACARBOXYLIC ACID
☐	Biphenyl-3-carboxylic acid
☐	MELLITIC ACID
☐	Erbium tris[3-(heptafluoropropylhydroxymethylene)-(,ãi)-camphorate]
☐	NITRIC ACID ACS REAGENT, 70%
☐	Dimethylcarbamyl chloride
☐	Trifluoroacetic Acid

Page 1 of 258 10 rows

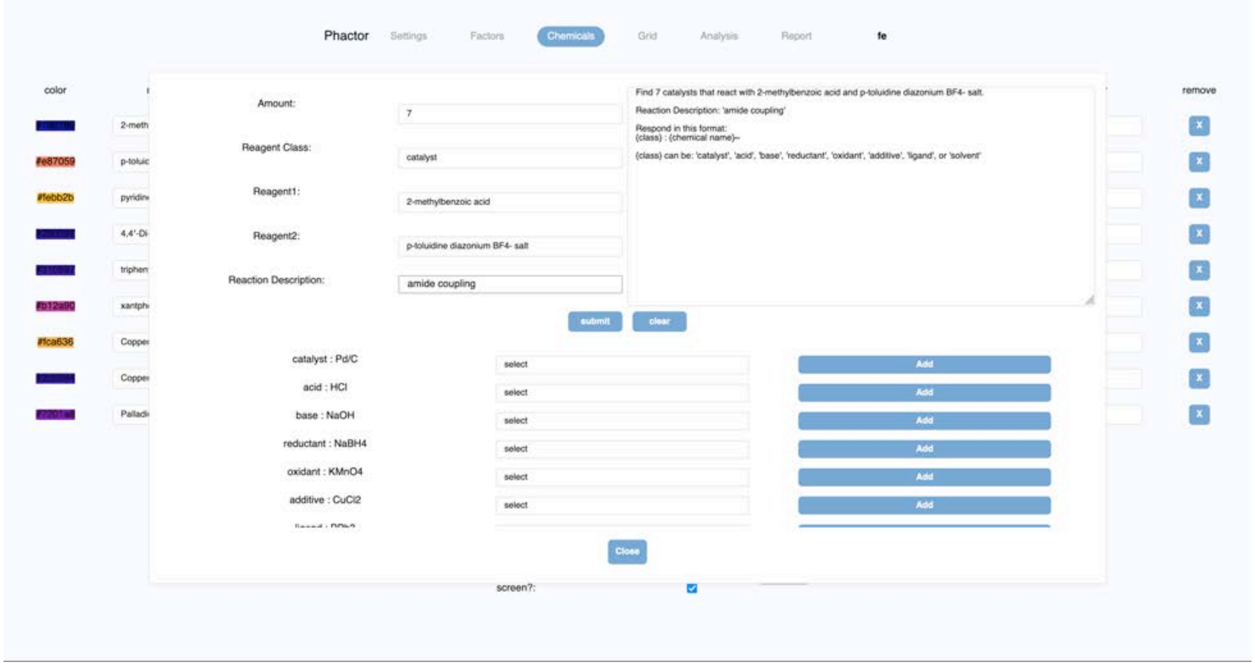
Molarity: 0.1

Type: select

Screen?: ☑

screen?: ☑

122-Fig. 6 | The connected inventory input pop-up page of phactor on the chemicals stage.



Amount: 7

Reagent Class: catalyst

Reagent1: 2-methylbenzoic acid

Reagent2: p-toluidine diazonium BF4- salt

Reaction Description: amide coupling

Find 7 catalysts that react with 2-methylbenzoic acid and p-toluidine diazonium BF4- salt.

Reaction Description: 'amide coupling'

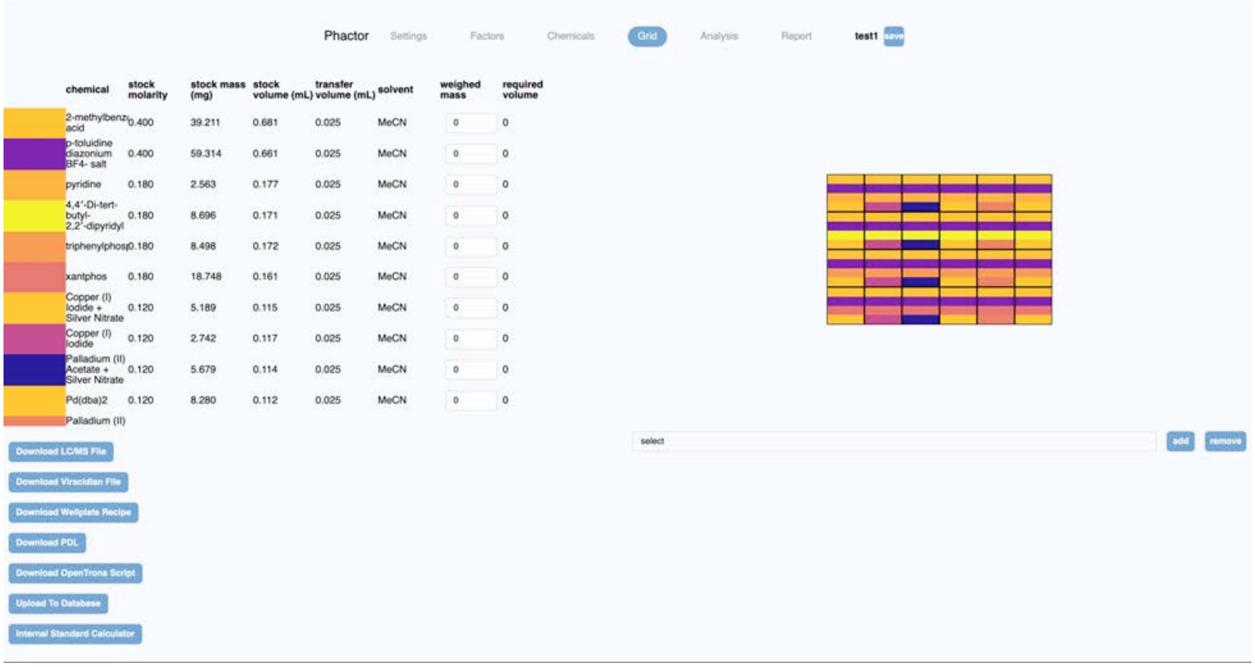
Respond in this format:
{class} : {chemical name}~

{class} can be: 'catalyst', 'acid', 'base', 'reductant', 'oxidant', 'additive', 'ligand', or 'solvent'

catalyst : Pd/C select
acid : HCl select
base : NaOH select
reductant : NaBH4 select
oxidant : KMnO4 select
additive : CuCl2 select

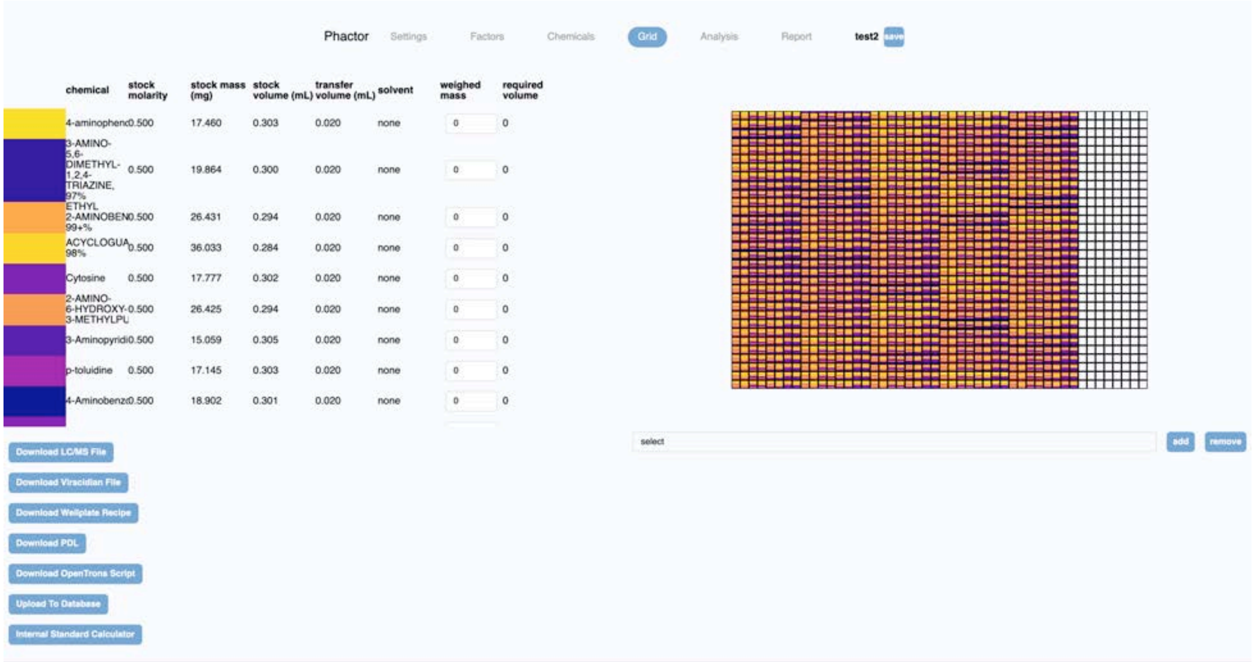
screen?: ☑

123-Fig. 7 | The connected ask GPT input pop-up page of phactor on the chemicals stage.



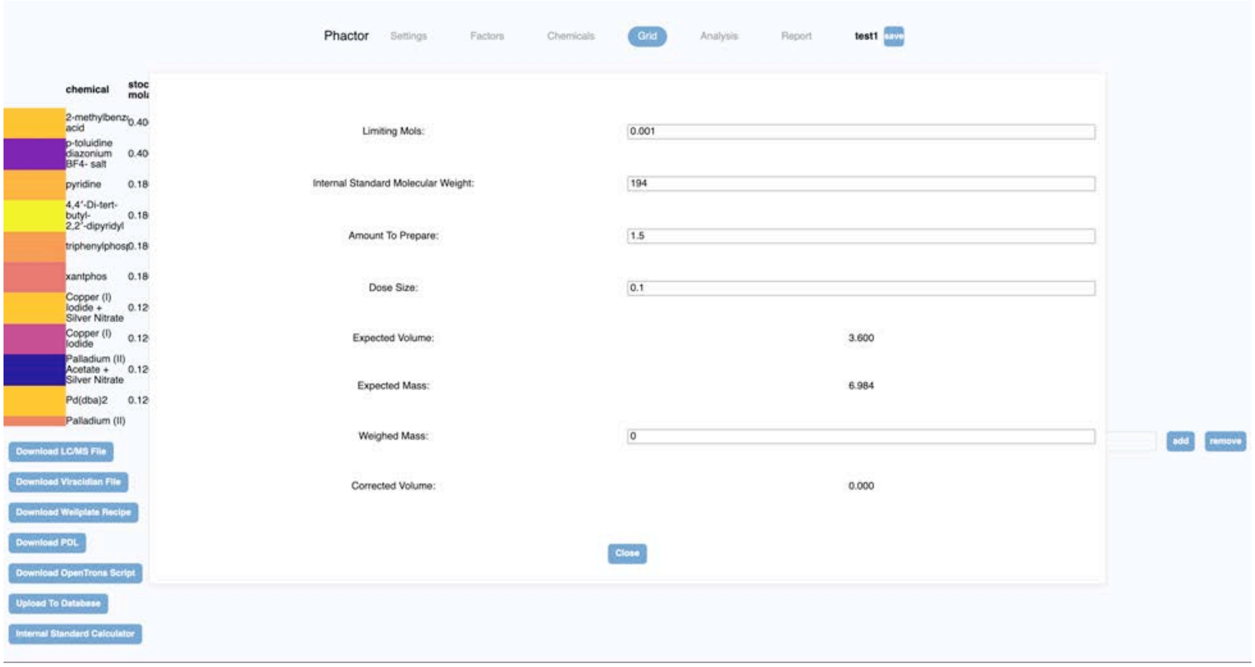
	chemical	stock molarity	stock mass (mg)	stock volume (mL)	transfer volume (mL)	solvent	weighed mass	required volume
🟨	2-methylbenz acid	0.400	39.211	0.681	0.025	MeCN	0	0
🟪	p-toluidine diazonium BF4- salt	0.400	59.314	0.661	0.025	MeCN	0	0
🟧	pyridine	0.180	2.563	0.177	0.025	MeCN	0	0
🟨	4,4'-Di-tert-butyl-2,2'-dipyridyl	0.180	8.696	0.171	0.025	MeCN	0	0
🟧	triphenylphos	0.180	8.498	0.172	0.025	MeCN	0	0
🟥	xantphos	0.180	18.748	0.161	0.025	MeCN	0	0
🟨	Copper (I) Iodide + Silver Nitrate	0.120	5.189	0.115	0.025	MeCN	0	0
🟪	Copper (I) Iodide	0.120	2.742	0.117	0.025	MeCN	0	0
🟦	Palladium (II) Acetate + Silver Nitrate	0.120	5.679	0.114	0.025	MeCN	0	0
🟨	Pd(dba)2	0.120	8.280	0.112	0.025	MeCN	0	0
🟥	Palladium (II)							

124-Fig. 8 | The grid stage of phactor displaying a 24-well reaction array.



	chemical	stock molarity	stock mass (mg)	stock volume (mL)	transfer volume (mL)	solvent	weighed mass	required volume
🟨	4-aminophenc	0.500	17.460	0.303	0.020	none	0	0
🟦	3-AMINO-5,6-DIMETHYL-1,2,4-TRIAZINE, 97%	0.500	19.864	0.300	0.020	none	0	0
🟧	ETHYL 2-AMINOBEN 99+%	0.500	26.431	0.294	0.020	none	0	0
🟨	ACYCLOGUA 98%	0.500	36.033	0.284	0.020	none	0	0
🟪	Cytosine	0.500	17.777	0.302	0.020	none	0	0
🟧	2-AMINO-6-HYDROXY-3-METHYLPU	0.500	26.425	0.294	0.020	none	0	0
🟪	3-Aminopyridi	0.500	15.059	0.305	0.020	none	0	0
🟪	p-toluidine	0.500	17.145	0.303	0.020	none	0	0
🟦	4-Aminobenzo	0.500	18.902	0.301	0.020	none	0	0

125-Fig. 9 | The grid stage of phactor displaying a 1536-well reaction array with 1,280 reactions.



chemical	stoc mol
2-methylbenz acid	0.40
p-toluidine diazonium BF4- salt	0.40
pyridine	0.18
4,4'-Di-tert-butyl-2,2'-dipyridyl	0.18
triphenylphos	0.18
xantphos	0.18
Copper (I) Iodide + Silver Nitrate	0.12
Copper (I) Iodide	0.12
Palladium (II) Acetate + Silver Nitrate	0.12
Pd(dba)2	0.12
Palladium (II)	

Limiting Mols: 0.001

Internal Standard Molecular Weight: 194

Amount To Prepare: 1.5

Dose Size: 0.1

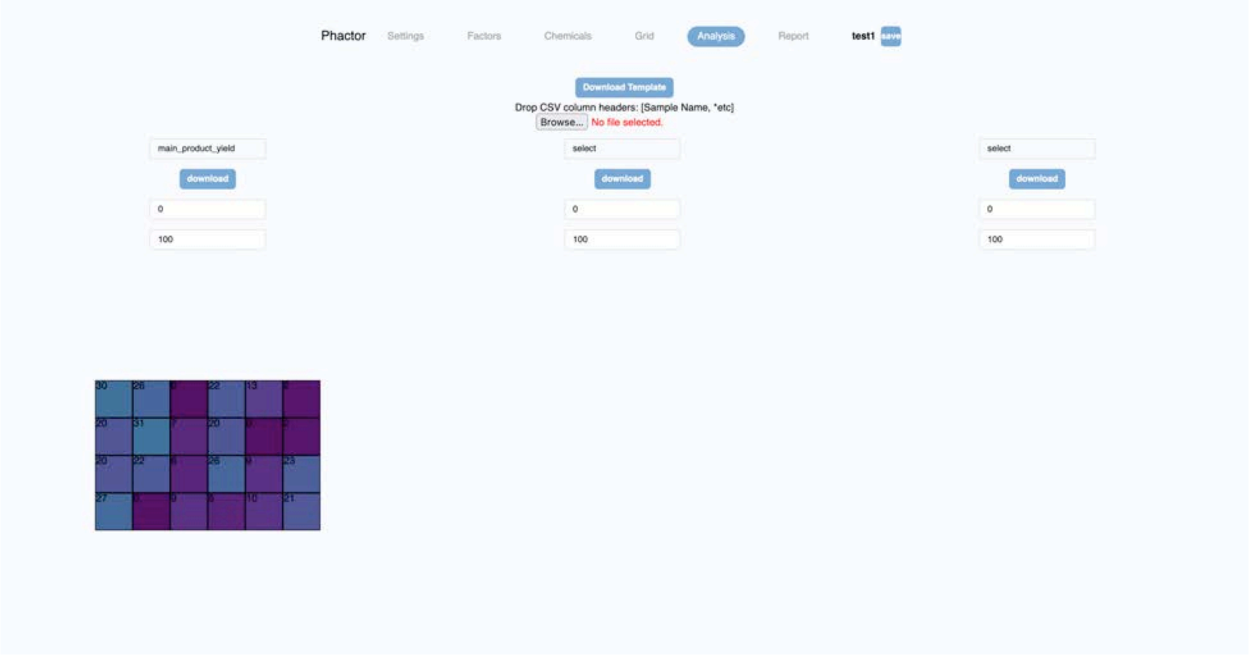
Expected Volume: 3.600

Expected Mass: 6.984

Weighed Mass: 0

Corrected Volume: 0.000

126-Fig. 10 | The connected internal standard calculator pop-up page of phactor on the grid stage.



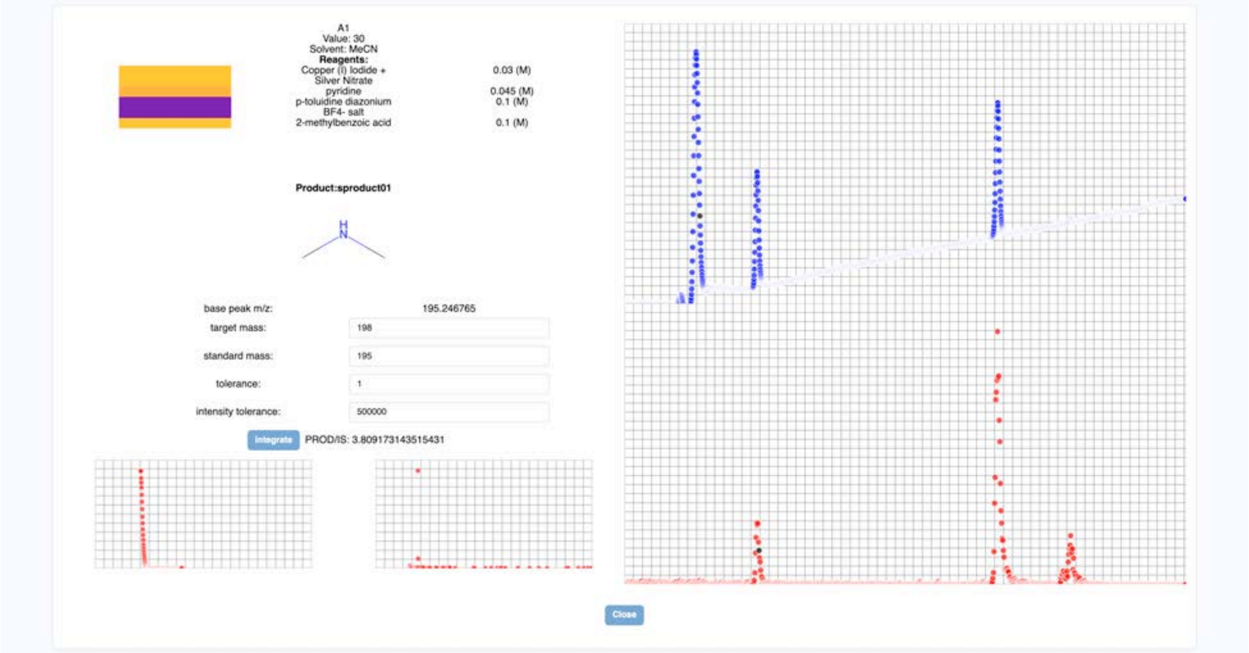
Drop CSV column headers: [Sample Name, *etc]

No file selected.

main_product_yield	select	select
0	0	0
100	100	100

30	26	[illegible]	22	13	[illegible]
20	31	7	20	[illegible]	[illegible]
20	22	[illegible]	26	9	23
27	[illegible]	9	[illegible]	10	21

127-Fig. 11 | The analysis stage of phactor displaying the results of a 24-well reaction array.



A1
Value: 30
Solvent: MeCN
Reagents:

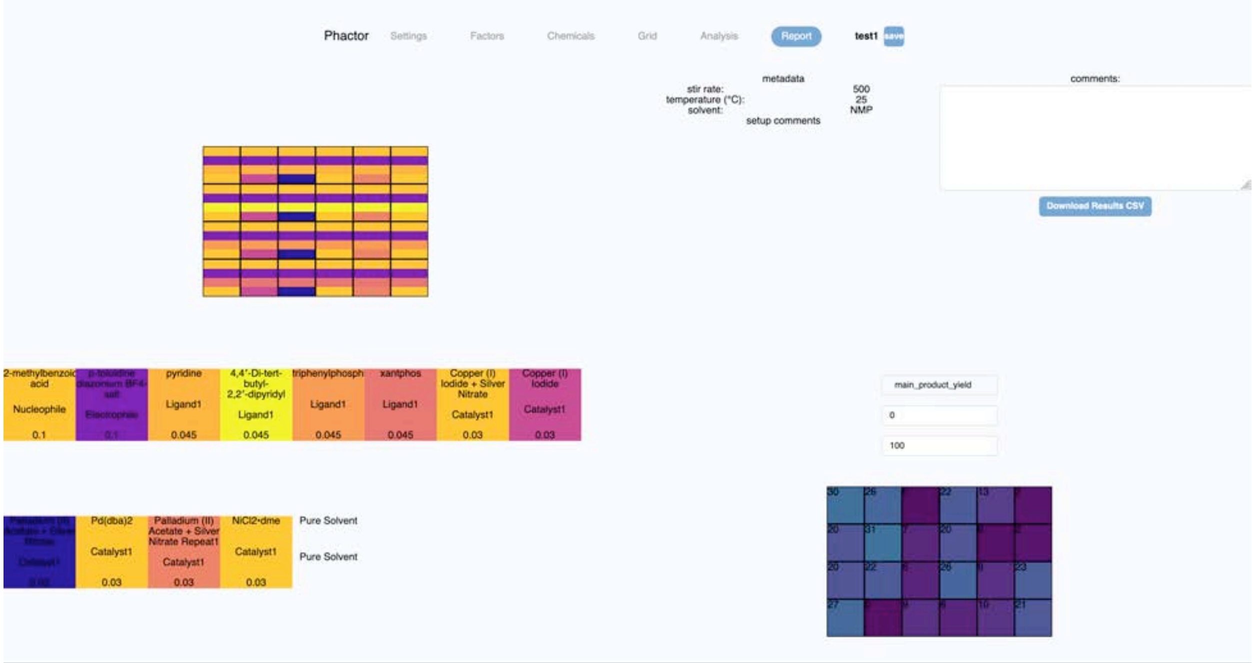
Copper (I) Iodide + Silver Nitrate	0.03 (M)
pyridine	0.045 (M)
p-toluidine diazonium BF4- salt	0.1 (M)
2-methylbenzoic acid	0.1 (M)

Product:sproduct01

base peak m/z:	195.246765
target mass:	198
standard mass:	195
tolerance:	1
intensity tolerance:	500000

PROD/IS: 3.809173143515431

128-Fig. 12 | The UPLC-MS analytics popup on the analysis stage of phactor displaying UPLC-MS trace of a completed reaction.



metadata

stir rate:	500
temperature (°C):	25
solvent:	NMP

setup comments

comments:

2-methylbenzoic acid	Nucleophile	0.1
p-toluidine diazonium BF4- salt	Electrophile	0.1
pyridine	Ligand1	0.045
4,4′-Di-tert-butyl-2,2′-dipyridyl	Ligand1	0.045
triphenylphosph	Ligand1	0.045
xantphos	Ligand1	0.045
Copper (I) Iodide + Silver Nitrate	Catalyst1	0.03
Copper (I) Iodide	Catalyst1	0.03
Palladium (II) Acetate + Silver Nitrate	Catalyst1	0.03
Pd(dba)2	Catalyst1	0.03
Palladium (II) Acetate + Silver Nitrate Repeat1	Catalyst1	0.03
NiCl2•dme	Catalyst1	0.03

Pure Solvent

Pure Solvent

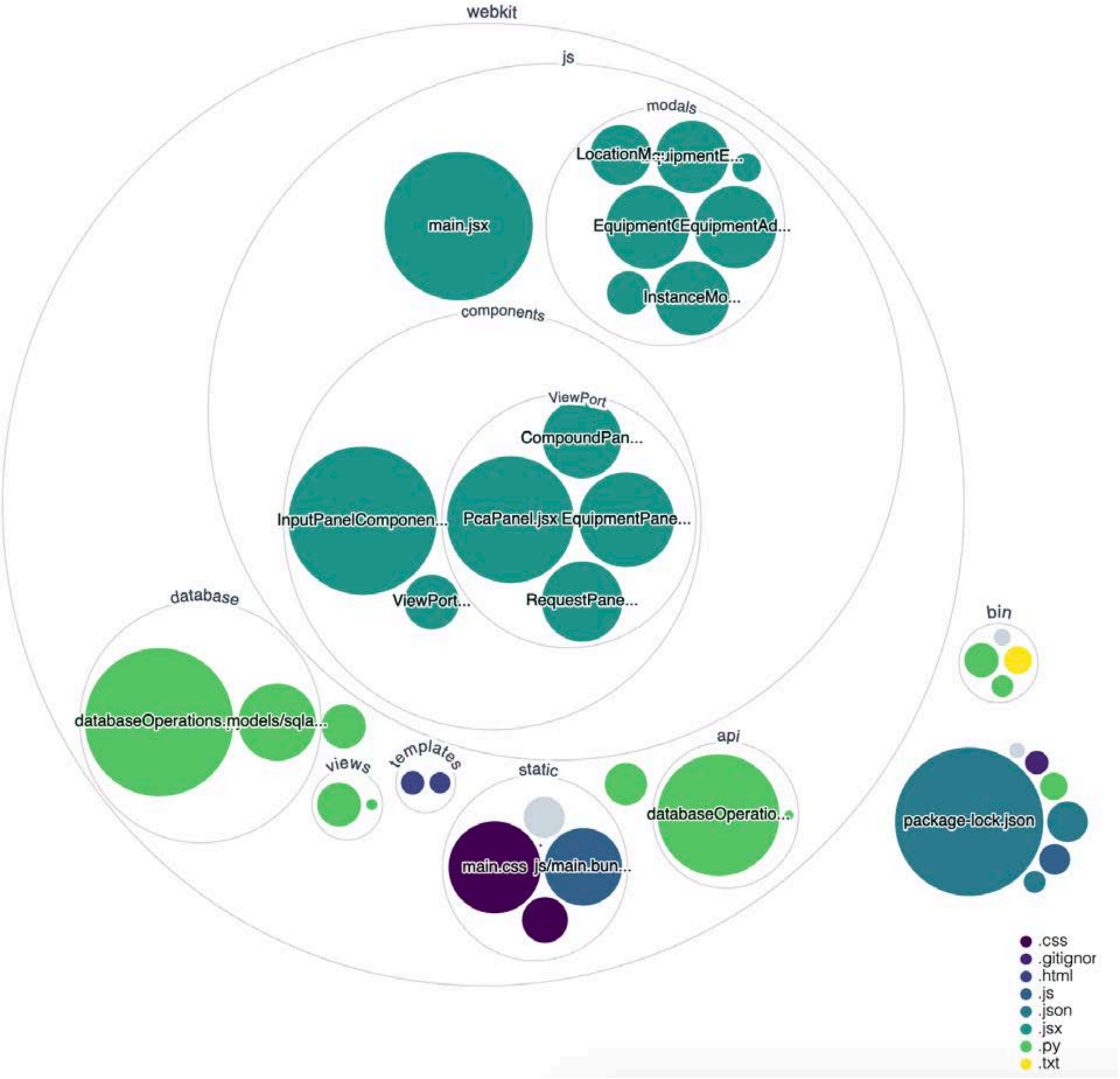
main_product_yield

0

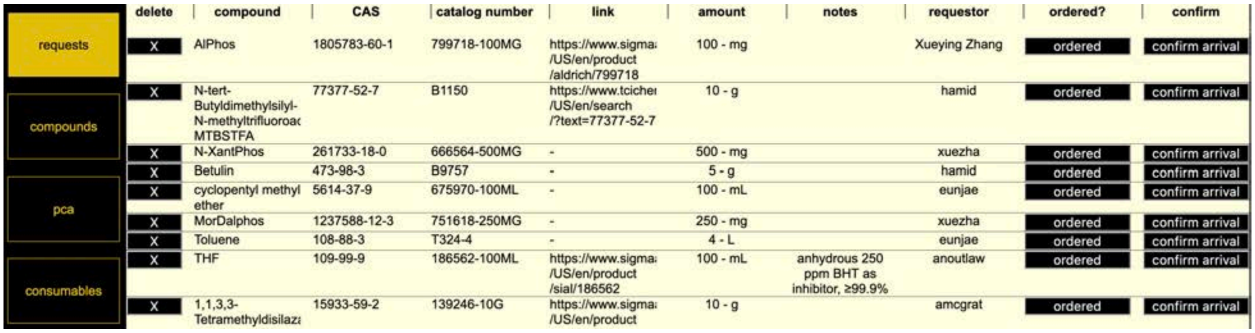
100

129-Fig. 13 | The report stage of phactor displaying the report of a completed 24-well reaction array.

phactor inventory

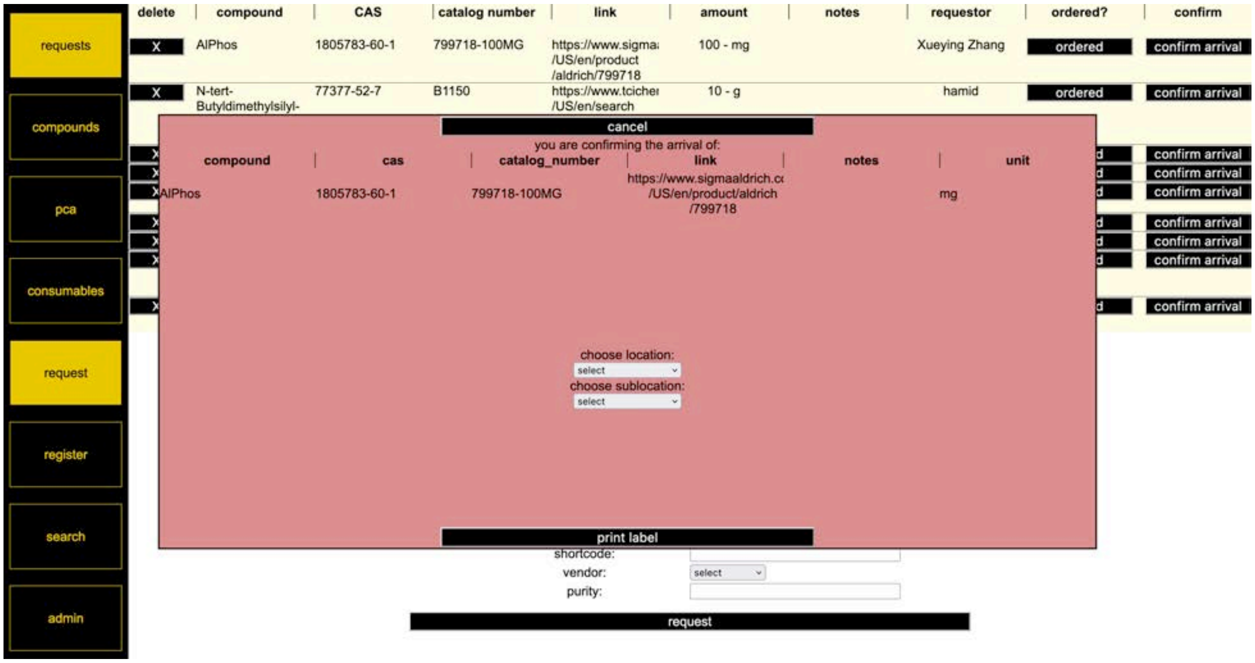


130-Fig. 1 | Visualization of the phactor inventory codebase.



delete	compound	CAS	catalog number	link	amount	notes	requestor	ordered?	confirm
X	AlPhos	1805783-60-1	799718-100MG	https://www.sigma: /US/en/product /aldrich/799718	100 - mg		Xueying Zhang	ordered	confirm arrival
X	N-tert-Butyldimethylsilyl-N-methyltrifluoroac MTBSTFA	77377-52-7	B1150	https://www.tciche /US/en/search /?text=77377-52-7	10 - g		hamid	ordered	confirm arrival
X	N-XantPhos	261733-18-0	666564-500MG	-	500 - mg		xuezha	ordered	confirm arrival
X	Betulin	473-98-3	B9757	-	5 - g		hamid	ordered	confirm arrival
X	cyclopentyl methyl ether	5614-37-9	675970-100ML	-	100 - mL		eunjae	ordered	confirm arrival
X	MorDalphos	1237588-12-3	751618-250MG	-	250 - mg		xuezha	ordered	confirm arrival
X	Toluene	108-88-3	T324-4	-	4 - L		eunjae	ordered	confirm arrival
X	THF	109-99-9	186562-100ML	https://www.sigma: /US/en/product /sial/186562	100 - mL	anhydrous 250 ppm BHT as inhibitor, ≥99.9%	anoutlaw	ordered	confirm arrival
X	1,1,3,3-Tetramethyldisilaz:	15933-59-2	139246-10G	https://www.sigma: /US/en/product	10 - g		amcgrat	ordered	confirm arrival

131-Fig. 2 | Request table of the phactor inventory.



you are confirming the arrival of:

compound	cas	catalog_number	link	notes	unit
AlPhos	1805783-60-1	799718-100MG	https://www.sigmaaldrich.c /US/en/product/aldrich /799718		mg

choose location: select

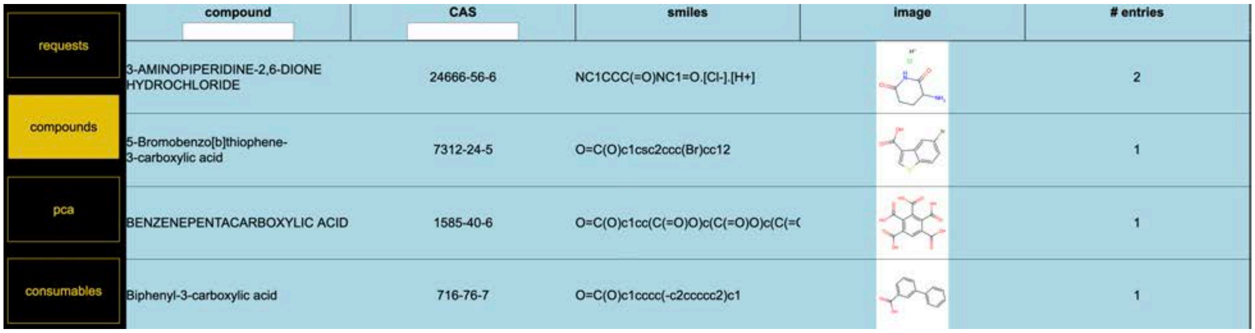
choose sublocation: select

shortcode:

vendor: select

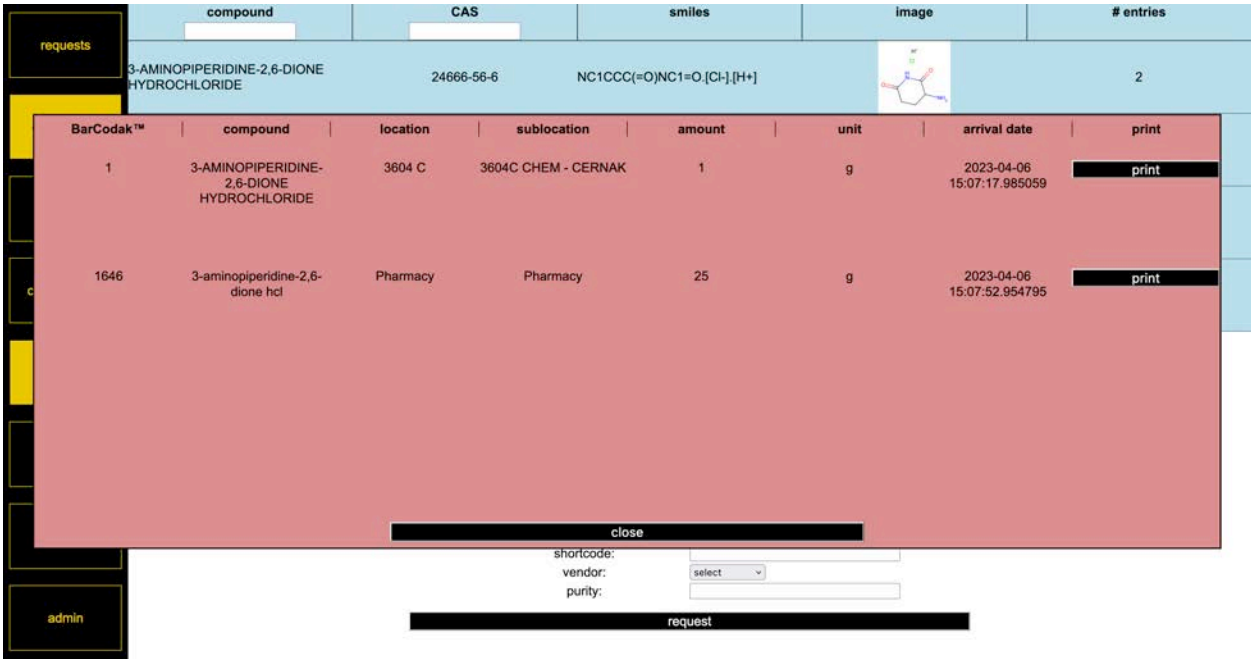
purity:

132-Fig. 3 | Confirm arrival pop-up of the request page.



compound	CAS	smiles	image	# entries
3-AMINOPIPERIDINE-2,6-DIONE HYDROCHLORIDE	24666-56-6	NC1CCC(=O)NC1=O.[Cl-].[H+]		2
5-Bromobenzo[b]thiophene-3-carboxylic acid	7312-24-5	O=C(O)c1csc2ccc(Br)cc12		1
BENZENEPENTACARBOXYLIC ACID	1585-40-6	O=C(O)c1cc(C(=O)O)c(C(=O)O)c(C(=(		1
Biphenyl-3-carboxylic acid	716-76-7	O=C(O)c1cccc(-c2ccccc2)c1		1

133-Fig. 4 | Compounds table of the phactor inventory.

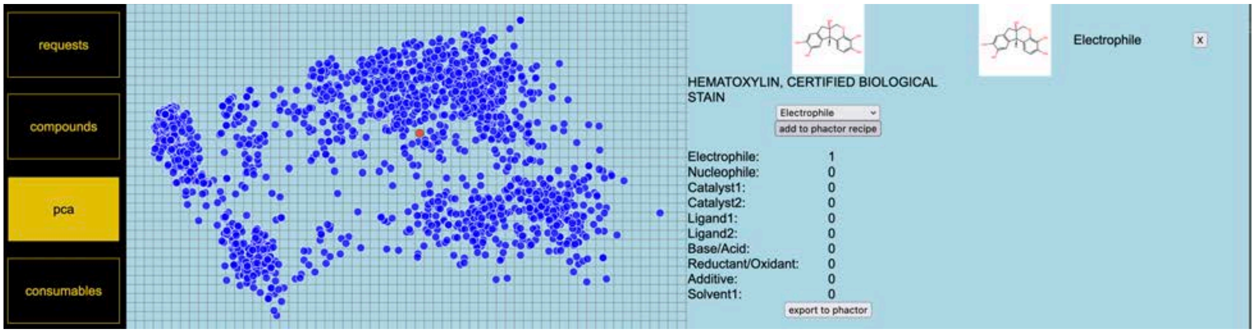


compound	CAS	smiles	image	# entries
3-AMINOPIPERIDINE-2,6-DIONE HYDROCHLORIDE	24666-56-6	NC1CCC(=O)NC1=O.[Cl-].[H+]		2

BarCodak™	compound	location	sublocation	amount	unit	arrival date	print
1	3-AMINOPIPERIDINE-2,6-DIONE HYDROCHLORIDE	3604 C	3604C CHEM - CERNAK	1	g	2023-04-06 15:07:17.985059	print
1646	3-aminopiperidine-2,6-dione hcl	Pharmacy	Pharmacy	25	g	2023-04-06 15:07:52.954795	print

shortcode:
vendor: select
purity:

134-Fig. 5 | BarCodak instance pop-up table of the phactor inventory.

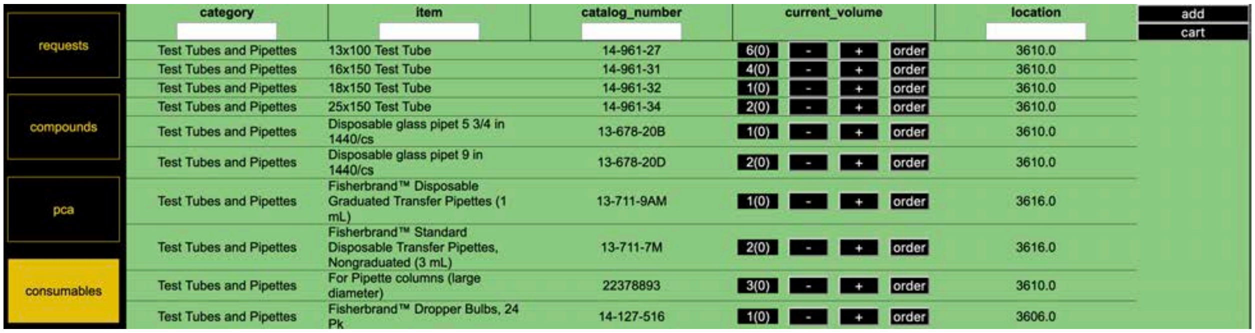


HEMATOXYLIN, CERTIFIED BIOLOGICAL STAIN

Electrophile

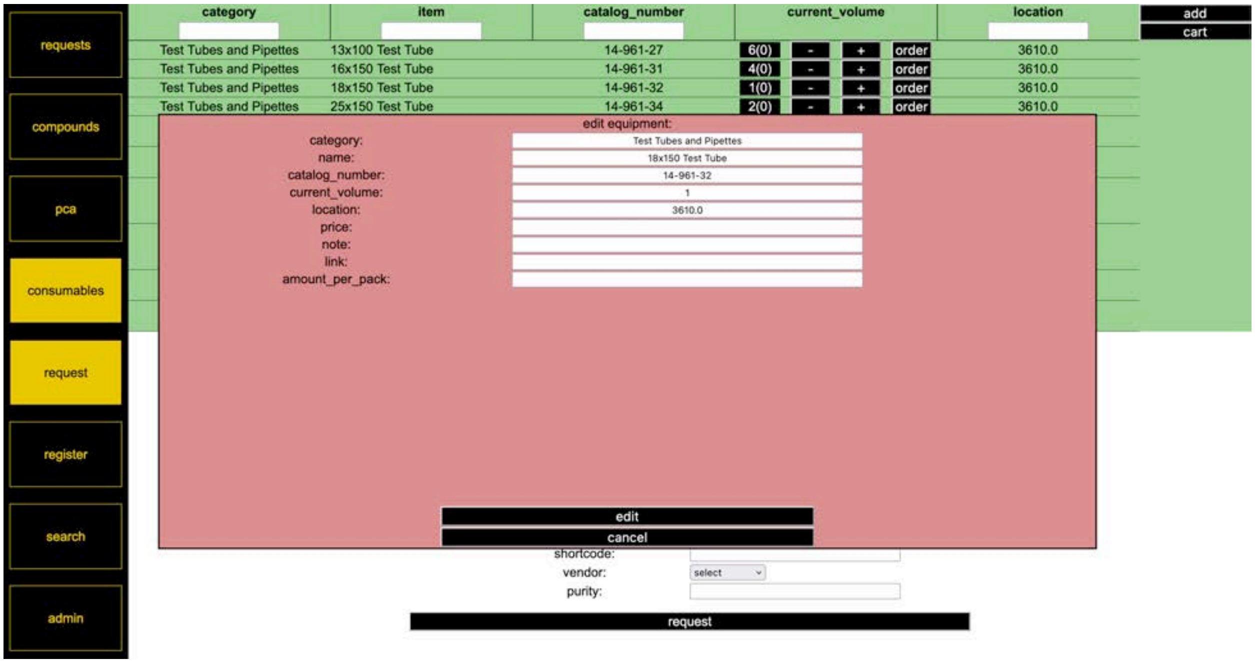
Electrophile:	1
Nucleophile:	0
Catalyst1:	0
Catalyst2:	0
Ligand1:	0
Ligand2:	0
Base/Acid:	0
Reductant/Oxidant:	0
Additive:	0
Solvent1:	0

135-Fig. 6 | PCA embedding page of the phactor inventory.



category	item	catalog_number	current_volume	location
Test Tubes and Pipettes	13x100 Test Tube	14-961-27	6(0) - + order	3610.0
Test Tubes and Pipettes	16x150 Test Tube	14-961-31	4(0) - + order	3610.0
Test Tubes and Pipettes	18x150 Test Tube	14-961-32	1(0) - + order	3610.0
Test Tubes and Pipettes	25x150 Test Tube	14-961-34	2(0) - + order	3610.0
Test Tubes and Pipettes	Disposable glass pipet 5 3/4 in 1440/cs	13-678-20B	1(0) - + order	3610.0
Test Tubes and Pipettes	Disposable glass pipet 9 in 1440/cs	13-678-20D	2(0) - + order	3610.0
Test Tubes and Pipettes	Fisherbrand™ Disposable Graduated Transfer Pipettes (1 mL)	13-711-9AM	1(0) - + order	3616.0
Test Tubes and Pipettes	Fisherbrand™ Standard Disposable Transfer Pipettes, Nongraduated (3 mL)	13-711-7M	2(0) - + order	3616.0
Test Tubes and Pipettes	For Pipette columns (large diameter)	22378893	3(0) - + order	3610.0
Test Tubes and Pipettes	Fisherbrand™ Dropper Bulbs, 24 Pk	14-127-516	1(0) - + order	3606.0

136-Fig. 7 | Consumables table of the phactor inventory.



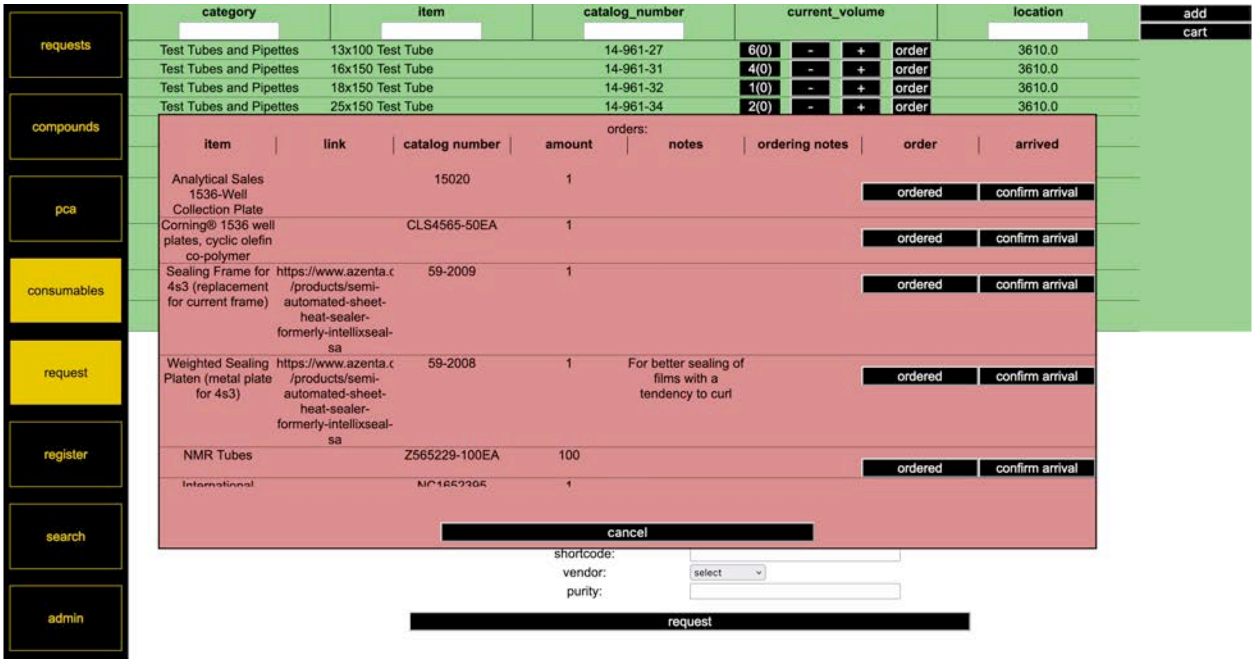
category	item	catalog_number	current_volume	location
Test Tubes and Pipettes	13x100 Test Tube	14-961-27	6(0)	3610.0
Test Tubes and Pipettes	16x150 Test Tube	14-961-31	4(0)	3610.0
Test Tubes and Pipettes	18x150 Test Tube	14-961-32	1(0)	3610.0
Test Tubes and Pipettes	25x150 Test Tube	14-961-34	2(0)	3610.0

edit equipment:

category: Test Tubes and Pipettes
name: 18x150 Test Tube
catalog_number: 14-961-32
current_volume: 1
location: 3610.0
price:
note:
link:
amount_per_pack:

shortcode:
vendor: select
purity:

137-Fig. 8 | Edit consumable form of the phactor inventory.



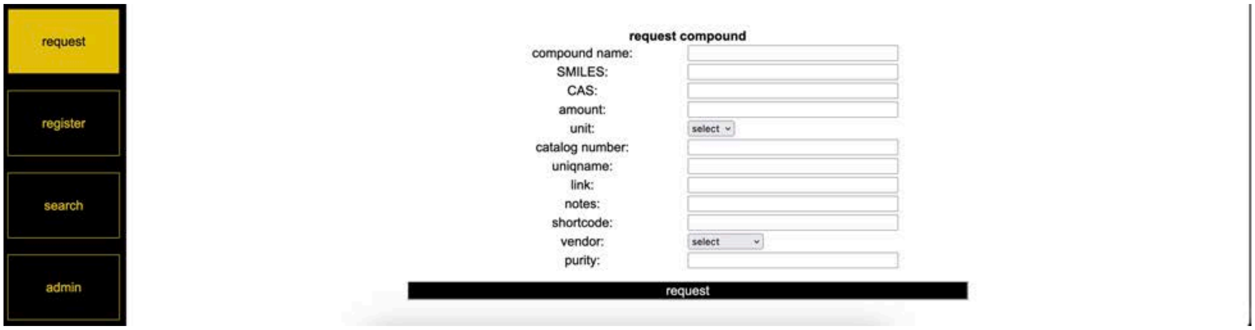
category	item	catalog_number	current_volume	location
Test Tubes and Pipettes	13x100 Test Tube	14-961-27	6(0)	3610.0
Test Tubes and Pipettes	16x150 Test Tube	14-961-31	4(0)	3610.0
Test Tubes and Pipettes	18x150 Test Tube	14-961-32	1(0)	3610.0
Test Tubes and Pipettes	25x150 Test Tube	14-961-34	2(0)	3610.0

orders:

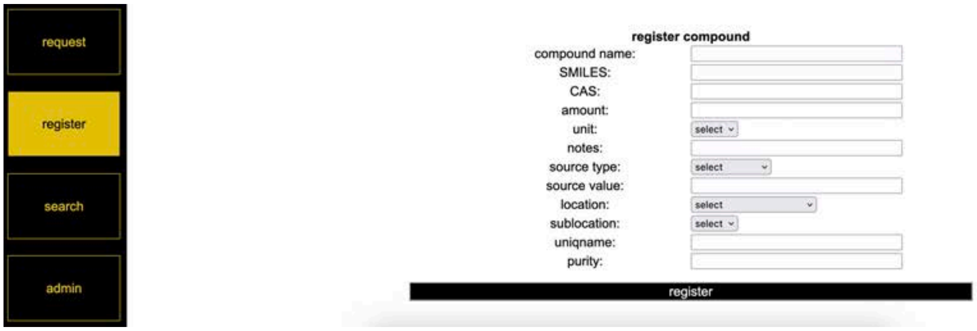
item	link	catalog number	amount	notes	ordering notes	order	arrived
Analytical Sales 1536-Well Collection Plate		15020	1			ordered	confirm arrival
Corning® 1536 well plates, cyclic olefin co-polymer		CLS4565-50EA	1			ordered	confirm arrival
Sealing Frame for 4s3 (replacement for current frame)	https://www.azenta.c /products/semi-automated-sheet-heat-sealer-formerly-intellixseal-sa	59-2009	1			ordered	confirm arrival
Weighted Sealing Platen (metal plate for 4s3)	https://www.azenta.c /products/semi-automated-sheet-heat-sealer-formerly-intellixseal-sa	59-2008	1	For better sealing of films with a tendency to curl		ordered	confirm arrival
NMR Tubes		Z565229-100EA	100			ordered	confirm arrival

shortcode:
vendor: select
purity:

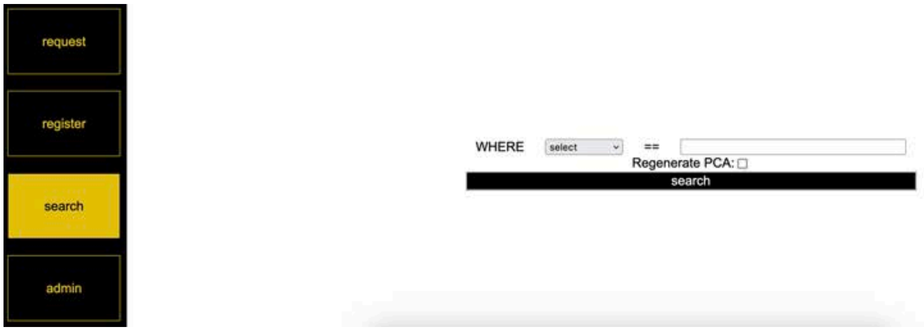
138-Fig. 9 | Consumable shopping cart table of the phactor inventory.



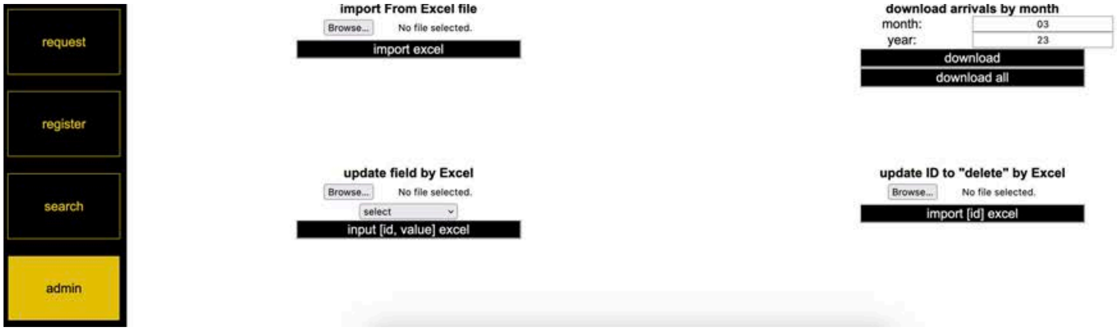
139-Fig. 10 | Request compound form of the phactor inventory.



140-Fig. 11 | Register compound form of the phactor inventory.



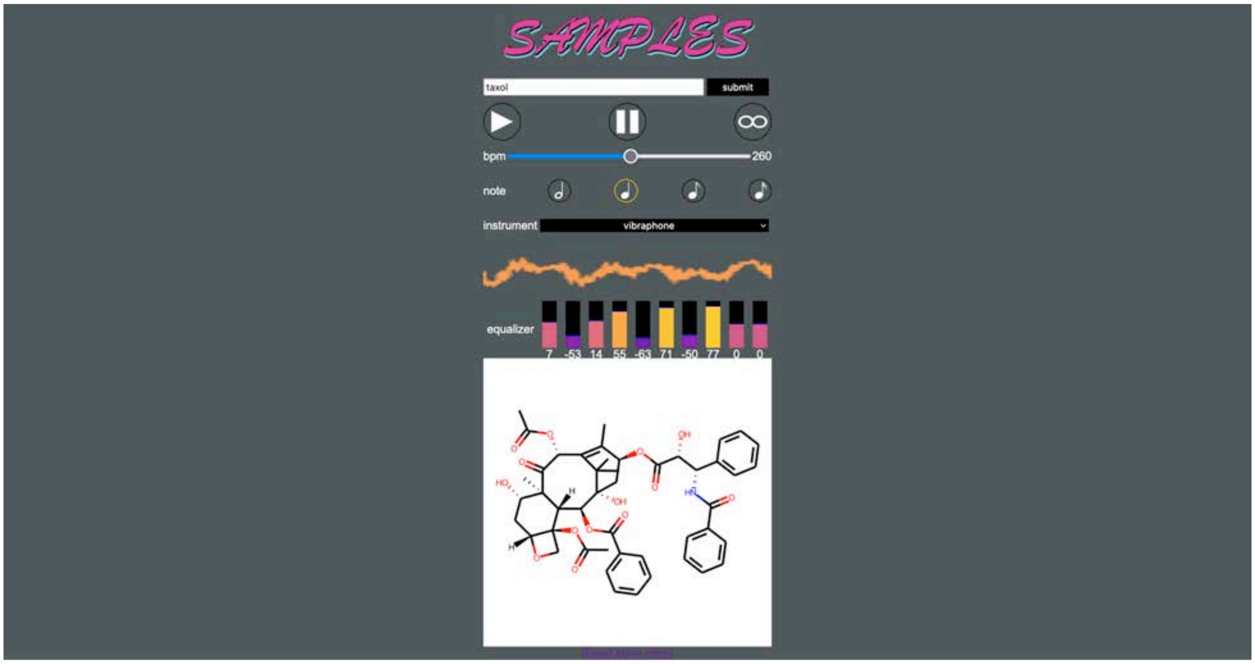
141-Fig. 12 | Search database form of the phactor inventory.



month: 03
year: 23

142-Fig. 13 | Admin panel of the phactor inventory.

samples



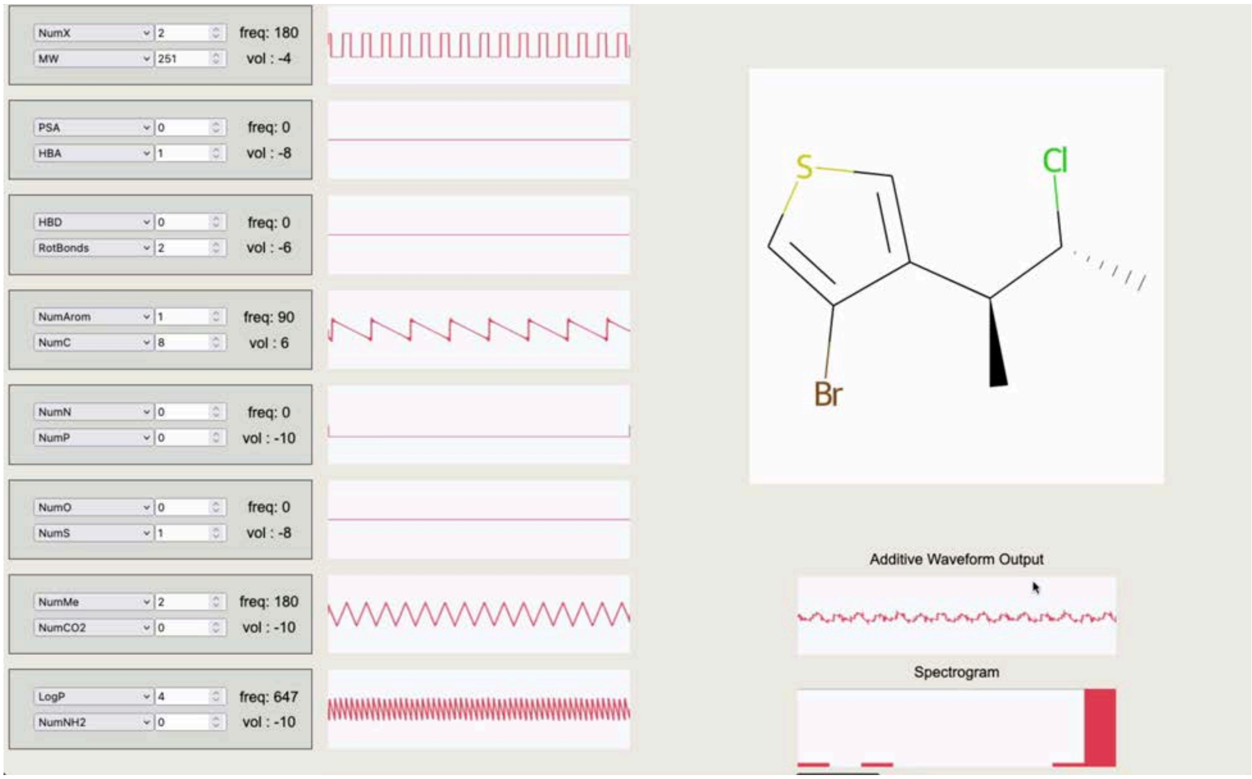
taxol

bpm 260

instrument vibraphone

equalizer 7 -53 14 55 -63 71 -50 77 0 0

143-Fig. 1 | Samples webapp playing taxol.



NumX 2 MW 251 freq: 180 vol : -4

PSA 0 HBA 1 freq: 0 vol : -8

HBD 0 RotBonds 2 freq: 0 vol : -6

NumArom 1 NumC 8 freq: 90 vol : 6

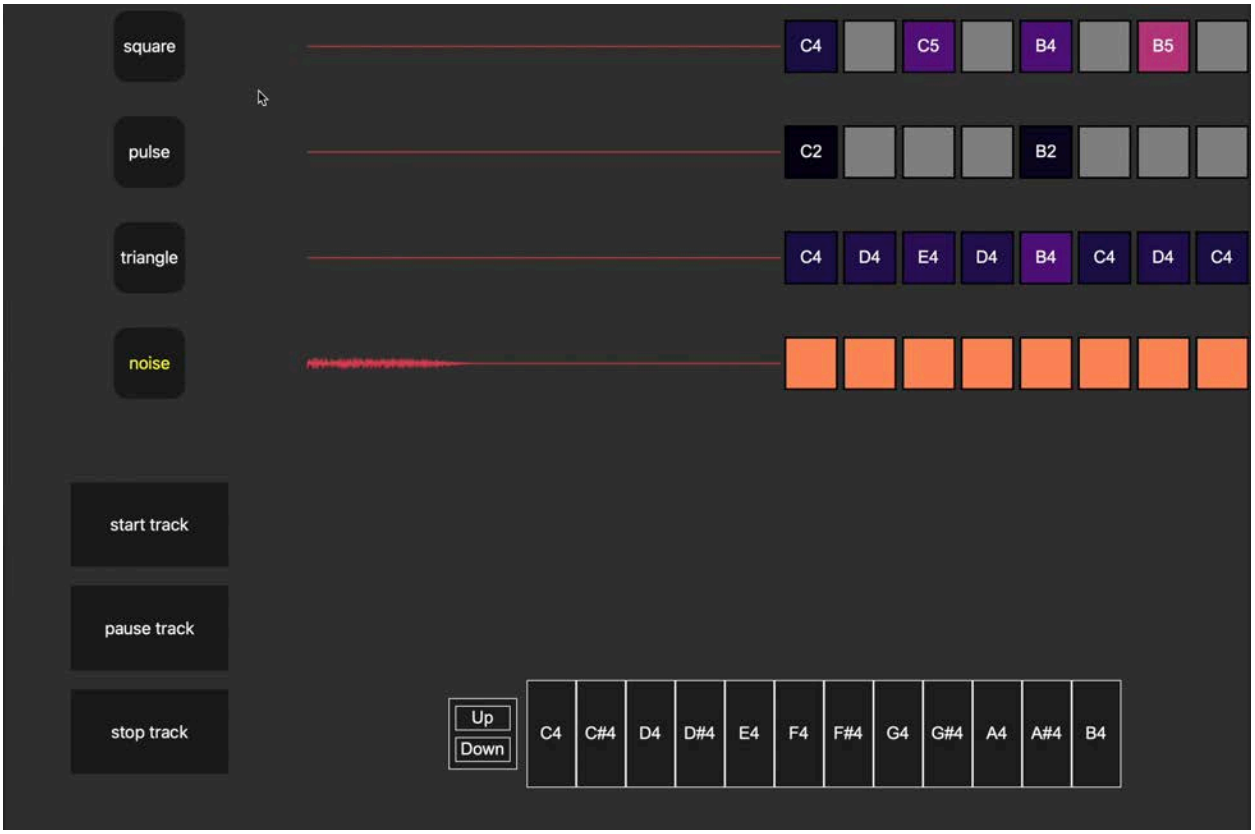
NumN 0 NumP 0 freq: 0 vol : -10

NumO 0 NumS 1 freq: 0 vol : -8

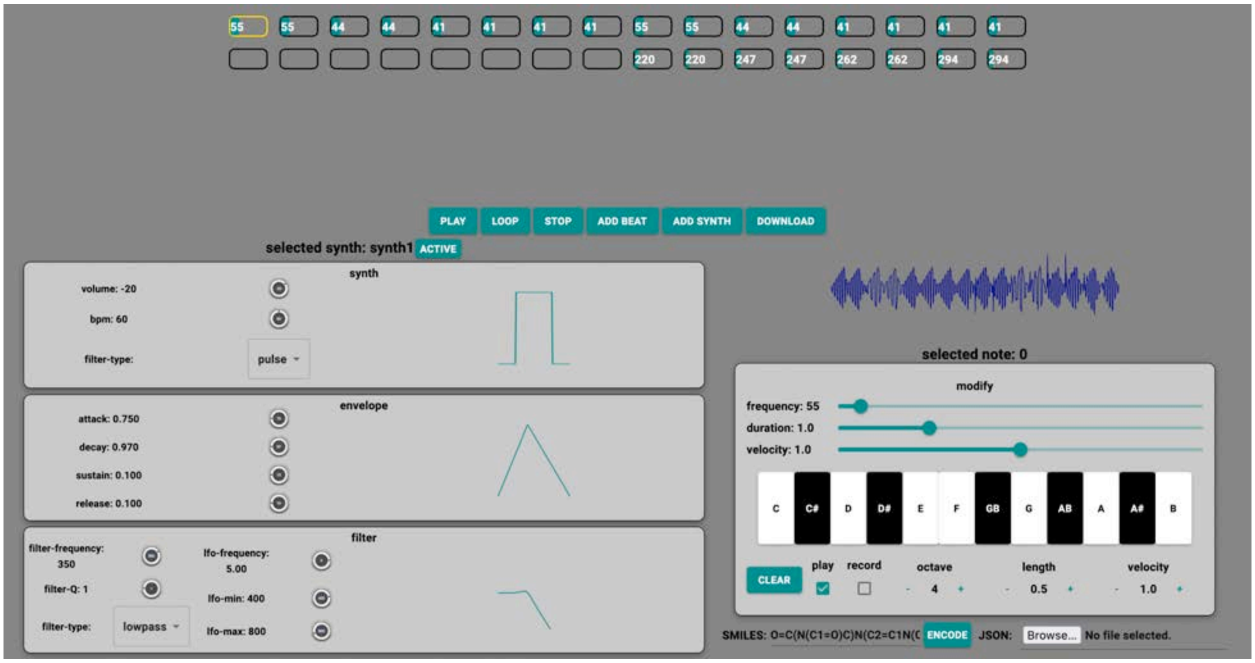
NumMe 2 NumCO2 0 freq: 180 vol : -10

LogP 4 NumNH2 0 freq: 647 vol : -10

144-Fig. 2 | Prototype of the SAMPLES additive synthesizer webapp.



145-Fig. 3 | 8-bit sequencing prototype implementation of samples.



selected synth: synth1

volume: -20

bpm: 60

filter-type: pulse

attack: 0.750

decay: 0.970

sustain: 0.100

release: 0.100

filter-frequency: 350

filter-Q: 1

filter-type: lowpass

lfo-frequency: 5.00

lfo-min: 400

lfo-max: 800

selected note: 0

frequency: 55

duration: 1.0

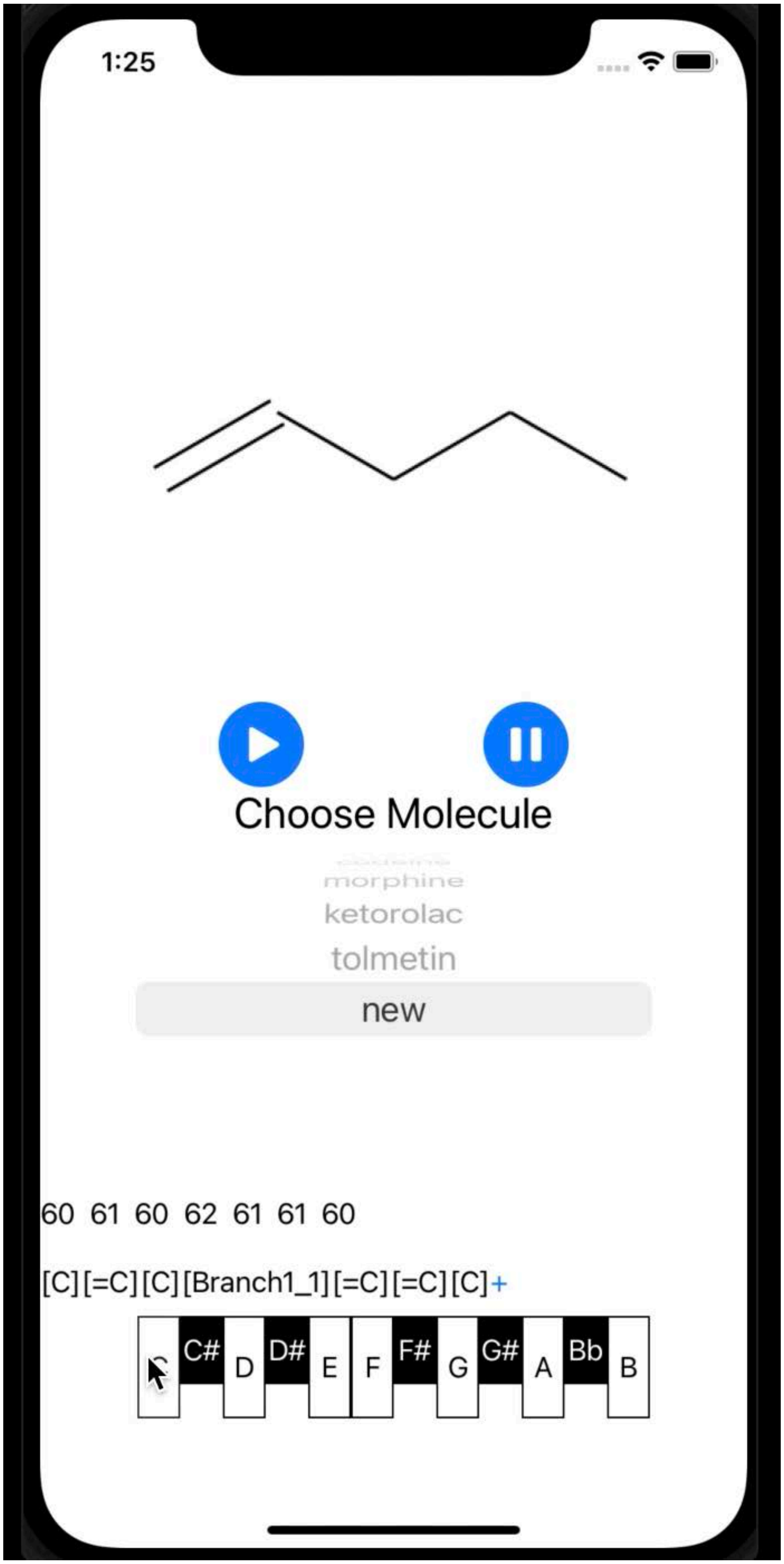
velocity: 1.0

play ☑ record ☐ octave - 4 + length - 0.5 + velocity - 1.0 +

SMILES: O=C(N(C1=O)C)N(C2=C1N(C

JSON: No file selected.

146-Fig. 4 | 8-bit sequencing prototype implementation of samples with filters and low frequency oscillators.

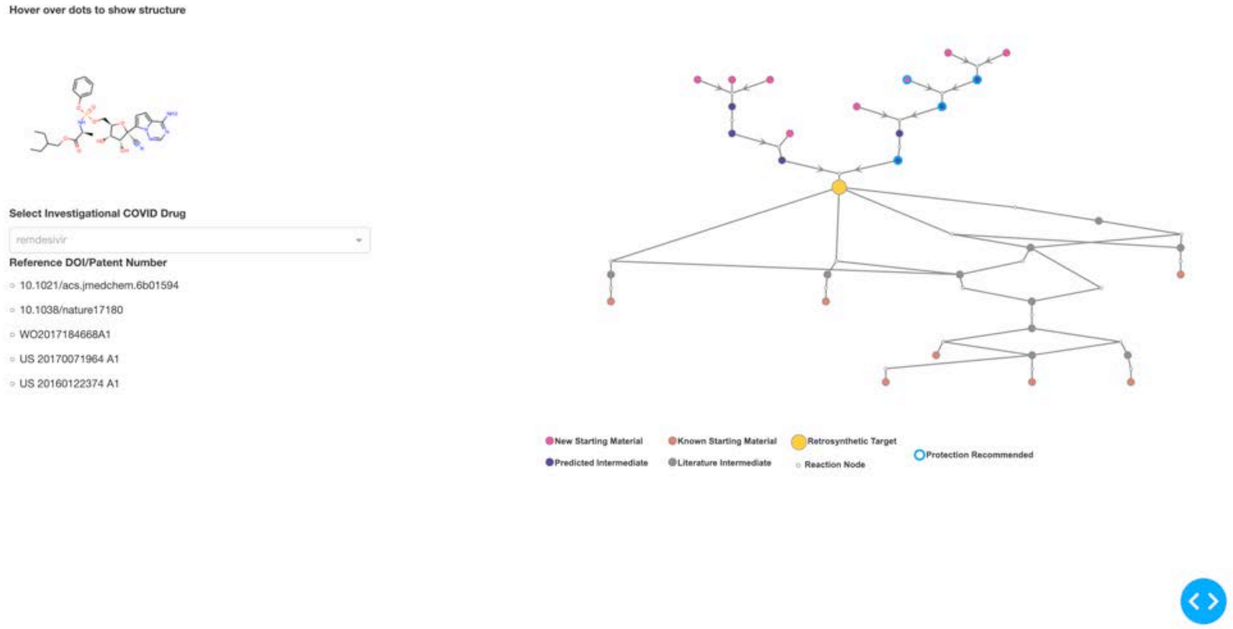


60 61 60 62 61 61 60

[C][=C][C][Branch1_1][=C][=C][C]+

147-Fig 5. | iPhone implementation of SAMPLES algorithm

retrosynthetic visualizer



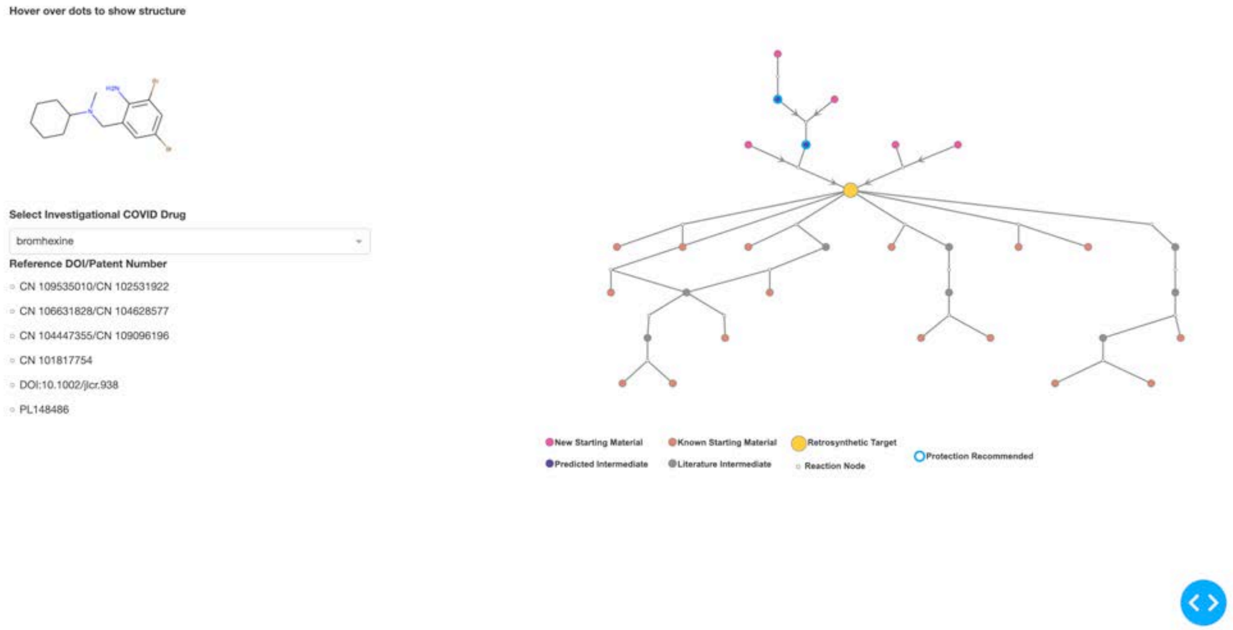
Select Investigational COVID Drug

remdesivir

Reference DOI/Patent Number

- 10.1021/acs.jmedchem.6b01594
- 10.1038/nature17180
- WO2017184668A1
- US 20170071964 A1
- US 20160122374 A1

148-Fig. 1 | Retrosynthetic analysis of remdesivir on the visualizer webapp.



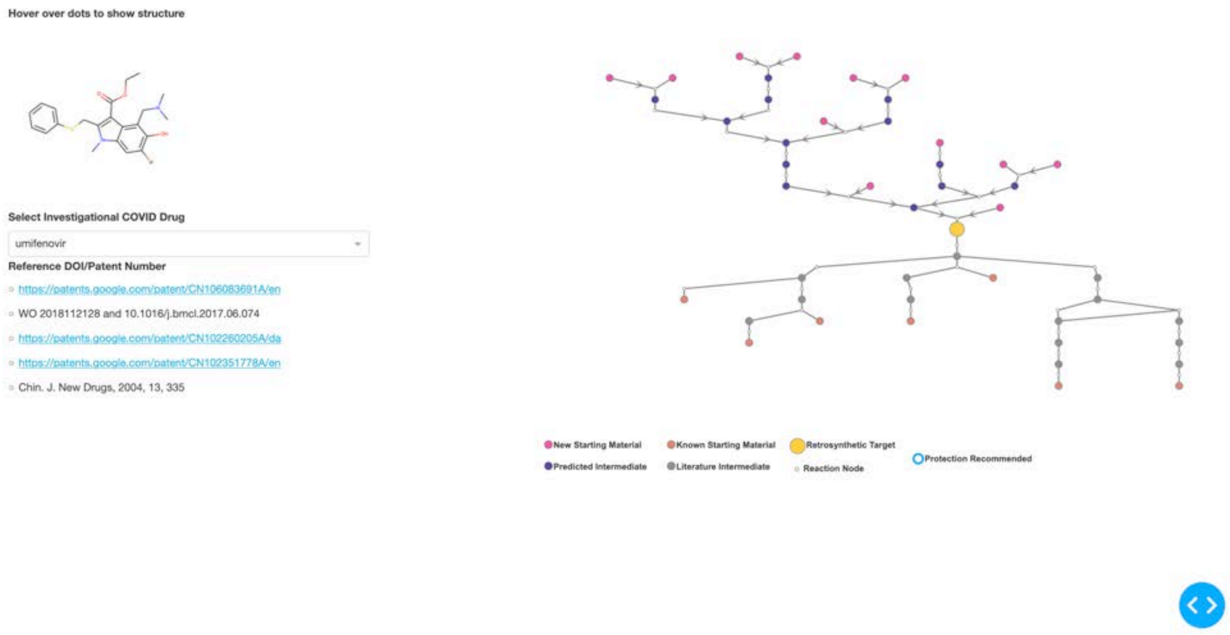
Select Investigational COVID Drug

bromhexine

Reference DOI/Patent Number

- CN 109535010/CN 102531922
- CN 106631828/CN 104628577
- CN 104447355/CN 109096196
- CN 101817754
- DOI:10.1002/jlcr.938
- PL148486

149-Fig. 2 | Retrosynthetic analysis of bromhexine on the visualizer webapp.



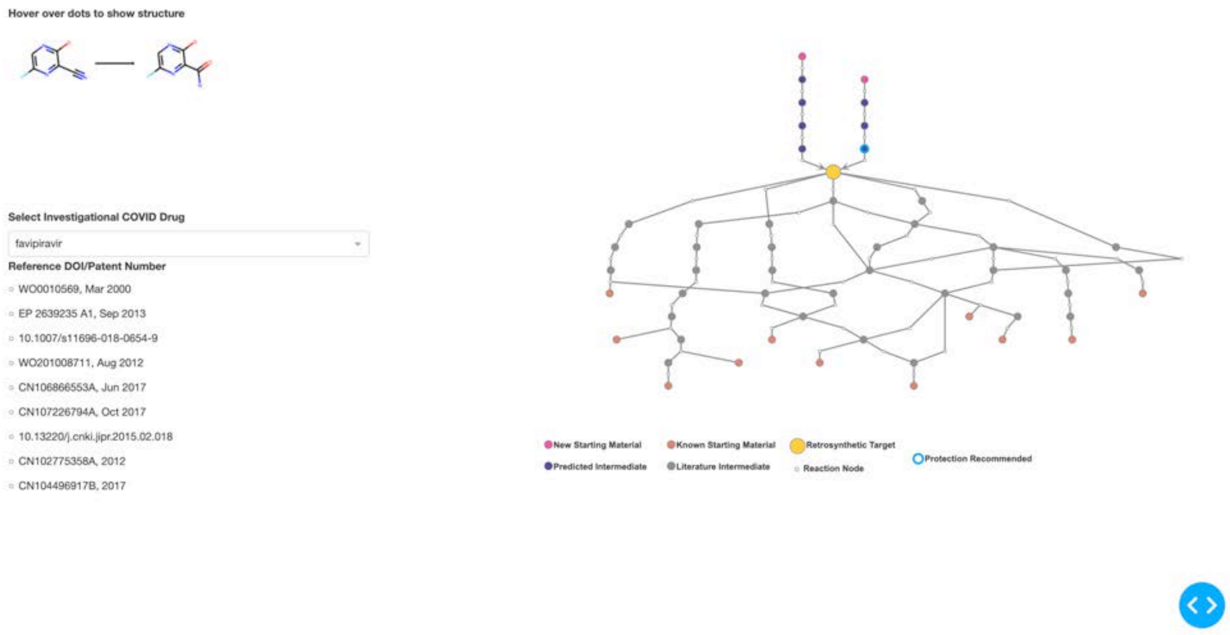
Hover over dots to show structure

Select Investigational COVID Drug: umifenovir

Reference DOI/Patent Number

- https://patents.google.com/patent/CN106083691A/en
- WO 2018112128 and 10.1016/j.bmcl.2017.06.074
- https://patents.google.com/patent/CN102260205A/da
- https://patents.google.com/patent/CN102351778A/en
- Chin. J. New Drugs, 2004, 13, 335

150-Fig. 3 | Retrosynthetic analysis of umifenovir on the visualizer webapp.



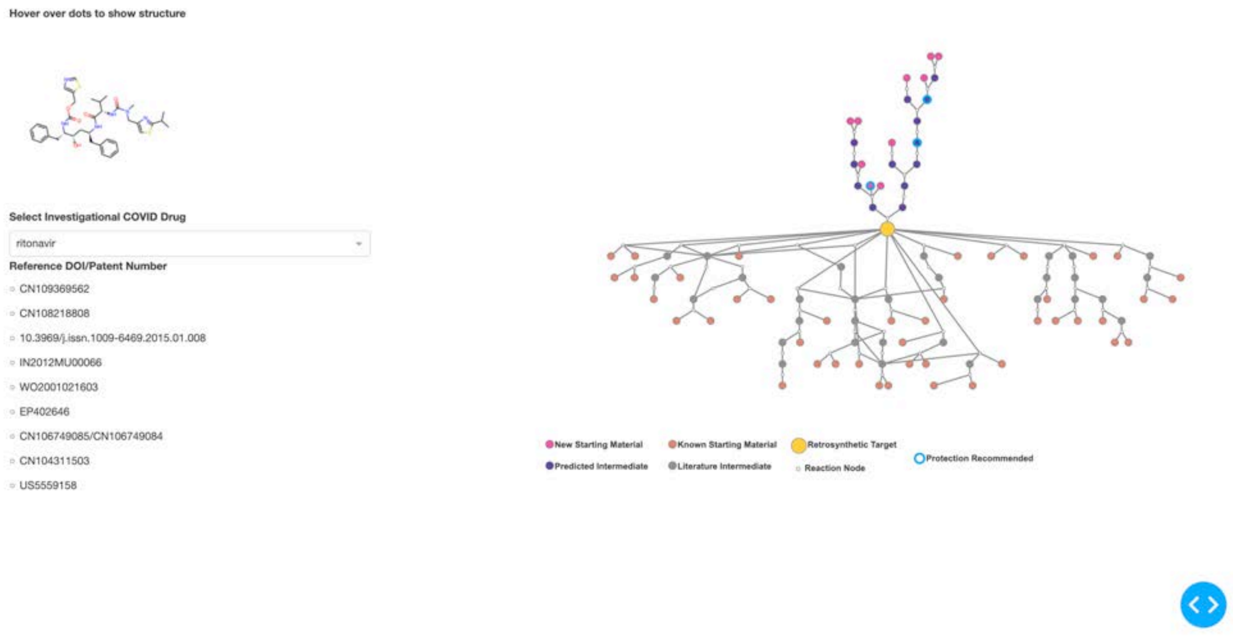
Hover over dots to show structure

Select Investigational COVID Drug: favipiravir

Reference DOI/Patent Number

- WO0010569, Mar 2000
- EP 2639235 A1, Sep 2013
- 10.1007/s11696-018-0654-9
- WO201008711, Aug 2012
- CN106866553A, Jun 2017
- CN107226794A, Oct 2017
- 10.13220/j.cnki.jipr.2015.02.018
- CN102775358A, 2012
- CN104496917B, 2017

151-Fig. 4 | Retrosynthetic analysis of favipiravir on the visualizer webapp.



Hover over dots to show structure

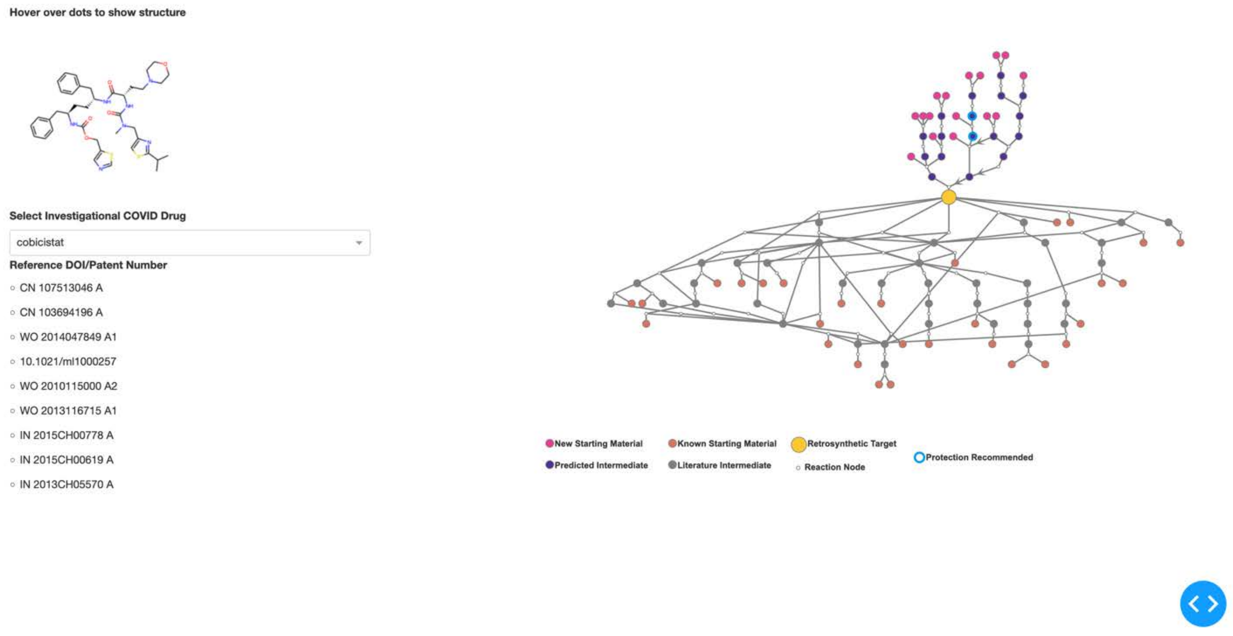
Select Investigational COVID Drug

ritonavir

Reference DOI/Patent Number

- CN109369562
- CN108218808
- 10.3969/j.issn.1009-6469.2015.01.008
- IN2012MU00066
- WO2001021603
- EP402646
- CN106749085/CN106749084
- CN104311503
- US5559158

152-Fig. 5 | Retrosynthetic analysis of ritonavir on the visualizer webapp.



Hover over dots to show structure

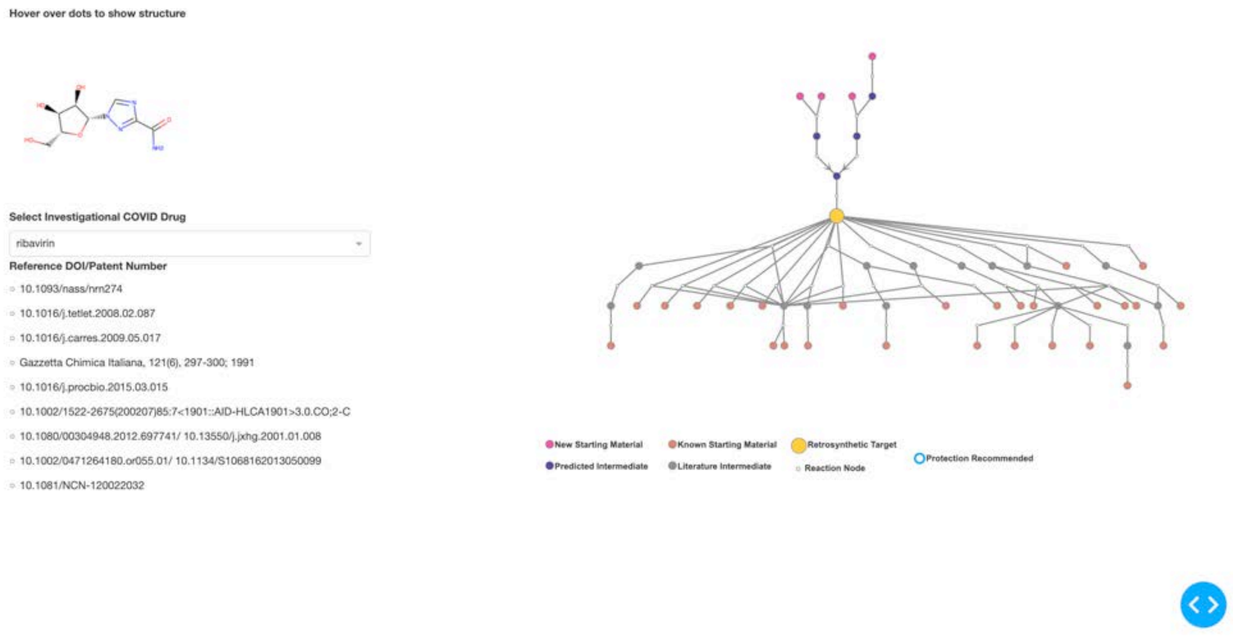
Select Investigational COVID Drug

cobicistat

Reference DOI/Patent Number

- CN 107513046 A
- CN 103694196 A
- WO 2014047849 A1
- 10.1021/ml1000257
- WO 2010115000 A2
- WO 2013116715 A1
- IN 2015CH00778 A
- IN 2015CH00619 A
- IN 2013CH05570 A

153-Fig. 6 | Retrosynthetic analysis of cobicistat on the visualizer webapp.



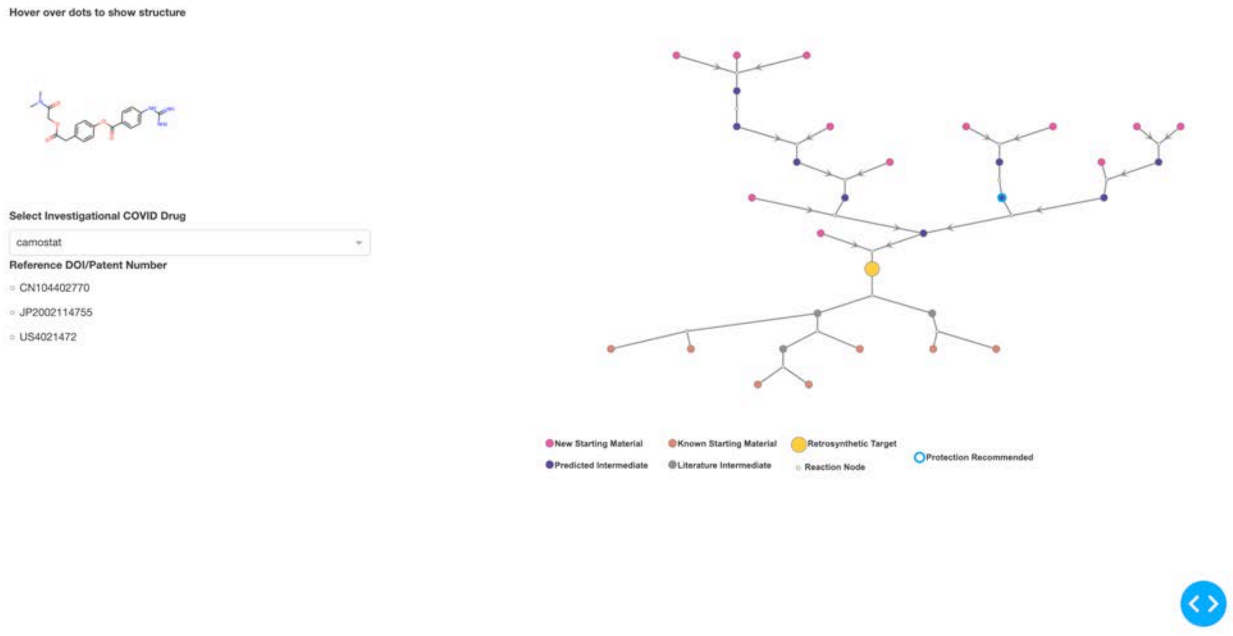
Select Investigational COVID Drug

ribavirin

Reference DOI/Patent Number

- 10.1093/nass/nrn274
- 10.1016/j.tetlet.2008.02.087
- 10.1016/j.carres.2009.05.017
- Gazzetta Chimica Italiana, 121(6), 297-300; 1991
- 10.1016/j.procbio.2015.03.015
- 10.1002/1522-2675(200207)85:7<1901::AID-HLCA1901>3.0.CO;2-C
- 10.1080/00304948.2012.697741/ 10.13550/j.jxhg.2001.01.008
- 10.1002/0471264180.or055.01/ 10.1134/S1068162013050099
- 10.1081/NCN-120022032

154-Fig. 7 | Retrosynthetic analysis of ribavirin on the visualizer webapp.



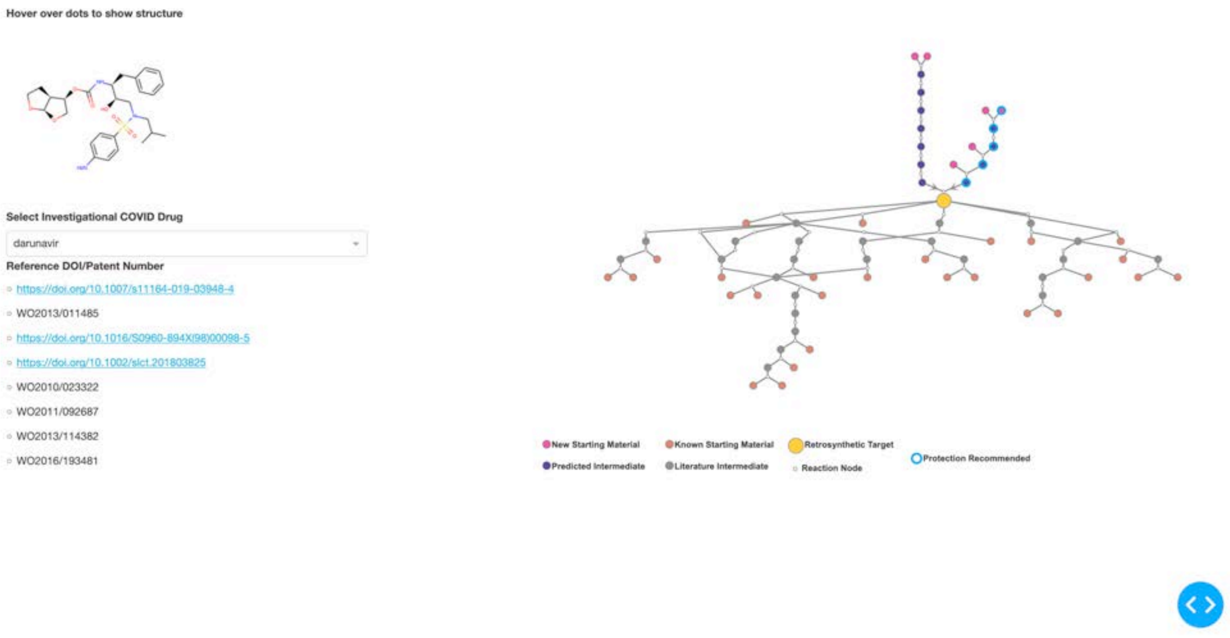
Select Investigational COVID Drug

camostat

Reference DOI/Patent Number

- CN104402770
- JP2002114755
- US4021472

155-Fig. 8 | Retrosynthetic analysis of camostat on the visualizer webapp.



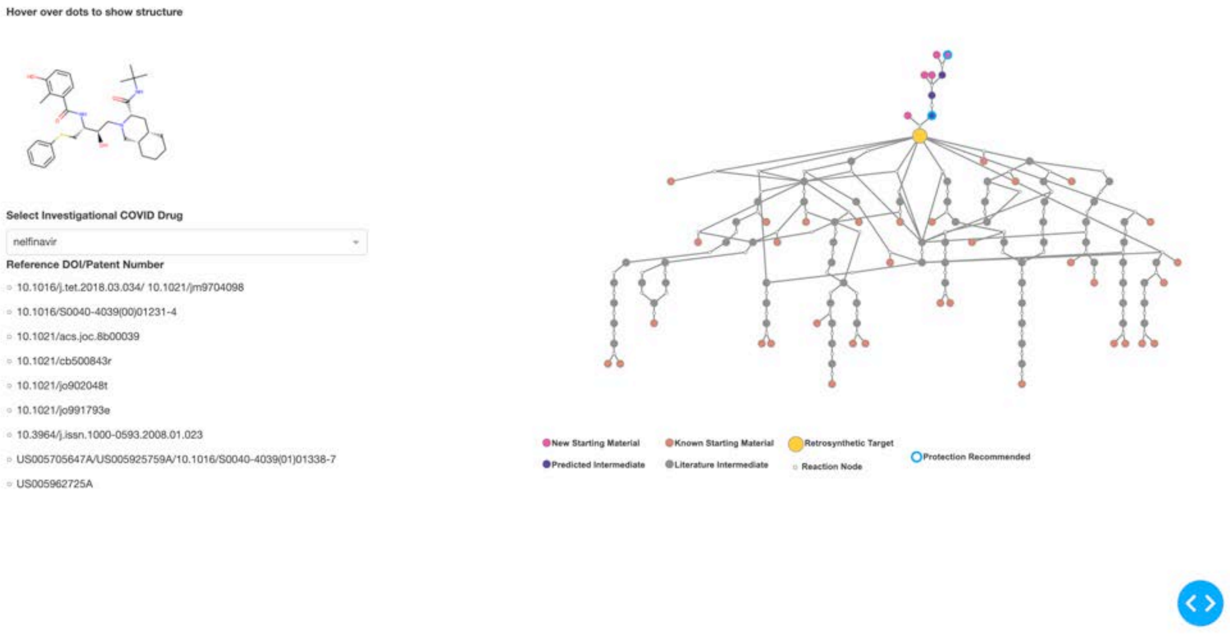
Select Investigational COVID Drug

darunavir

Reference DOI/Patent Number

- https://doi.org/10.1007/s11164-019-03948-4
- WO2013/011485
- https://doi.org/10.1016/S0960-894X(98)00098-5
- https://doi.org/10.1002/slct.201803825
- WO2010/023322
- WO2011/092687
- WO2013/114382
- WO2016/193481

156-Fig. 9 | Retrosynthetic analysis of darunavir on the visualizer webapp.



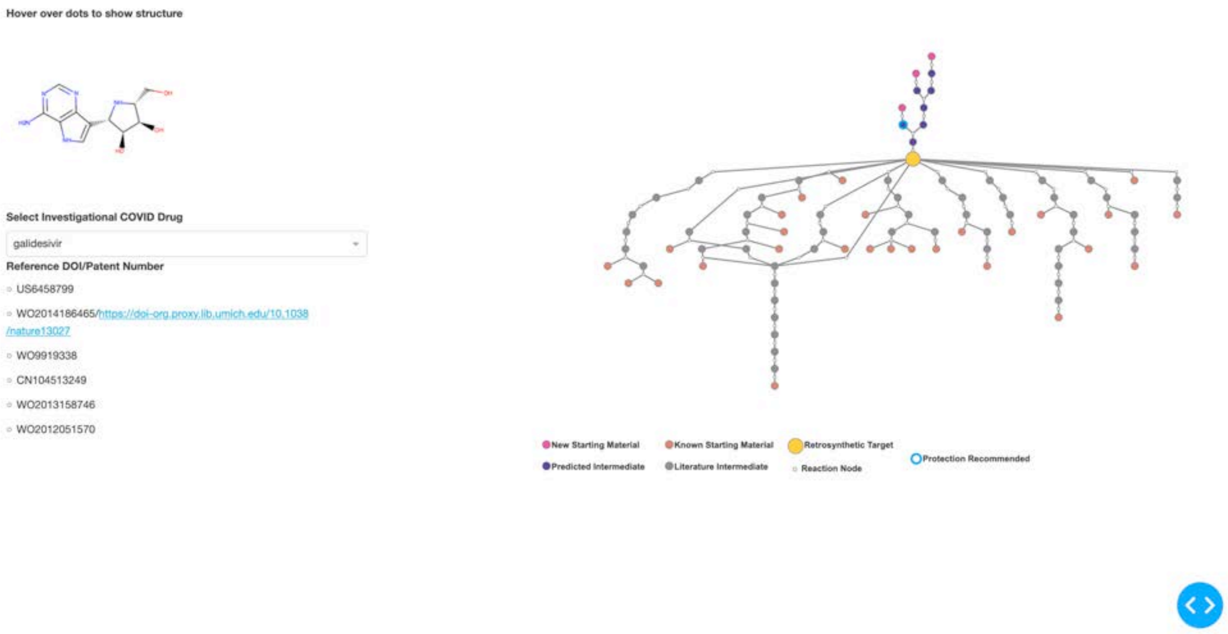
Select Investigational COVID Drug

nelfinavir

Reference DOI/Patent Number

- 10.1016/j.tet.2018.03.034/ 10.1021/jm9704098
- 10.1016/S0040-4039(00)01231-4
- 10.1021/acs.joc.8b00039
- 10.1021/cb500843r
- 10.1021/jo902048t
- 10.1021/jo991793e
- 10.3964/j.issn.1000-0593.2008.01.023
- US005705647A/US005925759A/10.1016/S0040-4039(01)01338-7
- US005962725A

157-Fig. 10 | Retrosynthetic analysis of nelfinavir on the visualizer webapp.



Hover over dots to show structure

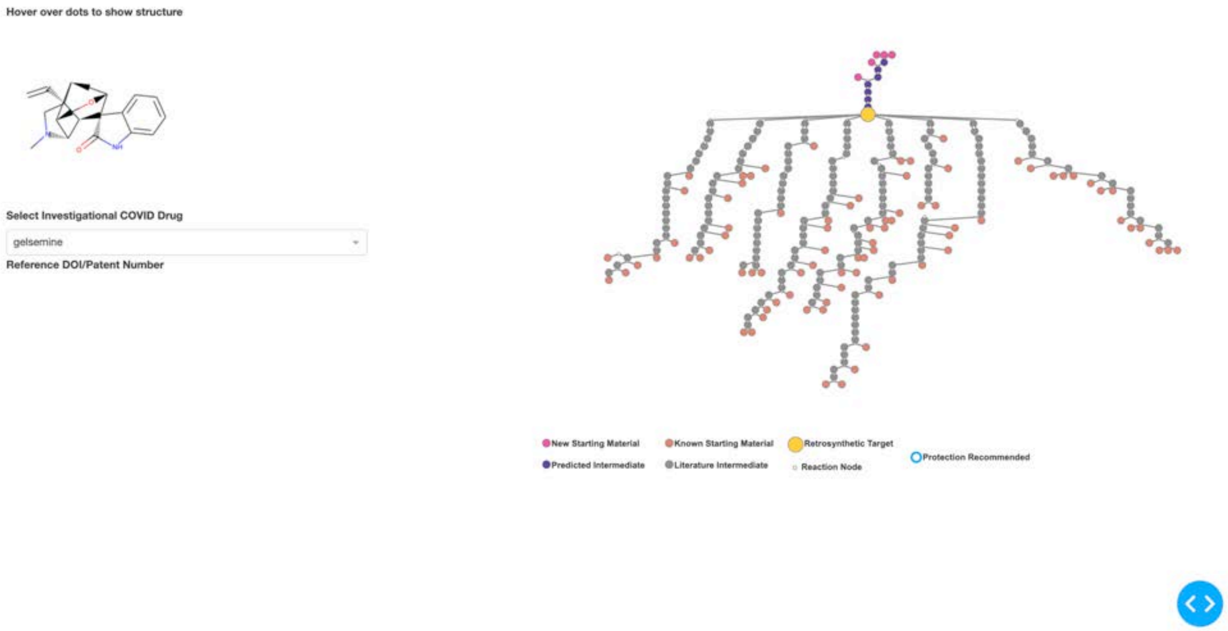
Select Investigational COVID Drug

galidesivir

Reference DOI/Patent Number

- US6458799
- WO2014186465/https://doi-org.proxy.lib.umich.edu/10.1038/nature13027
- WO9919338
- CN104513249
- WO2013158746
- WO2012051570

158-Fig. 11 | Retrosynthetic analysis of galidesivir on the visualizer webapp.



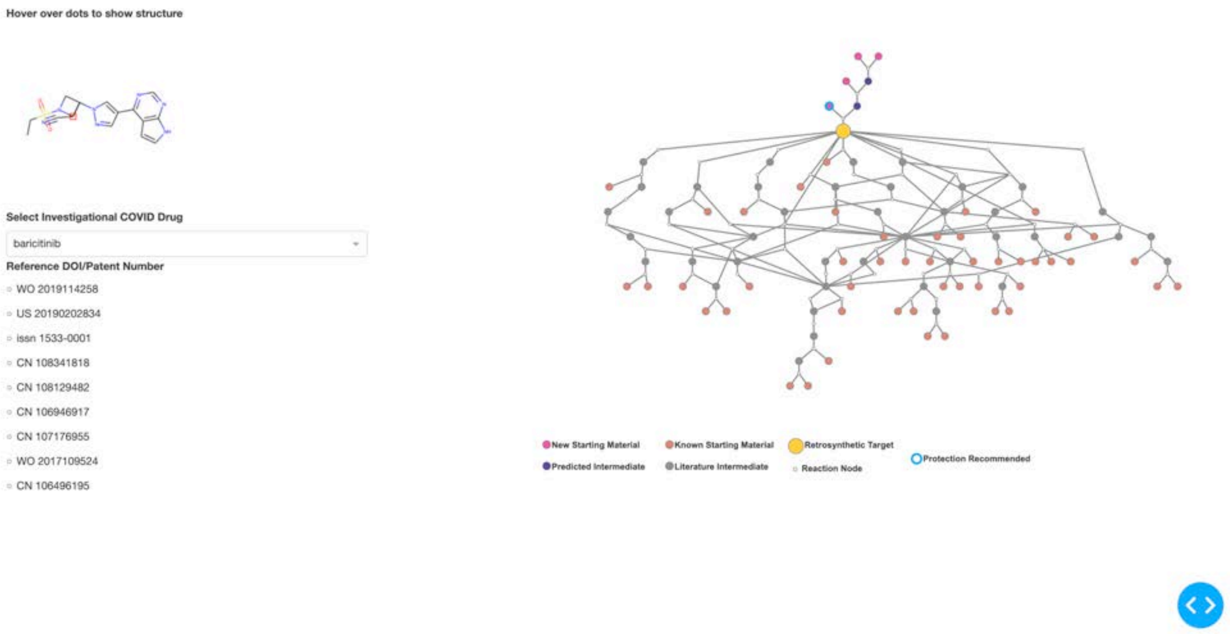
Hover over dots to show structure

Select Investigational COVID Drug

gelsemine

Reference DOI/Patent Number

159-Fig. 12 | Retrosynthetic analysis of gelsemine on the visualizer webapp.



Hover over dots to show structure

Select Investigational COVID Drug

baricitinib

Reference DOI/Patent Number

- WO 2019114258
- US 20190202834
- issn 1533-0001
- CN 108341818
- CN 108129482
- CN 106946917
- CN 107176955
- WO 2017109524
- CN 106496195

160-Fig. 13 | Retrosynthetic analysis of baricitinib on the visualizer webapp.

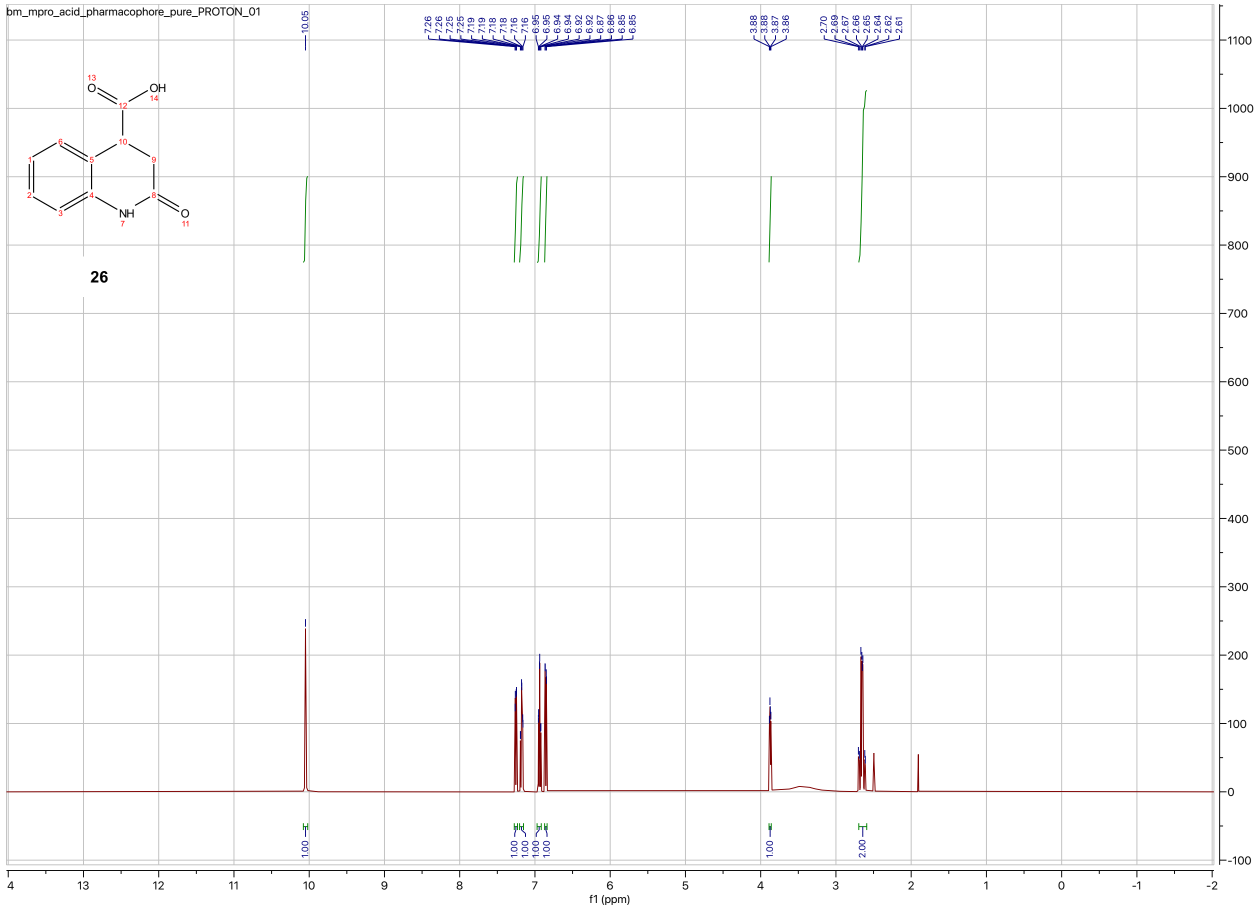
Chapter VIII. Conclusion

This thesis collates a large portion of the work I have completed throughout my Ph.D. studies. Throughout this experience, I helped create a synthetic laboratory from the ground up. I learned the difficulties in maintaining an organized lab and shared in the joy when scientific discoveries were made using the environment we had fostered. I became a better experimental chemist by learning how to run reactions efficiently in the lab and overcame the learning curve of column chromatography. I sat through weekly two-hour group meetings for five years and watched hundreds of chemical reactions presented and optimized. I cycled through several computers and wrote thousands and thousands of lines of code. I endured with the lab through phases of grant proposal struggles and celebrated winning many big-name grants at once after several cycles of edits. I watched Tim transition from an industrialist to an academic who is likely to soon win tenure. I too am not the same person that I was when I began working with Tim. When I had finished my undergraduate studies, I was not very exceptional on paper – Tim gave me a chance to distinguish myself and I took it. Now, as I head to MIT with a clearer vision of my future, I feel that my accolades better match my ego. Most importantly, what I had learned throughout my studies, under Tim's supervision, was why we as chemists perform and study synthesis. While it may be obvious to some, it was a hard-earned realization for me that all matter must be made. In fact, this is the overall theme of this thesis: a valuable molecule is a product just like any other item. It must be optimized and synthesized, and this is done through chemistry. A good reaction lets us create many molecules with what we have at hand, and these valuable reactions must be invented. I look forward in continuing aiding the creation of robust, automatable, eco-friendly, high-value reactions that will lead to the discovery of high-impact molecules.

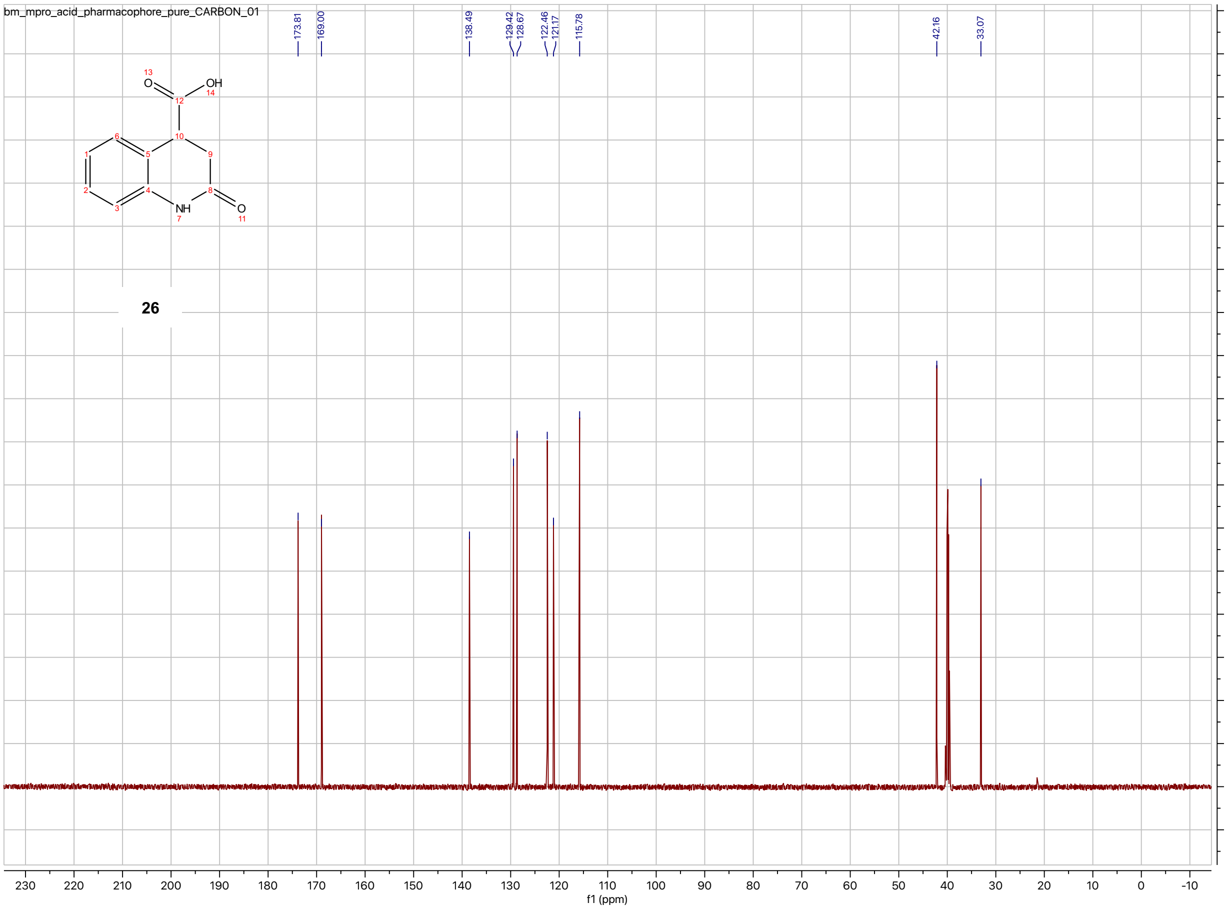
APPENDICIES

Appendix A

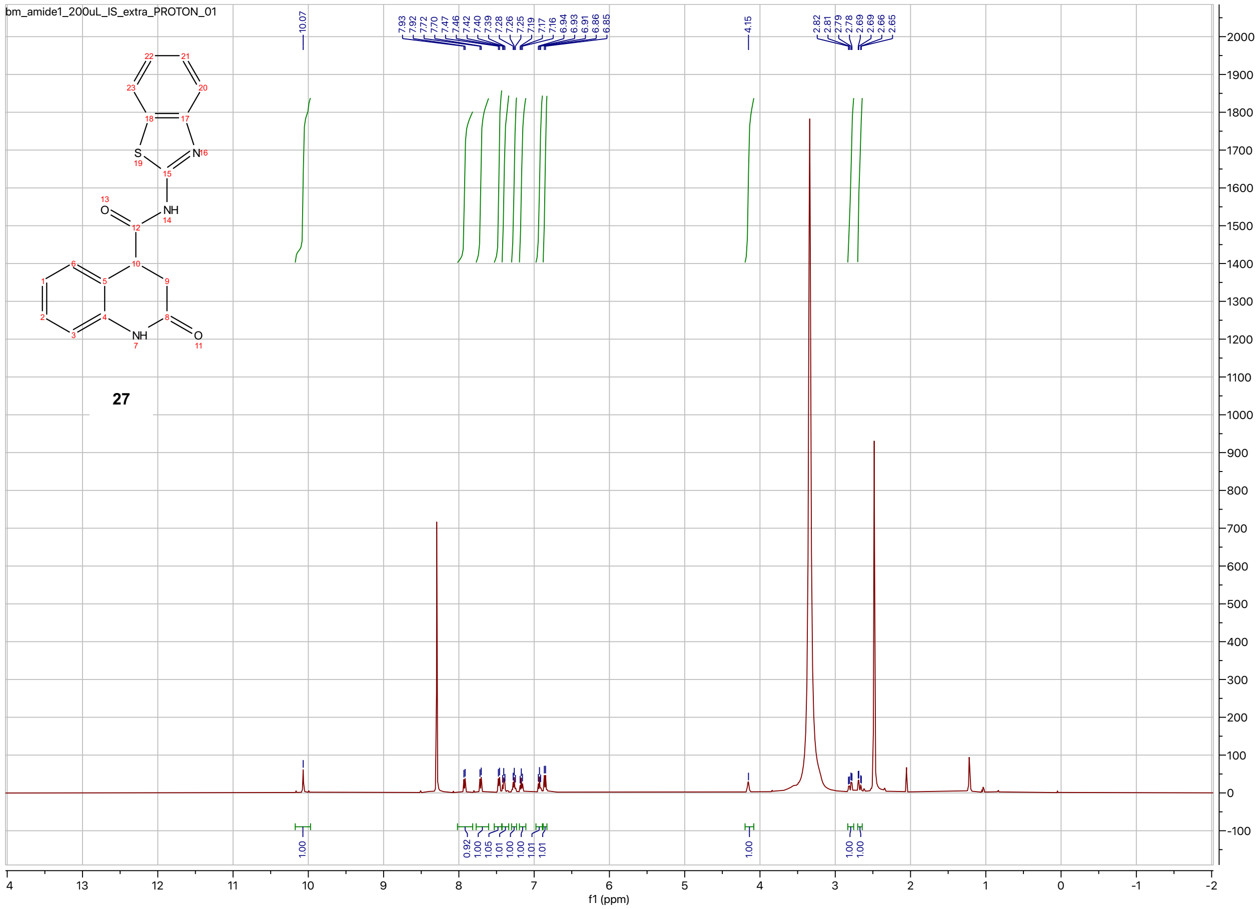
NMR Spectra of Isolated Compounds



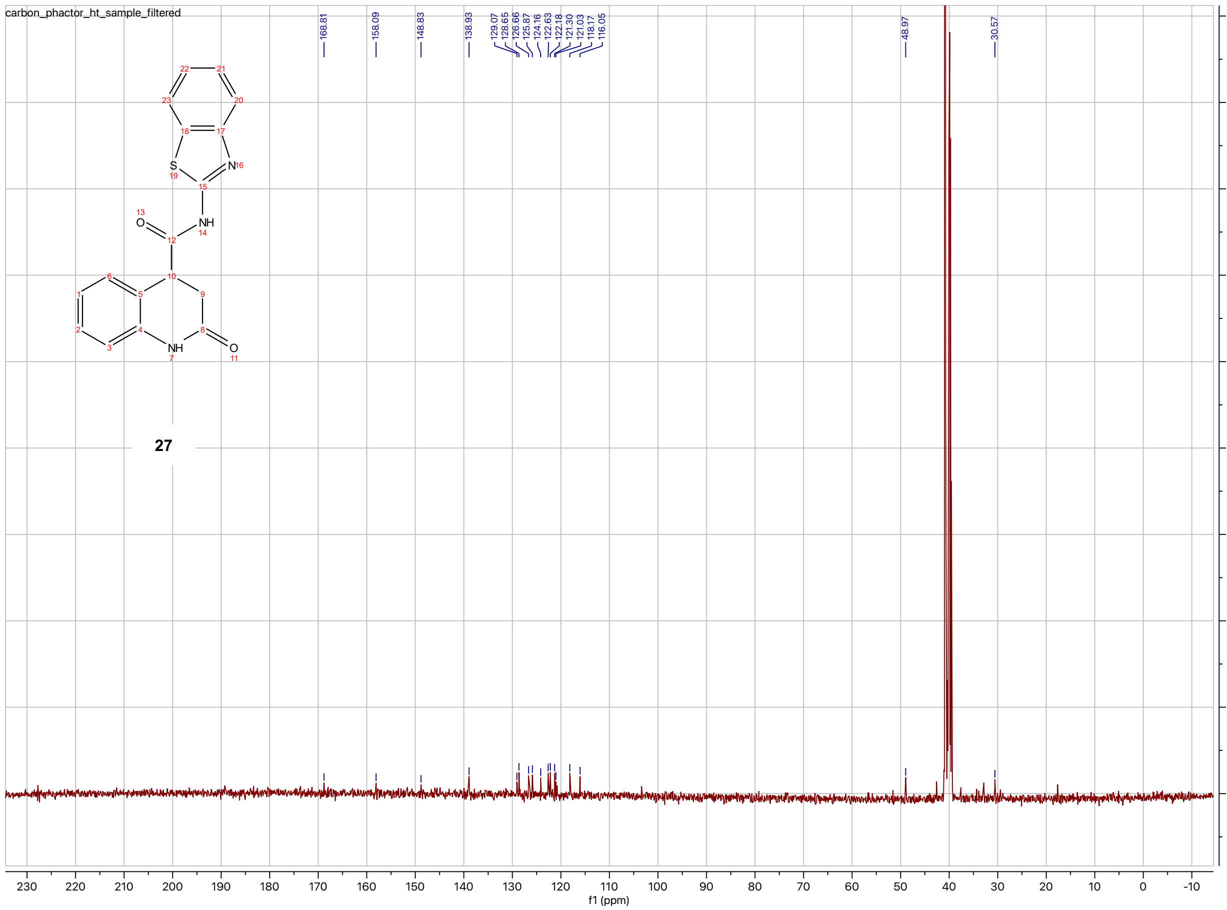
161-Fig. A1 | Proton NMR of 2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxylic acid.



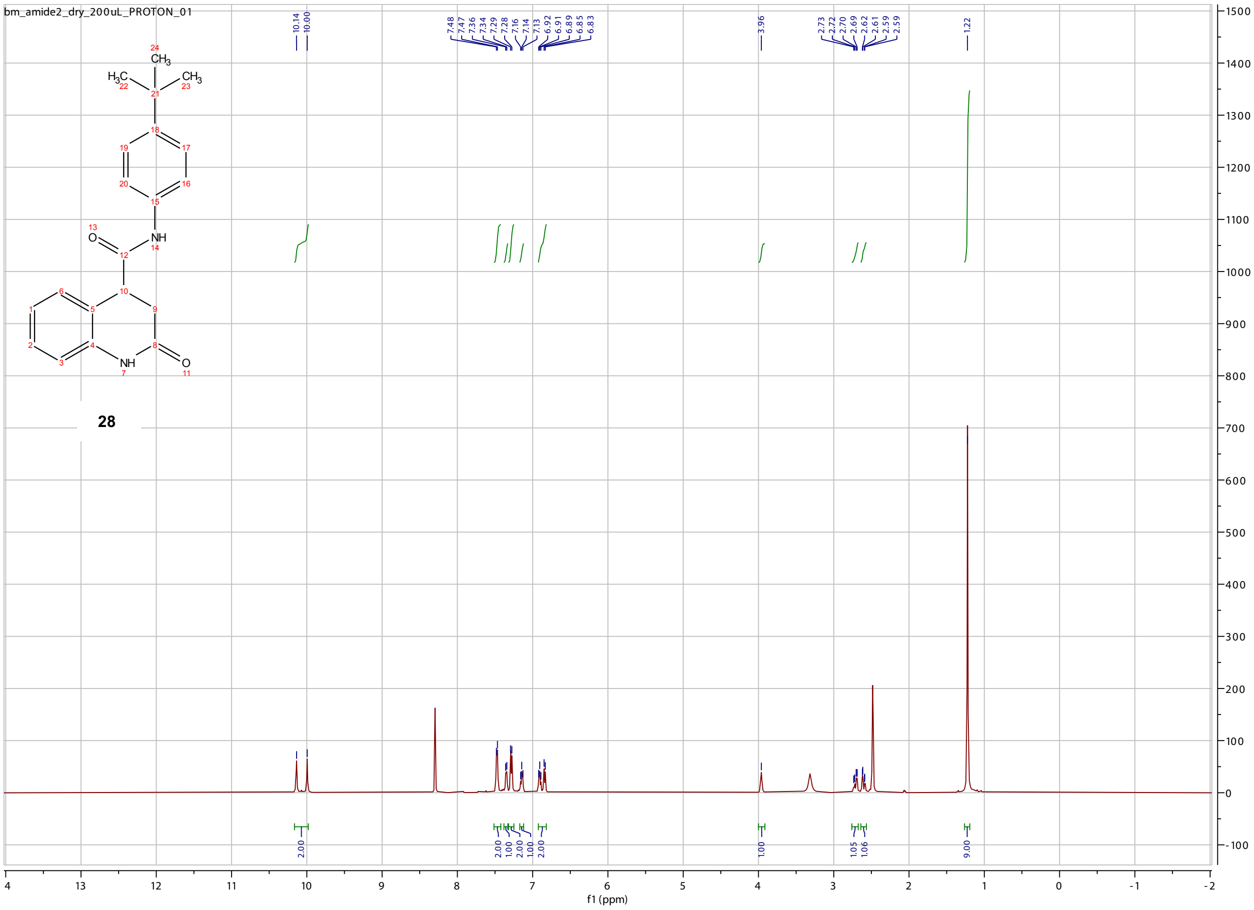
162-Fig. A2 | Carbon NMR of 2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxylic acid.



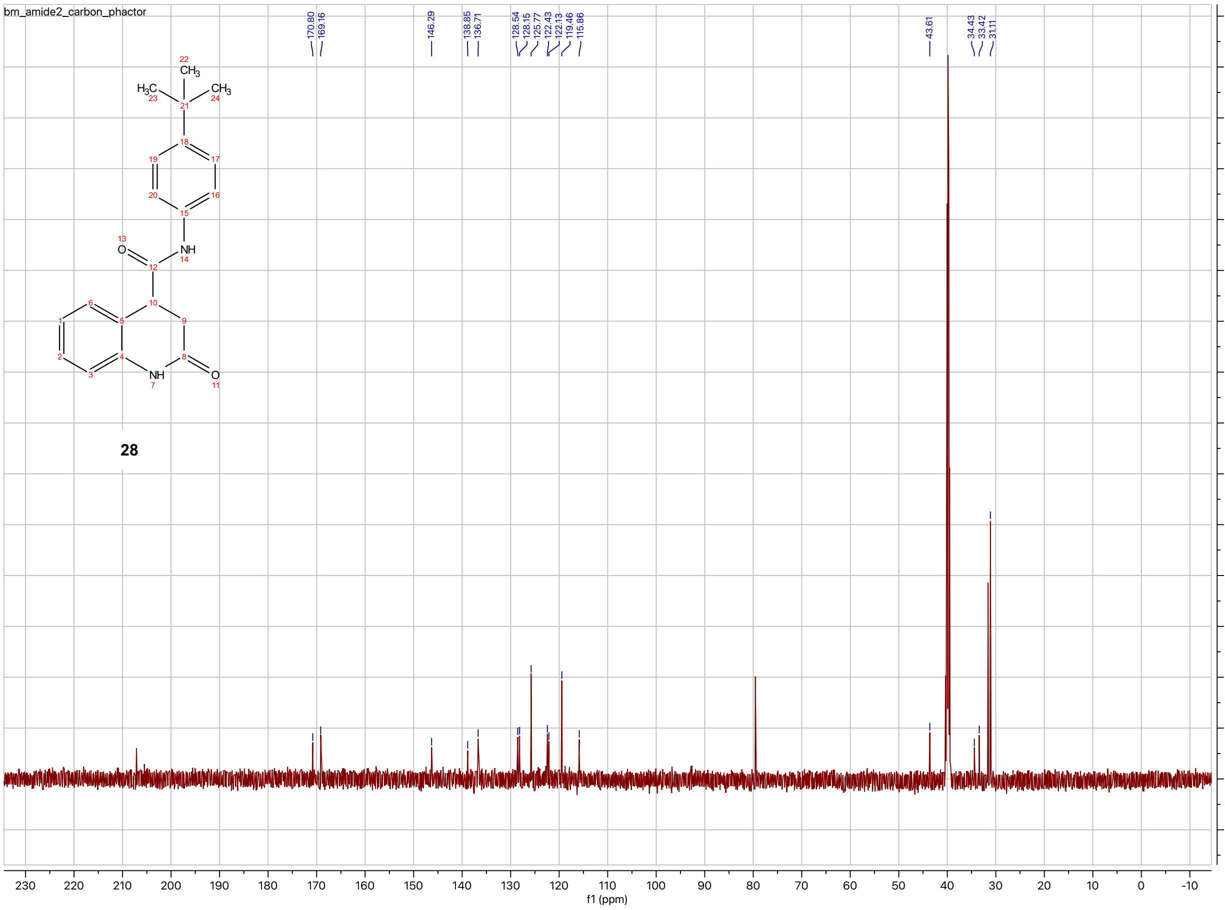
163-Fig. A3 | Proton NMR of N-(benzo[d]thiazol-2-yl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide.



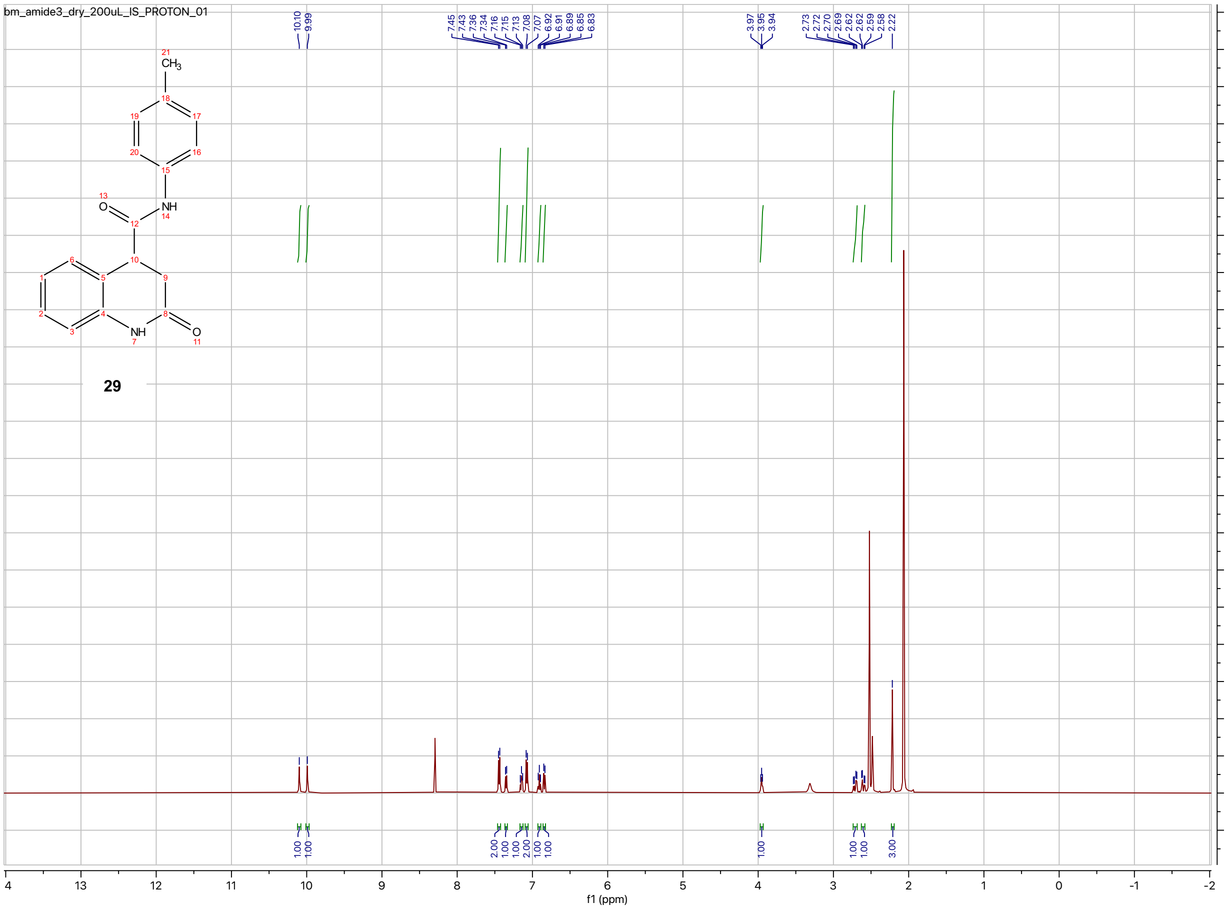
164-Fig. A4 | Carbon NMR of N-(benzo[d]thiazol-2-yl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide.



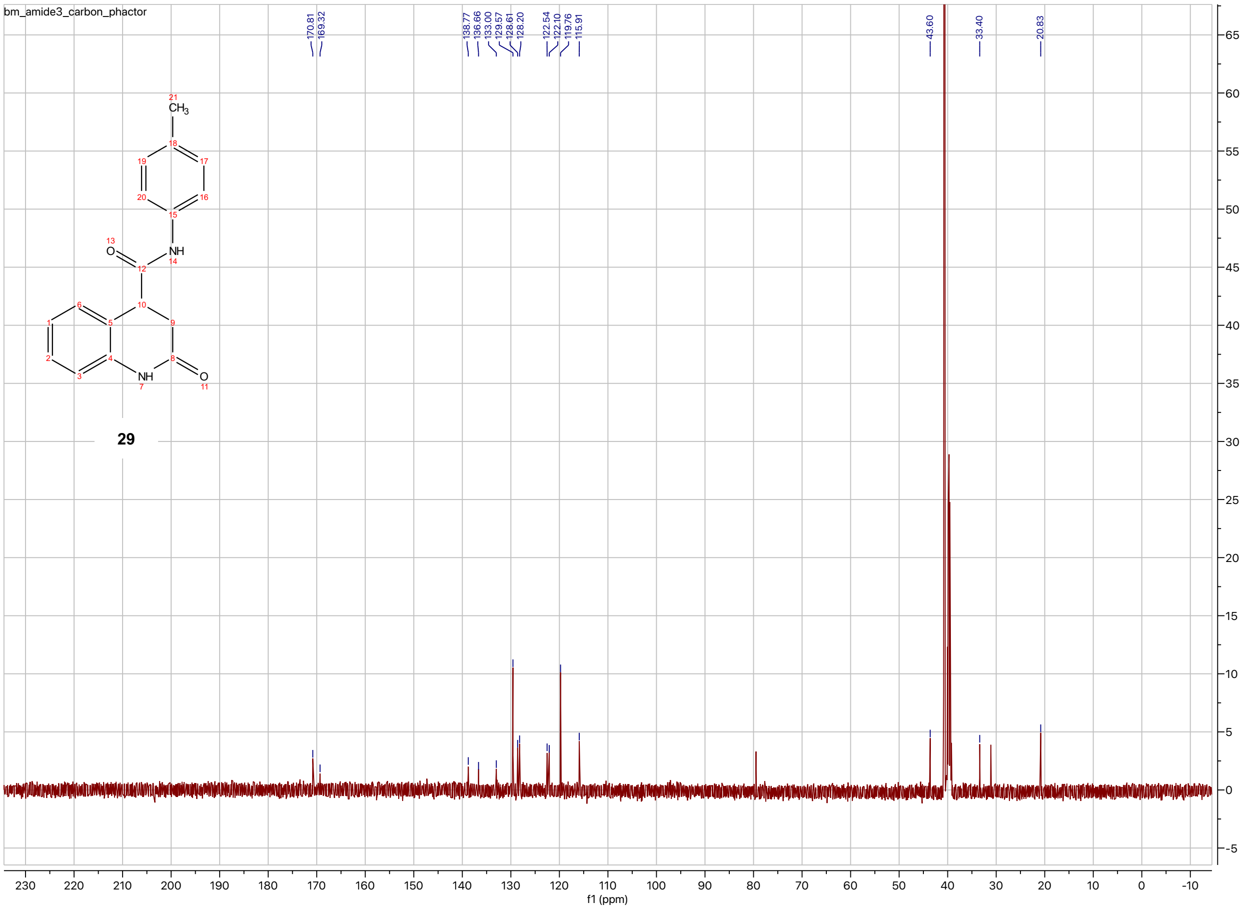
165-Fig. A5 | Proton NMR of N-(4-(tert-butyl)phenyl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide.



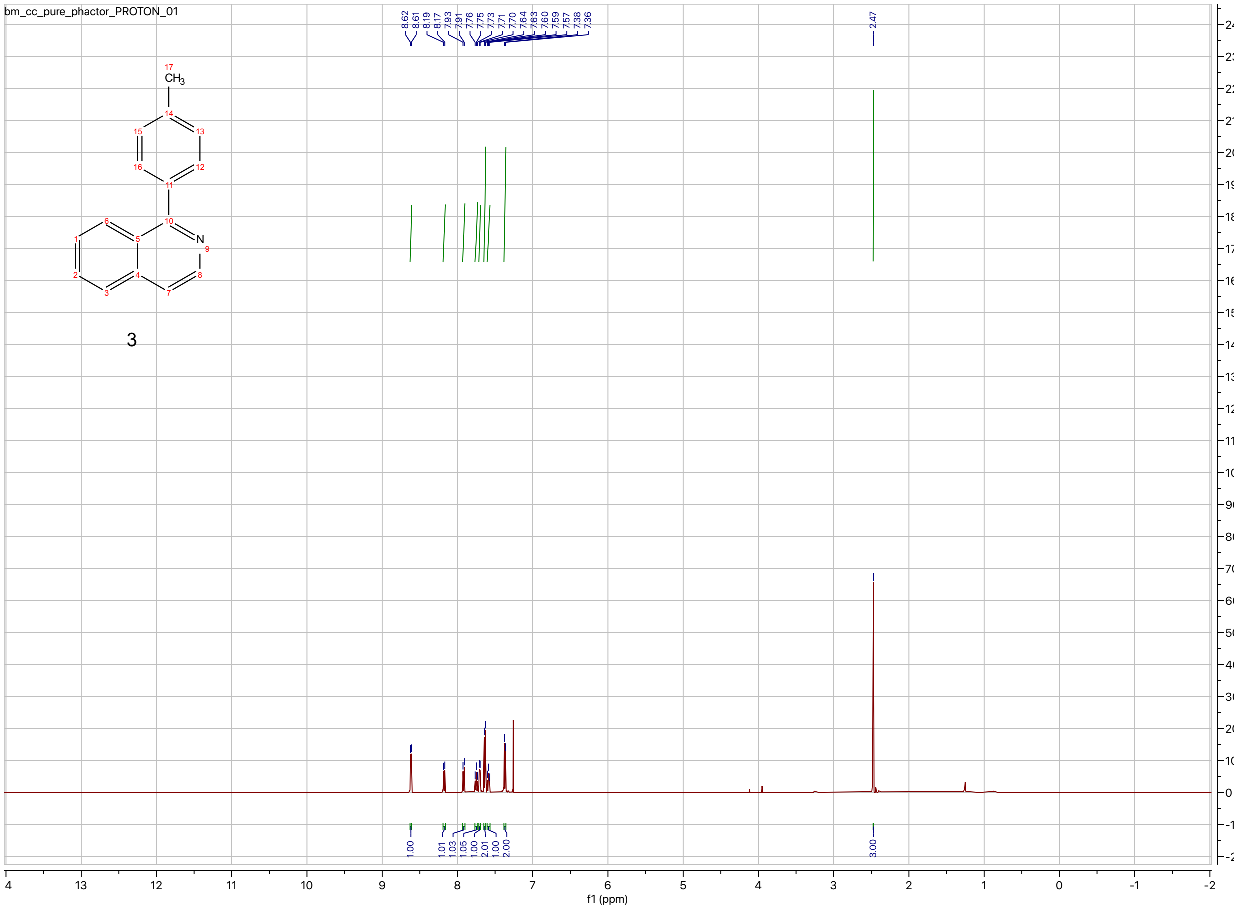
166-Fig. A6 | Proton NMR of N-(4-(tert-butyl)phenyl)-2-oxo-1,2,3,4-tetrahydroquinoline-4-carboxamide.



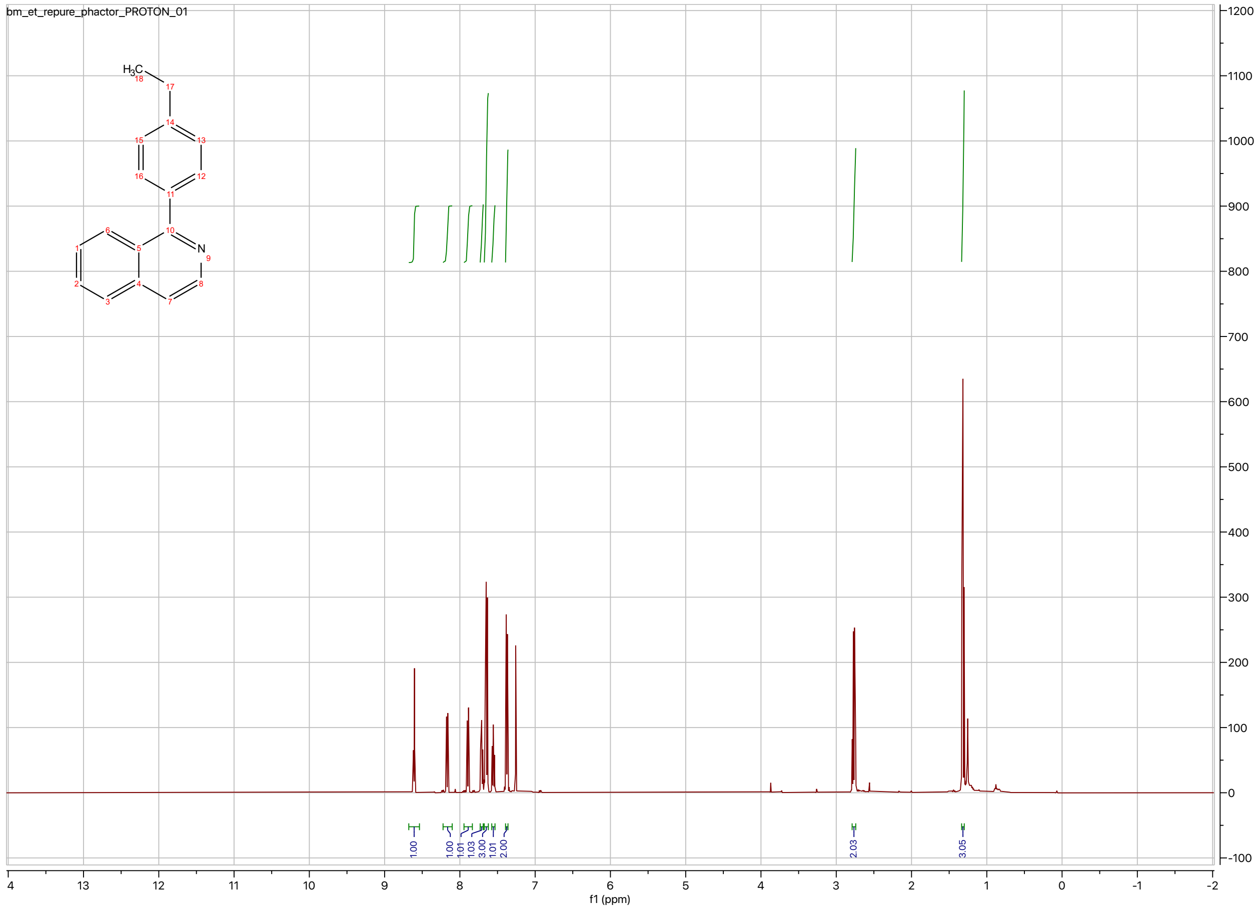
167-Fig. A7 | Proton NMR of 2-oxo-N-(p-tolyl)-1,2,3,4-tetrahydroquinoline-4-carboxamide.



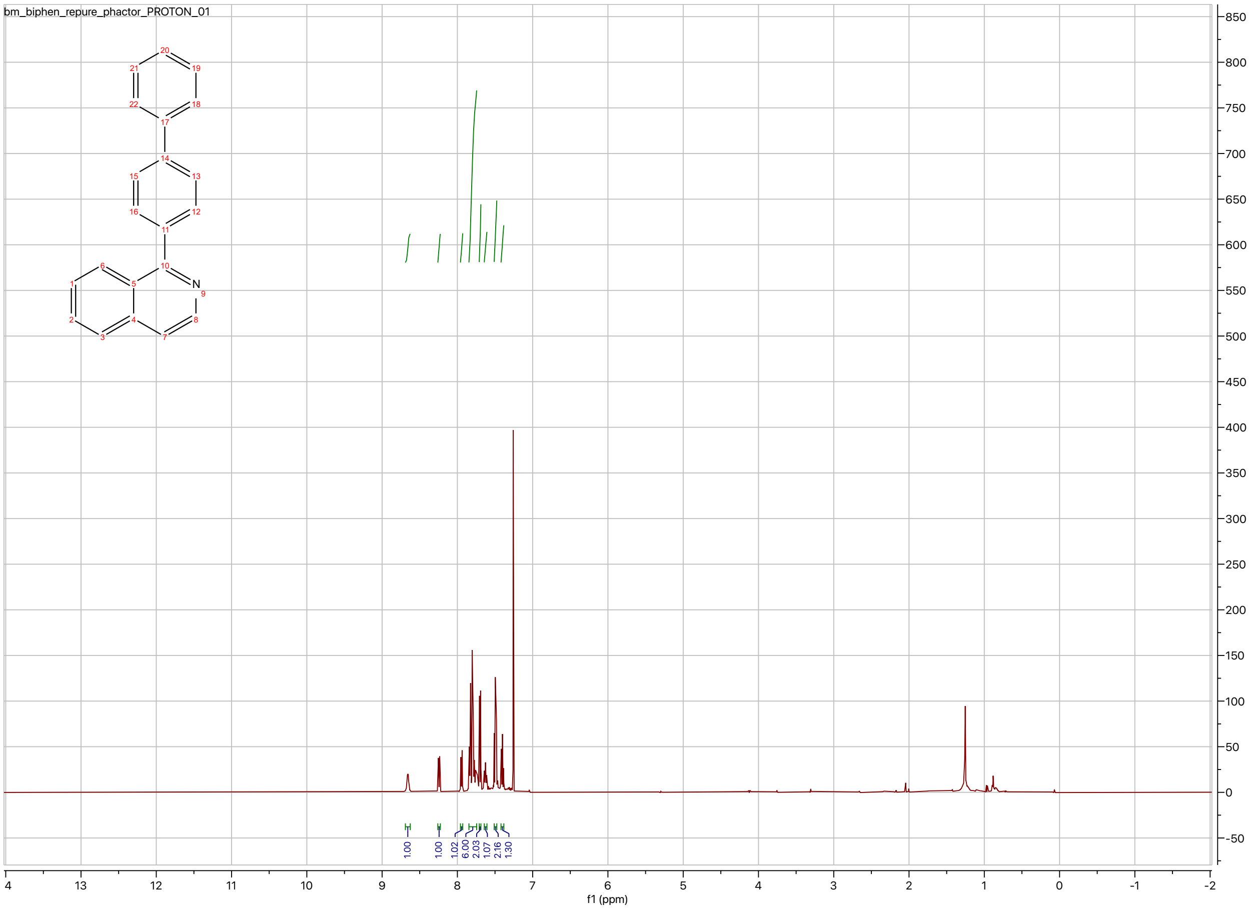
168-Fig. A8 | Carbon NMR of 2-oxo-N-(p-tolyl)-1,2,3,4-tetrahydroquinoline-4-carboxamide.



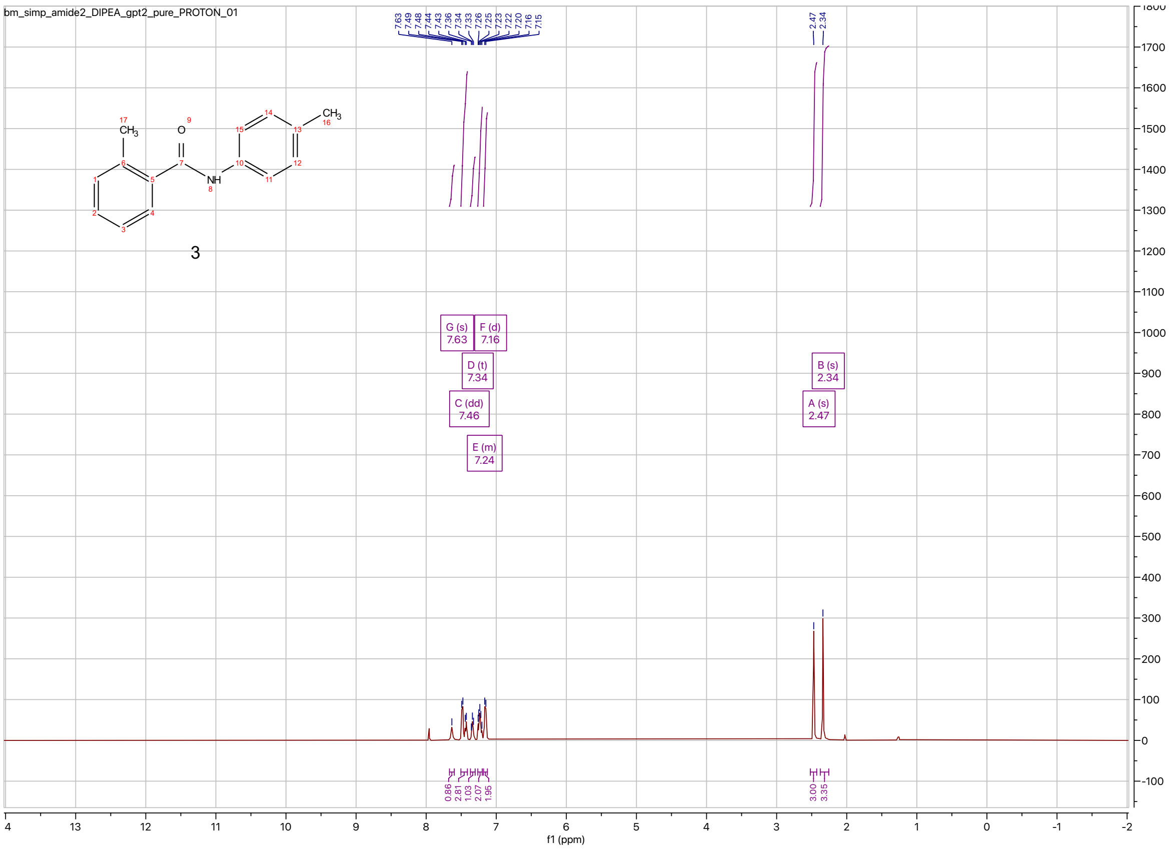
169-Fig. A9 | Proton NMR of 1-(*p*-tolyl)isoquinoline.



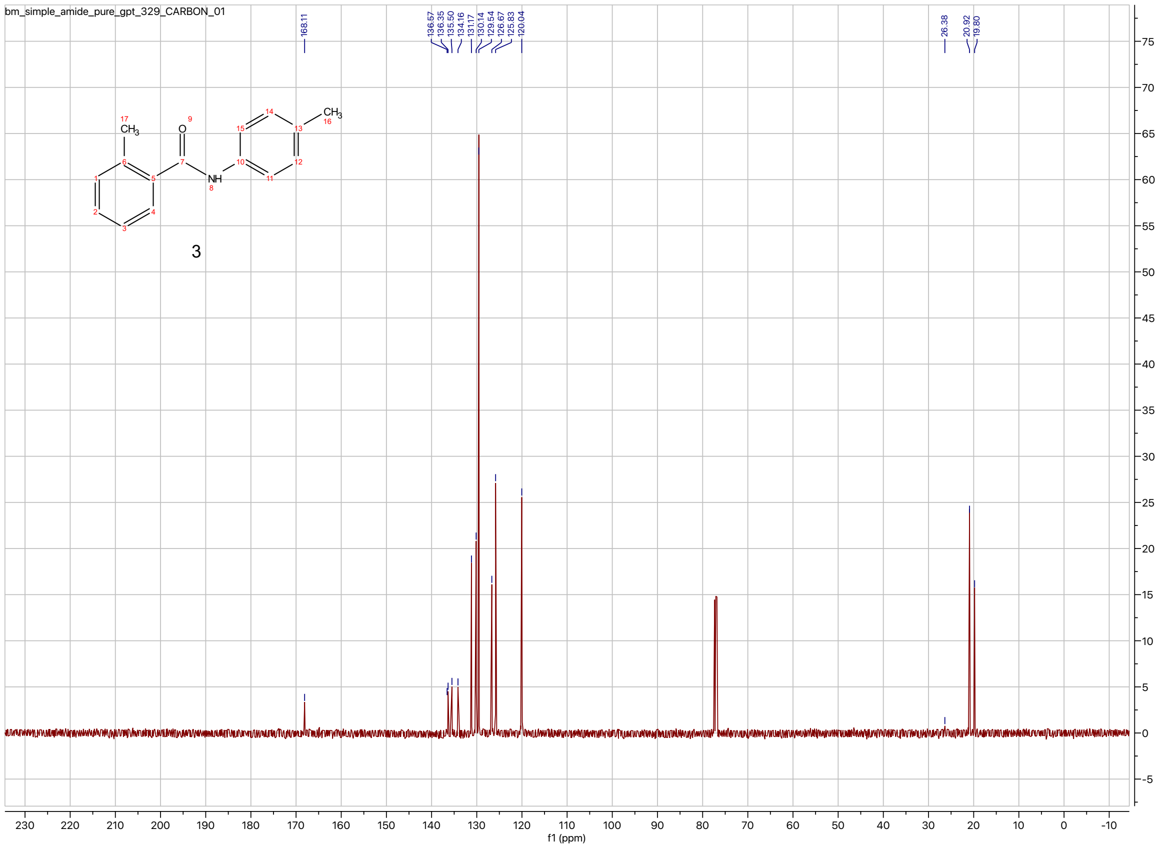
170-Fig. A10 | Proton NMR of 1-(4-ethylphenyl)isoquinoline.



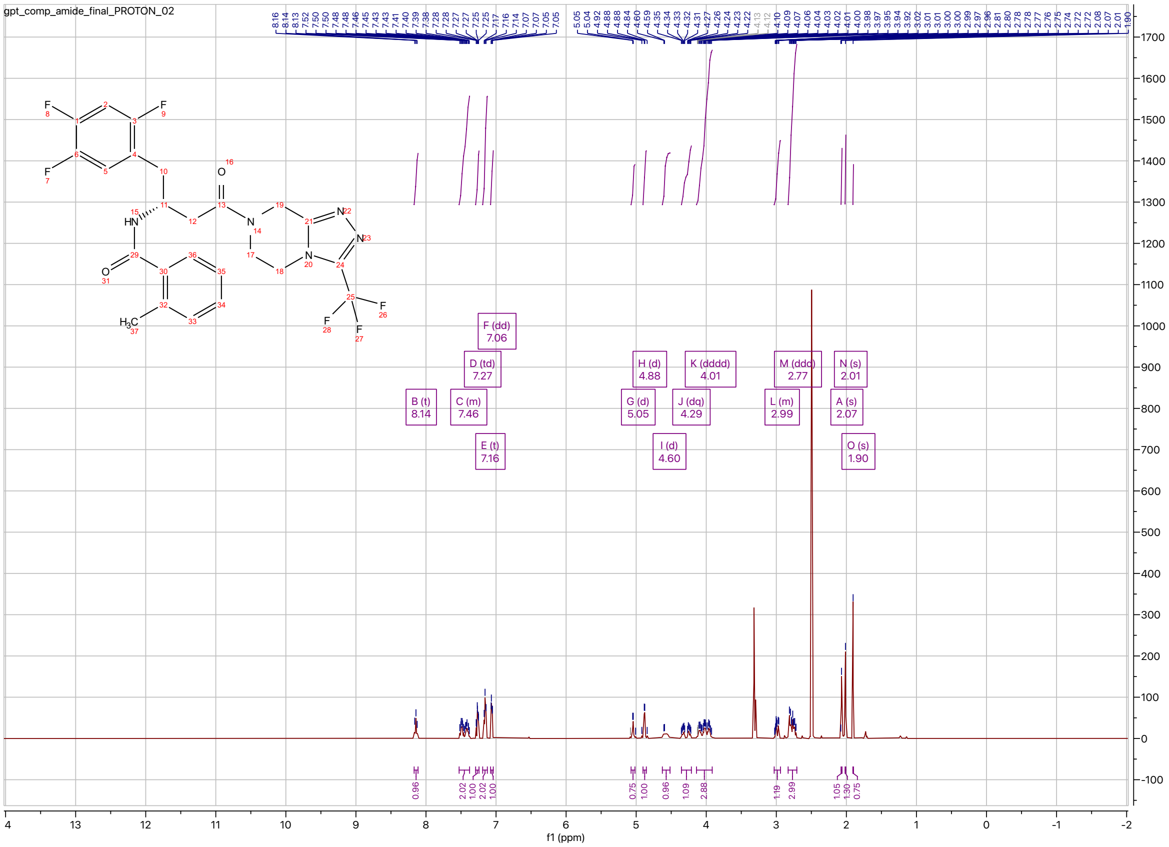
171-Fig. A11 | Proton NMR of 1-([1,1'-biphenyl]-4-yl)isoquinoline.



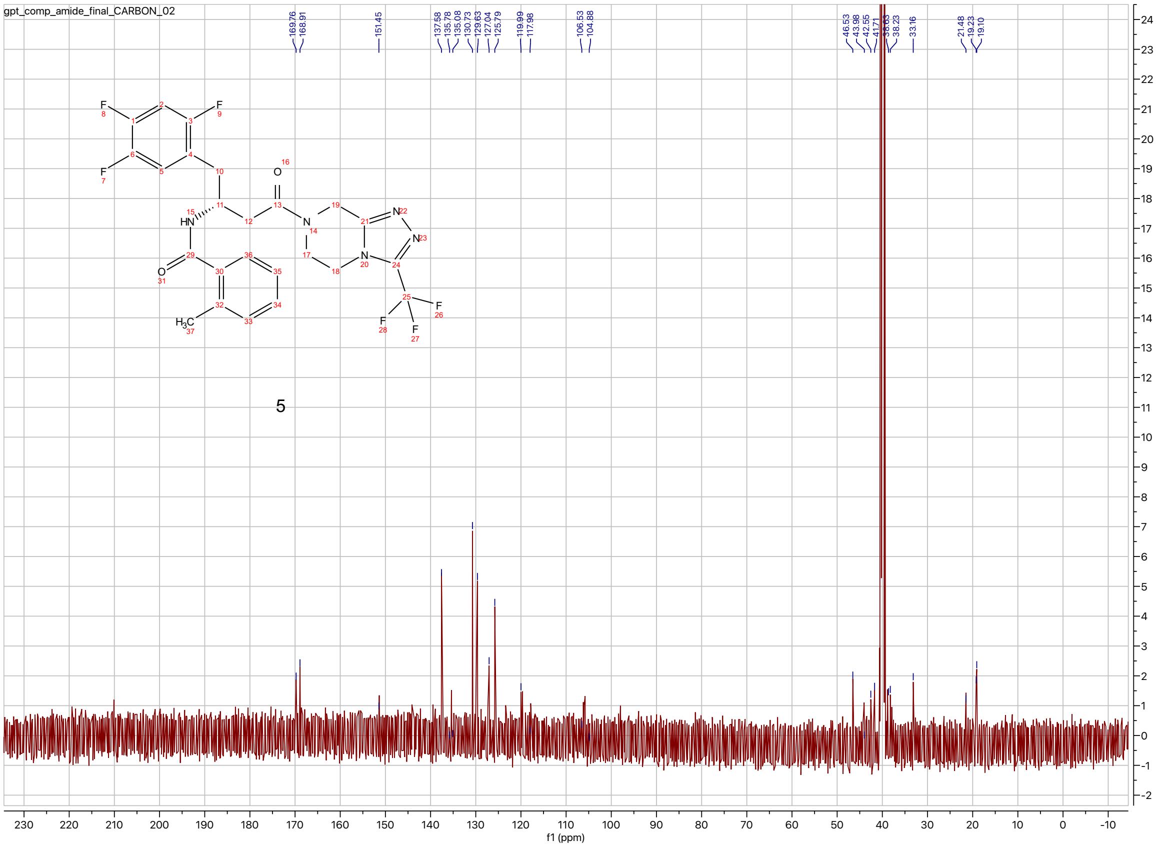
172-Fig. A12 | Proton NMR of 2-methyl-N-(p-tolyl)benzamide. ^{1}H NMR (500 MHz, $CDCl_3$) δ 7.63 (s, 1H), 7.46 (dd, *J* = 24.8, 7.7 Hz, 3H), 7.34 (t, *J* = 7.5 Hz, 1H), 7.26 – 7.20 (m, 2H), 7.16 (d, *J* = 7.9 Hz, 2H), 2.47 (s, 3H), 2.34 (s, 3H).



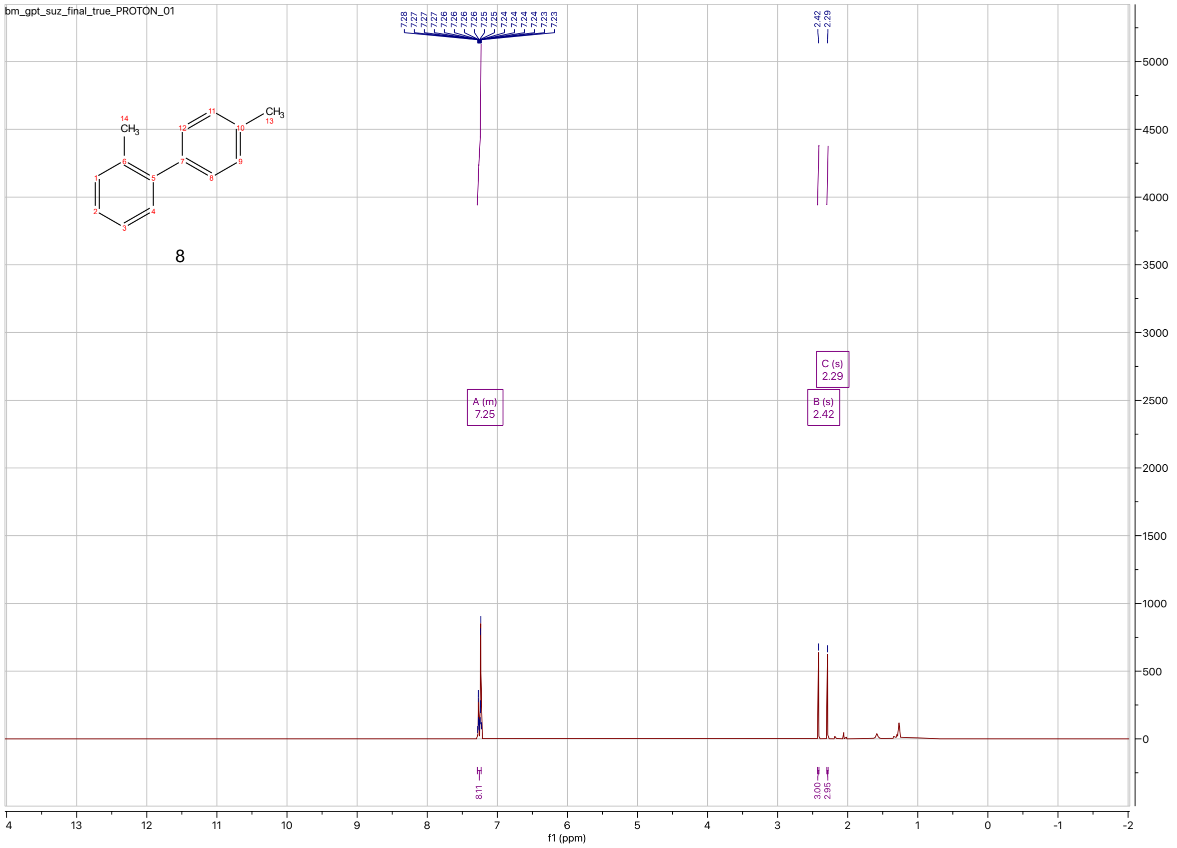
173-Fig. A13 | Carbon NMR of 2-methyl-N-(p-tolyl)benzamide. ^{13}C NMR (126 MHz, cdcl$_3$) δ 168.11, 136.57, 136.35, 135.50, 134.16, 131.17, 130.14, 129.54, 126.67, 125.83, 120.04, 26.38, 20.92, 19.80.



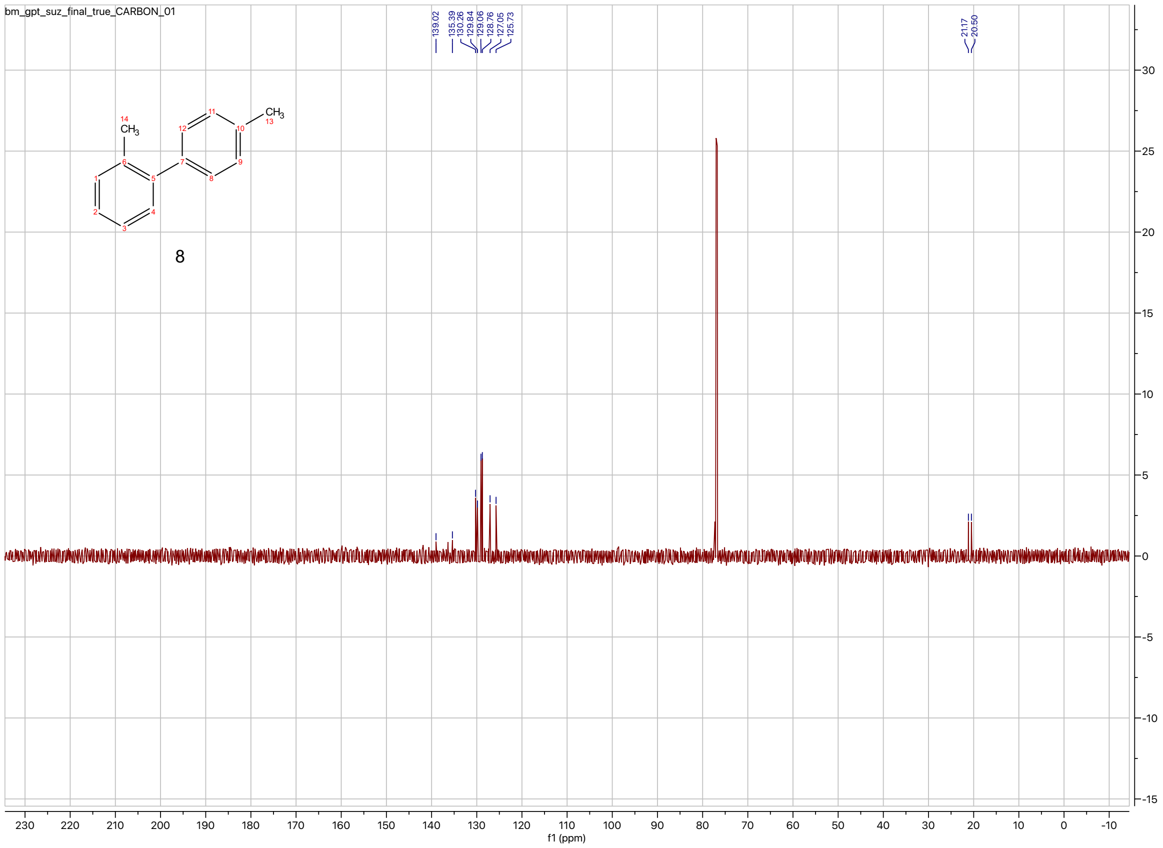
174-Fig. A14 | Proton NMR of (R)-2-methyl-N-(4-oxo-4-(3-(trifluoromethyl)-5,6-dihydro-[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl)-1-(2,4,5-trifluorophenyl)butan-2-yl)benzamide. ^{1}H NMR (500 MHz, dmso) δ 8.14 (t, *J* = 8.8 Hz, 1H), 7.53 – 7.38 (m, 2H), 7.27 (td, *J* = 7.6, 1.4 Hz, 1H), 7.16 (t, *J* = 6.8 Hz, 2H), 7.06 (dd, *J* = 7.8, 1.5 Hz, 1H), 5.05 (d, *J* = 3.9 Hz, 1H), 4.88 (d, *J* = 3.7 Hz, 1H), 4.60 (d, *J* = 4.8 Hz, 1H), 4.29 (dq, *J* = 38.5, 6.5 Hz, 1H), 4.01 (dddd, *J* = 38.2, 20.6, 14.5, 7.8 Hz, 3H), 3.03 – 2.94 (m, 1H), 2.77 (ddd, *J* = 23.4, 14.8, 8.4 Hz, 3H), 2.07 (s, 1H), 2.01 (s, 1H), 1.90 (s, 1H).



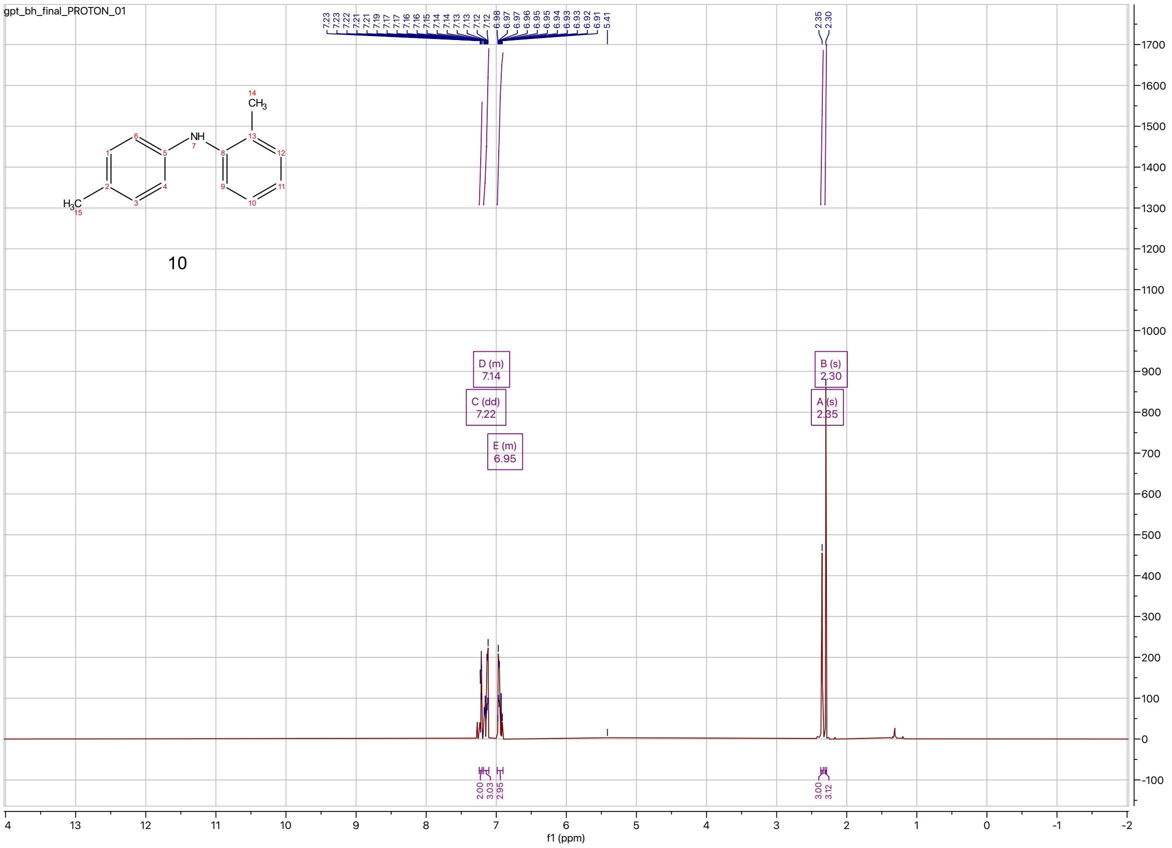
175-Fig. A15 | Carbon NMR of (R)-2-methyl-N-(4-oxo-4-(3-(trifluoromethyl)-5,6-dihydro-[1,2,4]triazolo[4,3-a]pyrazin-7(8H)-yl)-1-(2,4,5-trifluorophenyl)butan-2-yl)benzamide. ^{13}C NMR (126 MHz, dmso) δ 169.76, 168.91, 151.45, 137.58, 135.78, 135.08, 130.73, 129.63, 127.04, 125.79, 119.99, 117.98, 106.53, 104.88, 46.53, 43.98, 42.55, 41.71, 38.63, 38.23, 33.16, 21.48, 19.23, 19.10.



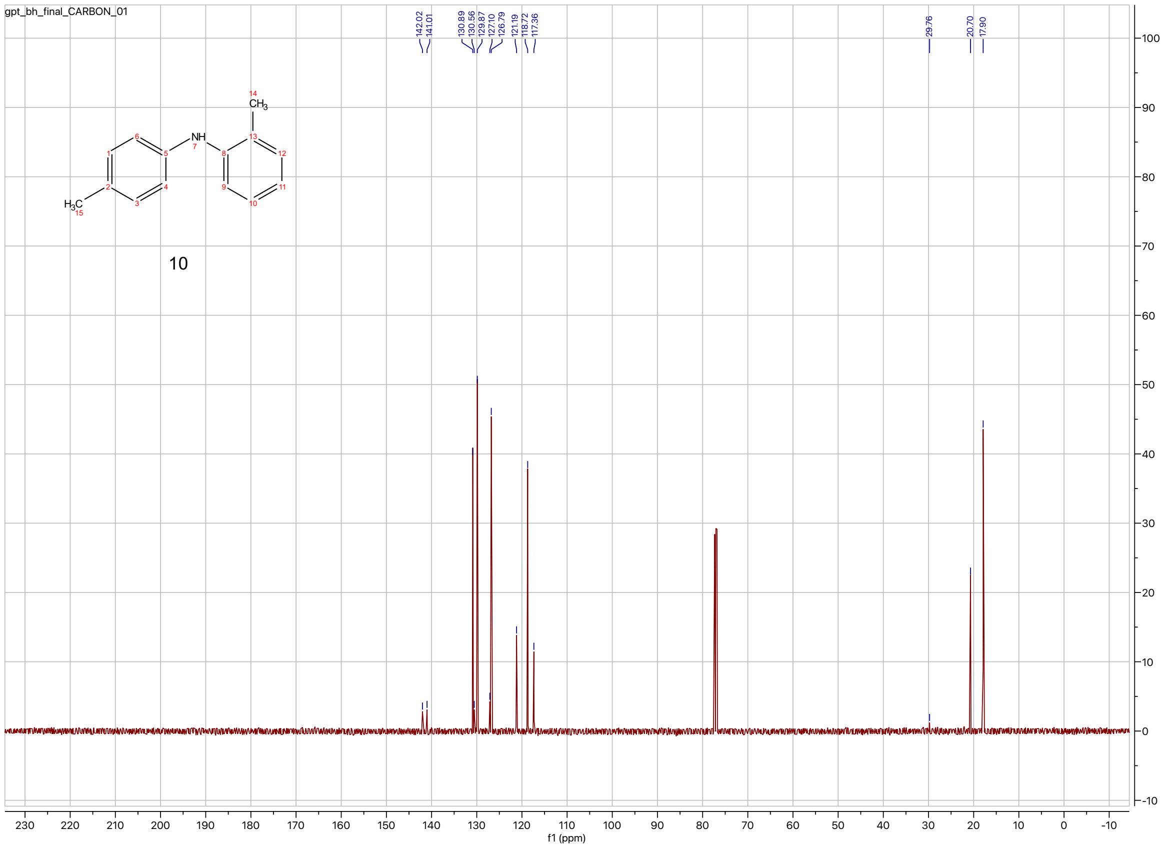
176-Fig. A16 | Proton NMR of 2,4'-dimethyl-1,1'-biphenyl. ^{1}H NMR (499 MHz, cdcl$_3$) δ 7.28 – 7.23 (m, 8H), 2.42 (s, 3H), 2.29 (s, 3H).



177-Fig. A17 | Carbon NMR of 2,4'-dimethyl-1,1'-biphenyl. ^{13}C NMR (126 MHz, cdcl$_3$) δ 139.02, 135.39, 130.26, 129.84, 129.06, 128.76, 127.05, 125.73, 21.17, 20.50.



178-Fig. A18 | Proton NMR of 2-methyl-N-(p-tolyl)aniline. ^{1}H NMR (499 MHz, cdcl$_3$) δ 7.22 (dd, *J* = 8.0, 1.3 Hz, 2H), 7.18 – 7.11 (m, 3H), 6.98 – 6.90 (m, 3H), 2.35 (s, 3H), 2.30 (s, 3H).



179-Fig. A19 | Carbon NMR of 2-methyl-N-(p-tolyl)aniline. ^{13}C NMR (126 MHz, cdcl$_3$) δ 142.02, 141.01, 130.89, 130.56, 129.87, 127.10, 126.79, 121.19, 118.72, 117.36, 29.76, 20.70, 17.90.

BIBLIOGRAPHY

(1) Mahjour, B.; Shen, Y.; Liu, W.; Cernak, T. A map of the amine–carboxylic acid coupling system. *Nature* **2020**, *580* (7801), 71-75. DOI: 10.1038/s41586-020-2142-y.

(2) Rains, E. M.; Sloane, N. J. On Cayley's enumeration of alkanes (or 4-valent trees). *Journal of Integer Sequences* **1999**, *2* (99.1), 1. DOI: 10.48550/arXiv.math/0207176.

(3) Henze, H. R.; Blair, C. M. THE NUMBER OF ISOMERIC HYDROCARBONS OF THE METHANE SERIES. *Journal of the American Chemical Society* **1931**, *53* (8), 3077–3085. DOI: 10.1021/ja01359a034.

(4) Henze, H. R.; Blair, C. M. THE NUMBER OF STRUCTURALLY ISOMERIC ALCOHOLS OF THE METHANOL SERIES. *Journal of the American Chemical Society* **1931**, *53* (8). DOI: 10.1021/ja01359a027.

(5) Perry, D. THE NUMBER OF STRUCTURAL ISOMERS OF CERTAIN HOMOLOGS OF METHANE AND METHANOL. *Journal of the American Chemical Society* **1932**, *54* (7). DOI: 10.1021/ja01346a035.

(6) Parks, C. A.; Hendrickson, J. B. Enumeration of monocyclic and bicyclic carbon skeletons. *Journal of Chemical Information and Computer Sciences* **1991**, *31* (2). DOI: 10.1021/ci00002a021.

(7) Bauer, J. IGOR2: a PC-program for generating new reactions and molecular structures. *Tetrahedron Computer Methodology* **1989**, *2* (5). DOI: 10.1016/0898-5529(89)90034-1.

(8) Spialter, L. The Atom Connectivity Matrix (ACM) and its Characteristic Polynomial (ACMCP): A New Computer-Oriented Chemical Nomenclature. *Journal of the American Chemical Society* **1963**, *85* (13), 2012-2013. DOI: 10.1021/ja00896a022.

(9) Vléduts, G. É. Concerning one system of classification and codification of organic reactions. *Information Storage and Retrieval* **1963**, *1* (2), 117-146. DOI: https://doi.org/10.1016/0020-0271(63)90013-5.

(10) Dugundji, J.; Ugi, I. An algebraic model of constitutional chemistry as a basis for chemical computer programs. In *Computers in Chemistry*, Berlin, Heidelberg, 1973//, 1973; Springer Berlin Heidelberg: pp 19-64.

(11) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New elements in the representation of the logical structure of chemistry by qualitative mathematical models and corresponding data structures. In *Computer Chemistry*, Springer Berlin Heidelberg, 1993; pp 199-233.

(12) Ugi, I.; Gillespie, P. Representation of Chemical Systems and Interconversions bybe Matrices and Their Transformation Properties. *Angewandte Chemie International Edition in English* **1971**, *10* (12). DOI: 10.1002/anie.197109141.

(13) Jochum, C.; Gasteiger, J.; Ugi, I. The Principle of Minimum Chemical Distance(PMCD). *Angewandte Chemie International Edition in English* **1980**, *19* (7). DOI: 10.1002/anie.198004953.

(14) Bauer, J.; Fontain, E.; Forstmeyer, D.; Ugi, I. Interactive generation of organic reactions by IGOR 2 and the PC-assisted discovery of a new reaction. *Tetrahedron Computer Methodology* **1988**, *1* (2). DOI: 10.1016/0898-5529(88)90017-6.

(15) Herges, R.; Ugi, I. Synthesis of Seven-Membered Rings by [(?2+?2)+?2] Cycloaddition to Homodienes. *Angewandte Chemie International Edition in English* **1985**, *24* (7). DOI: 10.1002/anie.198505941.

(16) Hendrickson, J. B. A Systematic Organization of Synthetic Reactions. *Journal of Chemical Information and Computer Sciences* **1979**, *19* (3). DOI: 10.1021/ci60019a004.

(17) Hendrickson, J. B. Comprehensive System for Classification and Nomenclature of Organic Reactions. *Journal of Chemical Information and Computer Sciences* **1997**, *37* (5). DOI: 10.1021/ci970040v.

(18) Hendrickson, J. B. Systematic Signatures for Organic Reactions. *Journal of Chemical Information and Modeling* **2010**, *50* (8). DOI: 10.1021/ci1000482.

(19) Hendrickson, J. B. Systematic synthesis design. IV. Numerical codification of construction reactions. *Journal of the American Chemical Society* **1975**, *97* (20). DOI: 10.1021/ja00853a023.

(20) Arens, J. F. A formalism for the classification and design of organic reactions. II. The classes of (+ −)n + and (− +)n − reactions. *Recueil des Travaux Chimiques des Pays-Bas* **2010**, *98* (6). DOI: 10.1002/recl.19790980606.

(21) Arens, J. F. A formalism for the classification and design of organic reactions. I. The class of (− +) <sub> <i>n</i> </sub> reactions. *Recueil des Travaux Chimiques des Pays-Bas* **1979**, *98* (4). DOI: 10.1002/recl.19790980403.

(22) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J Chem Inf Model* **2015**, *55* (1), 39-53. DOI: 10.1021/ci5006614. From NLM Medline.

(23) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *Journal of Chemical Information and Modeling* **2017**, *57* (8). DOI: 10.1021/acs.jcim.6b00601.

(24) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous

Representation of Molecules. *ACS Cent Sci* **2018**, *4* (2), 268-276. DOI: 10.1021/acscentsci.7b00572.

(25) Reymond, J.-L. The Chemical Space Project. *Accounts of Chemical Research* **2015**, *48* (3). DOI: 10.1021/ar500432k.

(26) Szymkuc, S. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*. DOI: 10.1002/anie.201506101.

(27) Klucznik, T. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chemistry* **2018**, *4*. DOI: 10.1016/j.chempr.2018.02.002.

(28) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W. Computational planning of the synthesis of complex natural products. *Nature* **2020**, *588* (7836), 83-88.

(29) Narumi, T.; Hayashi, R.; Tomita, K.; Kobayashi, K.; Tanahara, N.; Ohno, H.; Naito, T.; Kodama, E.; Matsuoka, M.; Oishi, S.; Fujii, N. Synthesis and biological evaluation of selective CXCR4 antagonists containing alkene dipeptide isosteres. *Organic & Biomolecular Chemistry* **2010**, *8* (3), 616-621, 10.1039/B917236J. DOI: 10.1039/B917236J.

(30) Boström, J.; Brown, D. G.; Young, R. J.; Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery* **2018**, *17* (10), 709-727. DOI: 10.1038/nrd.2018.116.

(31) Liu, J.; Patlewicz, G.; Williams, A. J.; Thomas, R. S.; Shah, I. Predicting Organ Toxicity Using in Vitro Bioactivity Data and Chemical Structure. *Chemical Research in Toxicology* **2017**, *30* (11), 2046-2059. DOI: 10.1021/acs.chemrestox.7b00084.

(32) Waring, M. J.; Arrowsmith, J.; Leach, A. R.; Leeson, P. D.; Mandrell, S.; Owen, R. M.; Pairaudeau, G.; Pennie, W. D.; Pickett, S. D.; Wang, J.; Wallace, O.; Weir, A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nature Reviews Drug Discovery* **2015**, *14* (7), 475-486. DOI: 10.1038/nrd4609.

(33) Liu, R.; Li, X.; Lam, K. S. Combinatorial chemistry in drug discovery. *Current Opinion in Chemical Biology* **2017**, *38*, 117-126. DOI: https://doi.org/10.1016/j.cbpa.2017.03.017.

(34) Burke, M. D.; Schreiber, S. L. A Planning Strategy for Diversity-Oriented Synthesis. *Angewandte Chemie International Edition* **2004**, *43* (1), 46-58. DOI: https://doi.org/10.1002/anie.200300626.

(35) Gesmundo, N. J.; Sauvagnat, B.; Curran, P. J.; Richards, M. P.; Andrews, C. L.; Dandliker, P. J.; Cernak, T. Nanoscale synthesis and affinity ranking. *Nature* **2018**, *557* (7704), 228-232. DOI: 10.1038/s41586-018-0056-8.

(36) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015**, *347* (6217), 49-53. DOI: doi:10.1126/science.1259203.

(37) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559* (7714), 377-381. DOI: 10.1038/s41586-018-0307-8.

(38) Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **2018**, *359* (6374), 429-434. DOI: doi:10.1126/science.aap9112.

(39) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365* (6453), eaax1566. DOI: doi:10.1126/science.aax1566.

(40) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019**, *363* (6424), eaau5631. DOI: doi:10.1126/science.aau5631.

(41) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186-190.

(42) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571* (7765), 343-348. DOI: 10.1038/s41586-019-1384-z.

(43) McNally, A.; Prier, C. K.; MacMillan, D. W. C. Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science* **2011**, *334* (6059), 1114-1117. DOI: 10.1126/science.1213920.

(44) Troshin, K.; Hartwig, J. F. Snap deconvolution: An informatics approach to high-throughput discovery of catalytic reactions. *Science* **2017**, *357* (6347), 175-181. DOI: doi:10.1126/science.aan1568.

(45) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry* **2012**, *4* (2), 90-98. DOI: 10.1038/nchem.1243.

(46) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach To Enable Alignment of Druglike Properties. *ACS Chemical Neuroscience* **2010**, *1* (6), 435-449. DOI: 10.1021/cn100008c.

(47) Hill, A. P.; Young, R. J. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug Discovery Today* **2010**, *15* (15), 648-655. DOI: https://doi.org/10.1016/j.drudis.2010.05.016.

(48) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **2017**, *46* (D1), D1074-D1082. DOI: 10.1093/nar/gkx1037.

(49) Fu, M.-C.; Shang, R.; Cheng, W.-M.; Fu, Y. Boron-Catalyzed N-Alkylation of Amines using Carboxylic Acids. *Angewandte Chemie International Edition* **2015**, *54* (31), 9042-9046. DOI: https://doi.org/10.1002/anie.201503879.

(50) Alla, S. K.; Sadhu, P.; Punniyamurthy, T. Organocatalytic Syntheses of Benzoxazoles and Benzothiazoles using Aryl Iodide and Oxone via C–H Functionalization and C–O/S Bond Formation. *The Journal of Organic Chemistry* **2014**, *79* (16), 7502-7511. DOI: 10.1021/jo501216h.

(51) Huang, L.; Hackenberger, D.; Gooßen, L. J. Iridium-Catalyzed ortho-Arylation of Benzoic Acids with Arenediazonium Salts. *Angewandte Chemie International Edition* **2015**, *54* (43), 12607-12611. DOI: https://doi.org/10.1002/anie.201505769.

(52) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods* **2000**, *44* (1), 235-249. DOI: https://doi.org/10.1016/S1056-8719(00)00107-6.

(53) Mao, R.; Frey, A.; Balon, J.; Hu, X. Decarboxylative C(sp3)–N cross-coupling via synergetic photoredox and copper catalysis. *Nature Catalysis* **2018**, *1* (2), 120-126. DOI: 10.1038/s41929-017-0023-z.

(54) Lin, Y.; Zhang, Z.; Mahjour, B.; Wang, D.; Zhang, R.; Shim, E.; McGrath, A.; Shen, Y.; Brugger, N.; Turnbull, R.; Trice, S.; Jasty, S.; Cernak, T. Reinforcing the supply chain of umifenovir and other antiviral drugs with retrosynthetic software. *Nature Communications* **2021**, *12* (1), 7327. DOI: 10.1038/s41467-021-27547-3.

(55) Temporary Policy for Preparation of Certain Alcohol-Based Hand Sanitizer Products During the Public Health Emergency (COVID-19) Guidance for Industry. https://www.fda.gov/media/136289/download (Food and Drug Administration, 2020).

(56) Garbe, L.; Rau, R.; Toppe, T. Influence of perceived threat of Covid-19 and HEXACO personality traits on toilet paper stockpiling. *PLoS ONE* **2020**, *15*. DOI: 10.1371/journal.pone.0234232.

(57) Ledford, H. Dozens of coronavirus drugs are in development—what happens next? *Nature* **2020**, *581*. DOI: 10.1038/d41586-020-01367-9.

(58) Guy, R. K.; DiPaola, R. S.; Romanelli, F.; Dutch, R. E. Rapid repurposing of drugs for COVID-19. *Science* **2020**, *368*. DOI: 10.1126/science.abb9332.

(59) Fragkou, P. C. Review of trials currently testing treatment and prevention of COVID-19. *Clin. Microbiol. Infect.* **2020**, *26*. DOI: 10.1016/j.cmi.2020.05.019.

(60) COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Available at https://coronavirus.jhu.edu/map.html. Accessed Dec 9, 2020.

(61) Hardy, M. A. Treating a global health crisis with a dose of synthetic chemistry. *ACS Cent. Sci.* **2020**, *6*. DOI: 10.1021/acscentsci.0c00637.

(62) “Working to Supply Remdesivir for COVID-19.” (I. Gilead Sciences, 2020). Available at https://www.gilead.com/purpose/advancing-global-health/covid-19/working-to-supply-remdesivir-for-covid-19.

(63) Huang, D. Efficacy and safety of umifenovir for coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis. *J. Med. Virol* **2020**, *93*. DOI: 10.1002/jmv.26256.

(64) Szymkuc, S. Computer-generated “synthetic contingency” plans at times of logistics and supply problems: scenarios for hydroxychloroquine and remdesivir. *Chem. Sci.* **2020**, *11*. DOI: 10.1039/D0SC01799J.

(65) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166*. DOI: 10.1126/science.166.3902.178.

(66) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*. DOI: 10.1038/nature25978.

(67) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. *Angew. Chem. Int. Ed. Engl.* **2020**, *59*. DOI: 10.1002/anie.201912083.

(68) Molga, K.; Dittwald, P.; Grzybowski, B. A. Navigating around patented routes by preserving specific motifs along computer-planned retrosynthetic pathways. *Chemistry* **2019**, *5*. DOI: 10.1016/j.chempr.2018.12.004.

(69) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018**, *51*. DOI: 10.1021/acs.accounts.8b00087.

(70) Lee, A. A. Molecular transformer unifies reaction prediction, and retrosynthesis across pharma chemical space. *Chem. Commun.* **2019**, *55*. DOI: 10.1039/C9CC05122H.

(71) Available at http://covidroutes.cernaklab.com.

(72) Lian, N. Umifenovir treatment is not associated with improved outcomes in patients with coronavirus disease 2019: a retrospective study. *Clin. Microbiol. Infect.* **2020**, *26*. DOI: 10.1016/j.cmi.2020.04.026.

(73) Würtz, S.; Rakshit, S.; Neumann, J. J.; Dröge, T.; Glorius, F. Palladium-catalyzed oxidative cyclization of N-aryl enamines: from anilines to indoles. *Angew. Chem. Int. Ed. Engl.* **2008**, *47*. DOI: 10.1002/anie.200802482.

(74) Newhouse, T.; Baran, P. S. If C–H bonds could talk: selective C–H bond oxidation. *Angew. Chem. Int. Ed. Engl.* **2011**, *50*. DOI: 10.1002/anie.201006368.

(75) Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **2016**, *45*, 546.

(76) Davies, H. M. L.; Morton, D. Collective approach to advancing C–H functionalization. *ACS Cent. Sci.* **2017**, *3*. DOI: 10.1021/acscentsci.7b00329.

(77) Corey, E. J. The Logic of Chemical Synthesis (Nobel Foundation, [Nobelstiftelsen], 1991).

(78) Nakatsuka, S.; Asano, O.; Ueda, K.; Goto, T. Introduction of a hydroxy group onto 5- and 6-position of indole nucleus by Friedel-Crafts acylation and subsequent Baeyer-Villiger oxidation. *Heterocycles* **1987**, *26*. DOI: 10.3987/R-1987-06-1471.

(79) Balakin, K. V. Arbidol: a quarter-century after. Past, present and future of the original Russian antiviral. *Russ. Chem. Rev.* **2018**, *87*. DOI: 10.1070/RCR4791.

(80) Wright, Z. V. F.; Wu, N. C.; Kadam, R. U.; Wilson, I. A.; Wolan, D. W. Structure-based optimization and synthesis of antiviral drug Arbidol analogues with significantly improved affinity to influenza hemagglutinin. *Bioorg. Med. Chem.* **2017**, *27*. DOI: 10.1016/j.bmcl.2017.06.074.

(81) Moriyama, K.; Takemura, M.; Togo, H. Direct and selective benzylic oxidation of alkylarenes via C-H abstraction using alkali metal bromides. *Org. Lett.* **2012**, *14*. DOI: 10.1021/ol300853z.

(82) Kanda, Y. Two-phase synthesis of taxol. *J. Am. Chem. Soc.* **2020**, *142*. DOI: 10.1021/jacs.0c03592.

(83) Krumpolc, M.; Roček, J. Synthesis of stable chromium(V) complexes of tertiary hydroxy acids. *J. Am. Chem. Soc.* **1979**, *101*. DOI: 10.1021/ja00506a013.

(84) Ueno, R.; Ikeda, Y.; Shirakawa, E. tert-Butoxy-radical-promoted α-arylation of alkylamines with aryl halides. *Eur. J. Org. Chem.* **2017**, *2017*. DOI: 10.1002/ejoc.201700548.

(85) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *Journal of Medicinal Chemistry* **2016**, *59* (10), 4443-4458. DOI: 10.1021/acs.jmedchem.5b01409.

(86) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *Journal of Medicinal Chemistry* **2011**, *54* (10), 3451-3479. DOI: 10.1021/jm200187y.

(87) Fitzgerald, P. R.; Paegel, B. M. DNA-Encoded Chemistry: Drug Discovery from a Few Good Reactions. *Chemical Reviews* **2021**, *121* (12), 7155-7177. DOI: 10.1021/acs.chemrev.0c00789.

(88) Wang, Y.; Haight, I.; Gupta, R.; Vasudevan, A. What is in Our Kit? An Analysis of Building Blocks Used in Medicinal Chemistry Parallel Libraries. *Journal of Medicinal Chemistry* **2021**, *64* (23), 17115-17122. DOI: 10.1021/acs.jmedchem.1c01139.

(89) Goldberg, F. W.; Kettle, J. G.; Kogej, T.; Perry, M. W. D.; Tomkinson, N. P. Designing novel building blocks is an overlooked strategy to improve compound quality. *Drug Discovery Today* **2015**, *20* (1), 11-17. DOI: 10.1016/j.drudis.2014.09.023.

(90) Helal, C. J.; Bundesmann, M.; Hammond, S.; Holmstrom, M.; Klug-McLeod, J.; Lefker, B. A.; McLeod, D.; Subramanyam, C.; Zakaryants, O.; Sakata, S. Quick Building Blocks (QBB): An Innovative and Efficient Business Model To Speed Medicinal Chemistry Analog Synthesis. *ACS Medicinal Chemistry Letters* **2019**, *10* (8), 1104-1109. DOI: 10.1021/acsmedchemlett.9b00205.

(91) Grygorenko, O. O.; Volochnyuk, D. M.; Vashchenko, B. V. Emerging Building Blocks for Medicinal Chemistry: Recent Synthetic Advances. *European Journal of Organic Chemistry* **2021**, *2021* (47), 6478-6510. DOI: 10.1002/ejoc.202100857.

(92) Pennington, L. D.; Aquila, B. M.; Choi, Y.; Valiulin, R. A.; Muegge, I. Positional Analogue Scanning: An Effective Strategy for Multiparameter Optimization in Drug Design. *Journal of Medicinal Chemistry* **2020**, *63* (17), 8956-8976. DOI: 10.1021/acs.jmedchem.9b02092.

(93) Tomberg, A.; Boström, J. Can "easy" chemistry produce complex, diverse and novel molecules? *Drug Discovery Today* **2020**, *25* (12), 2174-2181. DOI: 10.1016/j.drudis.2020.09.027.

(94) Ruddigkeit, L.; Van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52* (11), 2864-2875. DOI: 10.1021/ci300415d.

(95) Walters, W. P. Virtual chemical libraries: miniperspective. *Journal of Medicinal Chemistry* **2018**, *62* (3), 1116-1124. DOI: 10.1021/acs.jmedchem.8b01048.

(96) Stein, R. M.; Kang, H. J.; McCorvy, J. D.; Glatfelter, G. C.; Jones, A. J.; Che, T.; Slocum, S.; Huang, X.-P.; Savych, O.; Moroz, Y. S.; Stauch, B.; Johansson, L. C.;

Cherezov, V.; Kenakin, T.; Irwin, J. J.; Shoichet, B. K.; Roth, B. L.; Dubocovich, M. L. Virtual discovery of melatonin receptor ligands to modulate circadian rhythms. *Nature* **2020**, *579* (7800), 609-614. DOI: 10.1038/s41586-020-2027-0.

(97) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, *566* (7743), 224-229. DOI: 10.1038/s41586-019-0917-9.

(98) Lyu, J.; Irwin, J. J.; Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology* **2023**, *19*, 712-718. DOI: 10.1038/s41589-022-01234-w.

(99) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient exploration of chemical space with docking and deep learning. *Journal of Chemical Theory and Computation* **2021**, *17* (11), 7106-7119. DOI: 10.1021/acs.jctc.1c00810.

(100) Zhang, R.; Mahjour, B.; Cernak, T. Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing. *ChemRxiv* **2022**. DOI: 10.26434/chemrxiv-2022-917k5.

(101) Shen, Y.; Mahjour, B.; Cernak, T. Development of copper-catalyzed deaminative esterification using high-throughput experimentation. *Communications Chemistry* **2022**, *5*. DOI: 10.1038/s42004-022-00698-0.

(102) Douthwaite, J. L.; Zhao, R.; Shim, E.; Mahjour, B.; Zimmerman, P.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp3–sp2 Carbon–Carbon Bonds. *Journal of the American Chemical Society* **2023**, *145* (20), 10930–10937. DOI: 10.1021/jacs.2c11563.

(103) McGrath, A.; Zhang, R.; Shafiq, K.; Cernak, T. Repurposing amine and carboxylic acid building blocks with an automatable esterification reaction. *Chemical Communications* **2023**, *59*, 1026-1029. DOI: 10.1039/D2CC05670D.

(104) Zhang, Z.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp3–sp3 Carbon–Carbon Bonds. *Angewandte Chemie International Edition* **2021**, *60* (52), 27293-27298. DOI: 10.1002/anie.202112454.

(105) Lyon, W. L.; MacMillan, D. W. Expedient Access to Underexplored Chemical Space: Deoxygenative C (sp3)–C (sp3) Cross-Coupling. *Journal of the American Chemical Society* **2023**, *145* (14), 7736-7742. DOI: 10.1021/jacs.3c01488.

(106) Laudadio, G.; Palkowitz, M. D.; El-Hayek Ewing, T.; Baran, P. S. Decarboxylative Cross-Coupling: A Radical Tool in Medicinal Chemistry. *ACS Medicinal Chemistry Letters* **2022**, *13* (9), 1413-1420. DOI: 10.1021/acsmedchemlett.2c00286.

(107) Darzi, E. R.; White, B. M.; Loventhal, L. K.; Zakharov, L. N.; Jasti, R. An operationally simple and mild oxidative homocoupling of aryl boronic esters to access conformationally constrained macrocycles. *Journal of the American Chemical Society* **2017**, *139* (8), 3106-3114. DOI: 10.1021/jacs.6b12658.

(108) Mahjour, B.; Shen, Y.; Cernak, T. Ultrahigh-throughput experimentation for information-rich chemical synthesis. *Accounts of Chemical Research* **2021**, *54* (10), 2337-2346. DOI: 10.1021/acs.accounts.1c00119.

(109) Wong, H.; Cernak, T. Reaction miniaturization in eco-friendly solvents. *Current Opinion in Green and Sustainable Chemistry* **2018**, *11*, 91-98. DOI: 10.1016/j.cogsc.2018.06.001.

(110) Cernak, T.; Gesmundo, N. J.; Dykstra, K.; Yu, Y.; Wu, Z.; Shi, Z.-C.; Vachal, P.; Sperbeck, D.; He, S.; Murphy, B. A.; Sonatore, L.; Williams, S.; Madeira, M.; Verras, A.; Reiter, M.; Lee, C. H.; Cuff, J.; Sherer, E. C.; Kuethe, J.; Goble, S.; Perrotto, N.; Pinto, S.; Shen, D.-M.; Nargund, R.; Balkovec, J.; DeVita, R. J.; Dreher, S. D. Microscale High-Throughput Experimentation as an Enabling Technology in Drug Discovery: Application in the Discovery of (Piperidinyl)pyridinyl-1H-benzimidazole Diacylglycerol Acyltransferase 1 Inhibitors. *Journal of Medicinal Chemistry* **2017**, *60* (9), 3594-3605. DOI: 10.1021/acs.jmedchem.6b01543.

(111) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Acc Chem Res* **2017**, *50* (12), 2976-2985. DOI: 10.1021/acs.accounts.7b00428.

(112) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med Chem Lett* **2017**, *8* (6), 601-607. DOI: 10.1021/acsmedchemlett.7b00165.

(113) Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio de Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Organic Process Research & Development* **2019**, *23* (6), 1213-1242. DOI: 10.1021/acs.oprd.9b00140.

(114) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **2021**, *1* (1), 23. DOI: 10.1038/s43586-021-00022-5.

(115) Campos Kevin, R.; Coleman Paul, J.; Alvarez Juan, C.; Dreher Spencer, D.; Garbaccio Robert, M.; Terrett Nicholas, K.; Tillyer Richard, D.; Truppo Matthew, D.; Parmee Emma, R. The importance of synthetic chemistry in the pharmaceutical industry. *Science* **2019**, *363* (6424), eaat0805. DOI: 10.1126/science.aat0805.

(116) Davies, I. W. The digitization of organic synthesis. *Nature* **2019**, *570* (7760), 175-181. DOI: 10.1038/s41586-019-1288-y.

(117) Cernak, T.; Dykstra, K. D.; Tyagarajan, S.; Vachal, P.; Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem Soc Rev* **2016**, *45* (3), 546-576. DOI: 10.1039/c5cs00628g.

(118) Newhouse, T.; Baran, P. S. If C–H Bonds Could Talk: Selective C–H Bond Oxidation. *Angewandte Chemie International Edition* **2011**, *50* (15), 3362-3374. DOI: 10.1002/anie.201006368.

(119) Davies, H. M.; Morton, D. C-H functionalization: collaborative methods to redefine chemical logic. *Angew Chem Int Ed Engl* **2014**, *53* (39), 10256-10258. DOI: 10.1002/anie.201406633.

(120) Prier, C. K.; Rankic, D. A.; MacMillan, D. W. C. Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chemical Reviews* **2013**, *113* (7), 5322-5363. DOI: 10.1021/cr300503r.

(121) Yan, M.; Kawamata, Y.; Baran, P. S. Synthetic Organic Electrochemical Methods Since 2000: On the Verge of a Renaissance. *Chemical Reviews* **2017**, *117* (21), 13230-13319. DOI: 10.1021/acs.chemrev.7b00397.

(122) Cho, B. T.; Kang, S. K. Direct and indirect reductive amination of aldehydes and ketones with solid acid-activated sodium borohydride under solvent-free conditions. *Tetrahedron* **2005**, *61* (24), 5725-5734. DOI: 10.1016/j.tet.2005.04.039.

(123) Puleo, T. R.; Bandar, J. S. Base-catalyzed aryl halide isomerization enables the 4-selective substitution of 3-bromopyridines. *Chemical Science* **2020**, *11* (38), 10517-10522. DOI: 10.1039/D0SC02689A.

(124) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **2021**, *1* (1), 1-23. DOI: 10.1038/s43586-021-00022-5.

(125) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting reaction conditions from limited data through active transfer learning. *Chemical science* **2022**, *13* (22), 6655-6668.

(126) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **2018**, *51* (5), 1281-1289. DOI: 10.1021/acs.accounts.8b00087.

(127) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-aided key step generation in alkaloid total synthesis. *Science* **2023**, *379* (6631), 453-457.

(128) Jiang, Y.; Yu, Y.; Kong, M.; Mei, Y.; Yuan, L.; Huang, Z.; Kuang, K.; Wang, Z.; Yao, H.; Zou, J. Artificial Intelligence for Retrosynthesis Prediction. *Engineering* **2022**. DOI: 10.1016/j.eng.2022.04.021.

(129) Lin, M. H.; Tu, Z.; Coley, C. W. Improving the performance of models for one-step retrosynthesis through re-ranking. *Journal of cheminformatics* **2022**, *14* (1), 1-13.

(130) Tu, Z.; Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling* **2022**, *62* (15), 3503-3513. DOI: 10.1021/acs.jcim.2c00321.

(131) Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. Evaluating and clustering retrosynthesis pathways with learned strategy. *Chemical science* **2021**, *12* (4), 1469-1478.

(132) Somnath, V. R.; Bunne, C.; Coley, C.; Krause, A.; Barzilay, R. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems* **2021**, *34*, 9405-9415.

(133) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **2017**, *3* (12), 1237-1245.

(134) Gesmundo, N.; Dykstra, K.; Douthwaite, J.; Mahjour, B.; Ferguson, R.; Dreher, S.; Sauvagnat, B.; Sauri, J.; Cernak, T. Miniaturization of Popular Reactions from the Medicinal Chemists' Toolbox for Ultrahigh-Throughput Experimentation. **2022**.

(135) Zhang, R.; Mahjour, B.; Cernak, T. Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing. **2022**.

(136) Mahjour, B.; Zhang, R.; Shen, Y.; McGrath, A.; Zhao, R.; Mohamed, O. G.; Lin, Y.; Zhang, Z.; Douthwaite, J. L.; Tripathi, A.; Cernak, T. Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery. *Nature Communications* **2023**, *1* (14), 3924. DOI: 10.1038/s41467-023-39531-0.

(137) Mahjour, B.; Hoffstadt, J.; Cernak, T. Designing Chemical Reaction Arrays using phactor and ChatGPT. **2023**.

(138) Chen, S.-J.; Golden, D. L.; Krska, S. W.; Stahl, S. S. Copper-catalyzed cross-coupling of benzylic C–H bonds and azoles with controlled N-site selectivity. *Journal of the American Chemical Society* **2021**, *143* (36), 14438-14444.

(139) Hu, H.; Chen, S.-J.; Mandal, M.; Pratik, S. M.; Buss, J. A.; Krska, S. W.; Cramer, C. J.; Stahl, S. S. Copper-catalysed benzylic C–H coupling with alcohols via radical relay enabled by redox buffering. *Nature catalysis* **2020**, *3* (4), 358-367.

(140) Uehling, M. R.; King, R. P.; Krska, S. W.; Cernak, T.; Buchwald, S. L. Pharmaceutical diversification via palladium oxidative addition complexes. *Science* **2019**, *363* (6425), 405. DOI: 10.1126/science.aac6153.

(141) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS. *Science* **2018**, *361*, eaar6236.

(142) Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W. Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical science* **2016**, *7* (4), 2604-2613.

(143) Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. Screening for Generality in Asymmetric Catalysis. *Nature* **2022**. DOI: 10.1038/s41586-022-05263-2.

(144) Zuranski, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. Predicting reaction yields via supervised learning. *Accounts of chemical research* **2021**, *54* (8), 1856-1865.

(145) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine learning yield prediction from NiCOlit, a small-size literature data set of nickel catalyzed C–O couplings. *Journal of the American Chemical Society* **2022**, *144* (32), 14722-14730.

(146) Stevens, J. M.; Li, J.; Simmons, E. M.; Wisniewski, S. R.; DiSomma, S.; Fraunhoffer, K. J.; Geng, P.; Hao, B.; Jackson, E. W. Advancing Base Metal Catalysis through Data Science: Insight and Predictive Models for Ni-Catalyzed Borylation through Supervised Machine Learning. *Organometallics* **2022**, *41* (14), 1847-1864.

(147) Lexa, K. W.; Belyk, K. M.; Henle, J.; Xiang, B.; Sheridan, R. P.; Denmark, S. E.; Ruck, R. T.; Sherer, E. C. Application of Machine Learning and Reaction Optimization for the Iterative Improvement of Enantioselectivity of Cinchona-Derived Phase Transfer Catalysts. *Organic Process Research & Development* **2021**, *26* (3), 670-682.

(148) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine learning may sometimes simply capture literature popularity trends: A case study of heterocyclic suzuki–miyaura coupling. *Journal of the American Chemical Society* **2022**, *144* (11), 4819-4827.

(149) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The open reaction database. *Journal of the American Chemical Society* **2021**, *143* (45), 18820-18826.

(150) Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M.; Kovač, K. Electronic lab notebooks: can they replace paper? *Journal of cheminformatics* **2017**, *9* (1), 1-15.

(151) Higgins, S. G.; Nogiwa-Valdez, A. A.; Stevens, M. M. Considerations for implementing electronic laboratory notebooks in an academic research environment. *Nature Protocols* **2022**, *17* (2), 179-189.

(152) Tremouilhac, P.; Nguyen, A.; Huang, Y.-C.; Kotov, S.; Lütjohann, D. S.; Hübsch, F.; Jung, N.; Bräse, S. Chemotion ELN: an Open Source electronic lab notebook for chemists in academia. *Journal of Cheminformatics* **2017**, *9* (1), 54. DOI: 10.1186/s13321-017-0240-0.

(153) Barillari, C.; Ottoz, D. S. M.; Fuentes-Serna, J. M.; Ramakrishnan, C.; Rinn, B.; Rudolf, F. openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics* **2016**, *32* (4), 638-640. DOI: 10.1093/bioinformatics/btv606.

(154) Cole, J. M. The chemistry of errors. *Nature Chemistry* **2022**, *14* (9), 973-975. DOI: 10.1038/s41557-022-01028-6.

(155) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *Journal of Chemical Information and Modeling* **2012**, *52* (7), 1745-1756. DOI: 10.1021/ci300116p.

(156) Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The case for data science in experimental chemistry: examples and recommendations. *Nature Reviews Chemistry* **2022**, *6* (5), 357-370. DOI: 10.1038/s41570-022-00382-w.

(157) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An orchestration software to democratize autonomous discovery. *PLOS ONE* **2020**, *15* (4), e0229862. DOI: 10.1371/journal.pone.0229862.

(158) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A. Data-science driven autonomous process optimization. *Communications Chemistry* **2021**, *4* (1), 1-12.

(159) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. Olympus: a benchmarking framework for noisy optimization and experiment planning. *Machine Learning: Science and Technology* **2021**, *2* (3), 035021.

(160) Shiri, P.; Lai, V.; Zepel, T.; Griffin, D.; Reifman, J.; Clark, S.; Grunert, S.; Yunker, L. P. E.; Steiner, S.; Situ, H. Automated solubility screening platform using computer vision. *Iscience* **2021**, *24* (3), 102176.

(161) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590* (7844), 89-96.

(162) Daponte, J. A.; Guo, Y.; Ruck, R. T.; Hein, J. E. Using an automated monitoring platform for investigations of biphasic reactions. *ACS Catalysis* **2019**, *9* (12), 11484-11491.

(163) *Katalyst D2D*. https://www.acdlabs.com/products/spectrus-platform/katalyst-d2d/ (accessed.

(164) Labs, U. *LEA*. https://www.unchainedlabs.com/unchained-labs-makes-automation-workflows-cinch-releases-lea-9/ (accessed.

(165) Lu, J.; Paci, I.; Leitch, D. C. A broadly applicable quantitative relative reactivity model for nucleophilic aromatic substitution (SN Ar) using simple descriptors. *Chemical Science* **2022**, *13* (43), 12681-12695.

(166) Lu, J.; Donnecke, S.; Paci, I.; Leitch, D. C. A reactivity model for oxidative addition to palladium enables quantitative predictions for catalytic cross-coupling reactions. *Chemical Science* **2022**, *13* (12), 3477-3488.

(167) Allen, C. L.; Leitch, D. C.; Anson, M. S.; Zajac, M. A. The power and accessibility of high-throughput methods for catalysis research. *Nature Catalysis* **2019**, *2* (1), 2-4.

(168) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A. Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **2022**, *378* (6618), 399-405.

(169) Molga, K.; Szymkuć, S.; Gołębiowska, P.; Popik, O.; Dittwald, P.; Moskal, M.; Roszak, R.; Mlynarski, J.; Grzybowski, B. A. A computer algorithm to discover iterative sequences of organic reactions. *Nature Synthesis* **2022**, *1* (1), 49-58.

(170) Molga, K.; Szymkuc, S.; Grzybowski, B. A. Chemist ex machina: advanced synthesis planning by computers. *Accounts of chemical research* **2021**, *54* (5), 1094-1106.

(171) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. *Angewandte Chemie International Edition* **2020**, *59* (2), 725-730.

(172) Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic discovery of tactical combinations for advanced organic syntheses. *Chem* **2020**, *6* (1), 280-293.

(173) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of

atoms across both simple and complex chemical reactions. *Nature communications* **2019**, *10* (1), 1434.

(174) Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wołos, A.; Klucznik, T. Chematica: a story of computer code that started to think like a chemist. *Chem* **2018**, *4* (3), 390-398.

(175) Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties. *Accounts of Chemical Research* **2021**, *54* (4), 827-836.

(176) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *Journal of the American Chemical Society* **2022**, *144* (43), 19999-20007.

(177) Reker, D.; Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Reports Physical Science* **2020**, *1* (11), 100247.

(178) Hammer, A. J. S.; Leonov, A. I.; Bell, N. L.; Cronin, L. Chemputation and the Standardization of Chemical Informatics. *JACS Au* **2021**, *1* (10), 1572-1587. DOI: 10.1021/jacsau.1c00303.

(179) Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S. A comprehensive discovery platform for organophosphorus ligands for catalysis. *Journal of the American Chemical Society* **2022**, *144* (3), 1205-1217.

(180) Shen, Y.; Mahjour, B.; Cernak, T. Development of copper-catalyzed deaminative esterification using high-throughput experimentation. *Communications Chemistry* **2022**, *5* (1), 1-8.

(181) Zhang, Z.; Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp3–sp3 Carbon–Carbon Bonds. *Angewandte Chemie International Edition* **2021**, *60* (52), 27293-27298, https://doi.org/10.1002/anie.202112454. DOI: https://doi.org/10.1002/anie.202112454.

(182) McGrath, A.; Zhang, R.; Shafiq, K.; Cernak, T. Repurposing amine and carboxylic acid building blocks with an automatable esterification reaction. *Chemical Communications* **2023**.

(183) Wang, C.; Rakshit, S.; Glorius, F. Palladium-Catalyzed Intermolecular Decarboxylative Coupling of 2-Phenylbenzoic Acids with Alkynes via C−H and C−C Bond Activation. *Journal of the American Chemical Society* **2010**, *132* (40), 14006-14008. DOI: 10.1021/ja106130r.

(184) Rossetti, G. G.; Ossorio, M. A.; Rempel, S.; Kratzel, A.; Dionellis, V. S.; Barriot, S.; Tropia, L.; Gorgulla, C.; Arthanari, H.; Thiel, V.; Mohr, P.; Gamboni, R.; Halazonetis, T. D. Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits. *Scientific Reports* **2022**, *12* (1), 2505. DOI: 10.1038/s41598-022-06306-4.

(185) Malla, T. R.; Tumber, A.; John, T.; Brewitz, L.; Strain-Damerell, C.; Owen, C. D.; Lukacik, P.; Chan, H. T. H.; Maheswaran, P.; Salah, E. Mass spectrometry reveals potential of β-lactams as SARS-CoV-2 M pro inhibitors. *Chemical Communications* **2021**, *57* (12), 1430-1433.

(186) Zhang, J.-H.; Chung, T. D. Y.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *Journal of biomolecular screening* **1999**, *4* (2), 67-73.

(187) Lu, P.; Sanchez, C.; Cornella, J.; Larrosa, I. Silver-catalyzed protodecarboxylation of heteroaromatic carboxylic acids. *Organic letters* **2009**, *11* (24), 5710-5713.

(188) Gooßen, L. J.; Linder, C.; Rodríguez, N.; Lange, P. P.; Fromm, A. Silver-catalysed protodecarboxylation of carboxylic acids. *Chemical communications* **2009**, (46), 7173-7175.

(189) Dow, N. W.; Pedersen, P. S.; Chen, T. Q.; Blakemore, D. C.; Dechert-Schmitt, A.-M.; Knauber, T.; MacMillan, D. W. C. Decarboxylative borylation and cross-coupling of (hetero) aryl acids enabled by copper charge transfer catalysis. *Journal of the American Chemical Society* **2022**, *144* (14), 6163-6172.

(190) Xu, P.; López-Rojas, P.; Ritter, T. Radical decarboxylative carbometalation of benzoic acids: a solution to aromatic decarboxylative fluorination. *Journal of the American Chemical Society* **2021**, *143* (14), 5349-5354.

(191) Witzel, S.; Hoffmann, M.; Rudolph, M.; Kerscher, M.; Comba, P.; Dreuw, A.; Hashmi, A. S. K. Excitation of aryl cations as the key to catalyst-free radical arylations. *Cell Reports Physical Science* **2021**, *2* (2), 100325. DOI: 10.1016/j.xcrp.2021.100325.

(192) Moore, G. E. Cramming More Components Onto Integrated Circuits. *Proceedings of the IEEE* **1998**, *86* (1), 82-85. DOI: 10.1109/JPROC.1998.658762.

(193) Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Acc. Chem. Res.* **2017**, *50*, 2976.

(194) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8*, 601.

(195) Shultz, C. S.; Krska, S. W. Unlocking the potential of asymmetric hydrogenation at Merck. *Acc. Chem. Res.* **2007**, *40*, 1320.

(196) Allen, C. L.; Leitch, D. C.; Anson, M. S.; Zajac, M. A. The power and accessibility of high-throughput methods for catalysis research. *Nat. Catalysis* **2019**, *2*, 2.

(197) Wong, H.; Cernak, T. Reaction miniaturization in eco-friendly solvents. *Curr. Opin. Green Sus. Chem.* **2018**, *11*, 91.

(198) Cernak, T. Synthesis in the Chemical Space Age. *Chem.* **2016**, *1*, 6.

(199) Yu, Y.; Wu, Z.; Shi, Z.-C.; He, S.; Lai, Z.; Cernak, T. A.; Vachal, P.; Liu, M.; Liu, J.; Hong, Q.; Jian, T.; Guiadeen, D.; Krikorian, A.; Sperbeck, D. M.; Verras, A.; Sonatore, L. M.; Murphy, B. A.; Wiltsie, J.; Chung, C. C.; Gorski, J. N.; Liu, J.; Xiao, J.; Wolff, M.; Tong, S. X.; Madeira, M.; Karanam, B. V.; Shen, D.-M.; Balkovec, J. M.; De Vita, R. J.; Pinto, S.; Nargund, R. P. Accelerating the discovery of DGAT1 inhibitors through the application of parallel medicinal chemistry (PMC). *Bioorganic & Medicinal Chemistry Letters* **2019**, *29* (11), 1380-1385. DOI: 10.1016/j.bmcl.2019.03.039.

(200) He, S.; Lai, Z.; Hong, Q.; Shang, J.; Reibarkh, M.; Kuethe, J. T.; Liu, J.; Guiadeen, D.; Krikorian, A. D.; Cernak, T. A.; Dykstra, K. D.; Sperbeck, D. M.; Wu, Z.; Yu, Y.; Yang, G. X.; Jian, T.; Verras, A.; Sonatore, L. M.; Wiltsie, J.; Chung, C. C.; Murphy, B. A.; Gorski, J. N.; Liu, J.; Xiao, J.; Wolff, M.; Tong, S. X.; Madeira, M.; Karanam, B. V.; Shen, D. M.; Balkovec, J. M.; Pinto, S.; Nargund, R. P.; DeVita, R. J. Benzimidazole-based DGAT1 inhibitors with a [3.1.0] bicyclohexane carboxylic acid moiety. *Bioorg. Med. Chem. Lett.* **2019**, *29*, 1182.

(201) He, S.; Hong, Q.; Lai, Z.; Yang, D. X.; Ting, P. C.; Kuethe, J. T.; Cernak, T. A.; Dykstra, K. D.; Sperbeck, D. M.; Wu, Z.; Yu, Y.; Yang, G. X.; Jian, T.; Liu, J.; Guiadeen, D.; Krikorian, A. D.; Sonatore, L. M.; Wiltsie, J.; Liu, J.; Gorski, J. N.; Chung, C. C.; Gibson, J. T.; Lisnock, J.; Xiao, J.; Wolff, M.; Tong, S. X.; Madeira, M.; Karanam, B. V.; Shen, D. M.; Balkovec, J. M.; Pinto, S.; Nargund, R. P.; DeVita, R. J. Discovery of a Potent and Selective DGAT1 Inhibitor with a Piperidinyl-oxy-cyclohexanecarboxylic Acid Moiety. *ACS Med. Chem. Lett.* **2014**, *5*, 1082.

(202) Cho, J. Y.; Tse, M. K.; Holmes, D.; Maleczka, R. E.; Smith, M. R. Remarkably selective iridium catalysts for the elaboration of aromatic C-H bonds. *Science* **2002**, *295*, 305.

(203) Preshlock, S. M.; Ghaffari, B.; Maligres, P. E.; Krska, S. W.; Maleczka, R. E.; Smith, M. R. High-throughput optimization of Ir-catalyzed C-H borylation: a tutorial for practical applications. *J. Am. Chem. Soc.* **2013**, *135*, 7572.

(204) Ishiyama, T.; Takagi, J.; Ishida, K.; Miyaura, N.; Anastasi, N. R.; Hartwig, J. F. Mild Iridium-Catalyzed Borylation of Arenes. High Turnover Numbers, Room Temperature Reactions, and Isolation of a Potential Intermediate. *J. Am. Chem. Soc.* **2002**, *124*, 390.

(205) Liu, K.; Kurukulasuriya, R.; Dykstra, K.; DiMichelle, L.; Liu, J.; Vachal, P.; Ogawa, A.; DeVita, R. J.; Shen, D. M.; Tan, Q.; Chen, Y.; Gauthier, D.; Verras, A.; Crespo, A.; Zamlynny, B.; Madwed, J.; Hoek, M.; Bateman, T.; Yang, Y. F.; Houk, K. N.; Krska, S.;

Cernak, T. Development of indazole mineralocorticoid receptor antagonists and investigation into their selective late-stage functionalization. *Bioorg. Med. Chem. Lett.* **2019**, *29*, 1854.

(206) Takatsy, G. Uj modszer sorozatos higitasok gyors es pontos elvegzesere [A rapid and accurate method for serial dilutions]. *Kiserl. Orvostud.* **1950**, *5*, 393.

(207) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59*, 4443.

(208) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016**, *59*, 4385.

(209) Keserü, G. M.; Soos, T.; Kappe, C. O. Anthropogenic reaction parameters--the missing link between chemical intuition and the available chemical space. *Chem. Soc. Rev.* **2014**, *43*, 5387.

(210) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.

(211) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54*, 3451.

(212) Buitrago Santanilla, A.; Christensen, M.; Campeau, L. C.; Davies, I. W.; Dreher, S. D. P2Et Phosphazene: A Mild, Functional Group Tolerant Base for Soluble, Room Temperature Pd-Catalyzed C-N, C-O, and C-C Cross-Coupling Reactions. *Org. Lett.* **2015**, *17*, 3370.

(213) Uehling, M. R.; King, R. P.; Krska, S. W.; Cernak, T.; Buchwald, S. L. Pharmaceutical diversification via palladium oxidative addition complexes. *Science* **2019**, *363*, 405.

(214) Vinogradova, E. V.; Zhang, C.; Spokoyny, A. M.; Pentelute, B. L.; Buchwald, S. L. Organometallic palladium reagents for cysteine bioconjugation. *Nature* **2015**, *526*, 687.

(215) Annis, D. A.; Nickbarg, E.; Yang, X.; Ziebell, M. R.; Whitehurst, C. E. Affinity selection-mass spectrometry screening techniques for small molecule drug discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 518.

(216) Gabriel, J.; Höfner, G.; Wanner, K. T. A Library Screening Strategy Combining the Concepts of MS Binding Assays and Affinity Selection Mass Spectrometry. *Front. Chem.* **2019**, *7*, 665.

(217) Prudent, R.; Annis, D. A.; Dandliker, P. J.; Ortholand, J. Y.; Roche, D. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nat. Rev. Chem.* **2021**, *5*, 62.

(218) Mahjour, B.; Cernak, T. PhactorTM—a High Throughput Experimentation Management System. *ChemRxiv* **2020**.

(219) Mahjour, B.; Shen, Y.; Liu, W.; Cernak, T. A map of the amine-carboxylic acid coupling system. *Nature* **2020**, *580*, 71.

(220) Wleklinski, M.; Loren, B. P.; Ferreira, C. R.; Jaman, Z.; Avramova, L.; Sobreira, T. J. P.; Thompson, D. H.; Cooks, R. G. High throughput reaction screening using desorption electrospray ionization mass spectrometry. *Chem. Sci.* **2018**, *9*, 1647.

(221) Sutanto, F.; Shaabani, S.; Neochoritis, C. G.; Zarganes-Tzitzikas, T.; Patil, P.; Ghonchepour, E.; Dömling, A. Multicomponent reaction-derived covalent inhibitor space. *Sci. Adv.* **2021**, *7*, eabd9307.

(222) Osipyan, A.; Shaabani, S.; Warmerdam, R.; Shishkina, S. V.; Boltz, H.; Dömling, A. Automated, Accelerated Nanoscale Synthesis of Iminopyrrolidines. *Angew. Chem., Int. Ed.* **2020**, *59*, 12423.

(223) Wang, Y.; Shaabani, S.; Ahmadianmoghaddam, M.; Gao, L.; Xu, R.; Kurpiewska, K.; Kalinowska-Tluscik, J.; Olechno, J.; Ellson, R.; Kossenjans, M.; Helan, V.; Groves, M.; Dömling, A. Acoustic Droplet Ejection Enabled Automated Reaction Scouting. *ACS Cent. Sci.* **2019**, *5*, 451.

(224) Sobreira, T. J. P.; Avramova, L.; Szilagyi, B.; Logsdon, D. L.; Loren, B. P.; Jaman, Z.; Hilger, R. T.; Hosler, R. S.; Ferreira, C. R.; Koswara, A.; Thompson, D. H.; Cooks, R. G.; Nagy, Z. K. High-throughput screening of organic reactions in microdroplets using desorption electrospray ionization mass spectrometry (DESI-MS): hardware and software implementation. *Anal. Methods* **2020**, *12*, 3654.

(225) Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. The evolution of data-driven modeling in organic chemistry. *ACS central science* **2021**, *7* (10), 1622-1637.

(226) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *Journal of the American Chemical Society* **2018**, *140* (15), 5004-5008.

(227) Gao, W.; Raghavan, P.; Coley, C. W. Autonomous platforms for data-driven organic synthesis. *Nature Communications* **2022**, *13* (1), 1-4.

(228) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive chemistry: Machine learning for reaction deployment, reaction development, and reaction discovery. *Chemical Science* **2022**, *14* (2), 226-244.

(229) Luo, R.; Sun, L.; Xia, Y.; Qin, T.; Zhang, S.; Poon, H.; Liu, T.-Y. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* **2022**, *23* (6).

(230) Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 is a deep unsupervised language model for protein design. *Nature communications* **2022**, *13* (1), 4348.

(231) https://openai.com/.

(232) Zhang, C.; Trudell, M. L. Palladium-bisimidazol-2-ylidene complexes as catalysts for general and efficient Suzuki cross-coupling reactions of aryl chlorides with arylboronic acids. *Tetrahedron Letters* **2000**, *41* (5), 595-598.

(233) Choy, P. Y.; Yuen, O. Y.; Leung, M. P.; Chow, W. K.; Kwong, F. Y. A Highly Efficient Monophosphine Ligand for Parts per Million Levels Pd-Catalyzed Suzuki–Miyaura Coupling of (Hetero) Aryl Chlorides. *European Journal of Organic Chemistry* **2020**, *2020* (19), 2846-2853.

(234) Thomas, A. A.; Denmark, S. E. Pre-transmetalation intermediates in the Suzuki-Miyaura reaction revealed: The missing link. *Science* **2016**, *352* (6283), 329-332.

(235) Zhao, X.-Y.; Zhou, Q.; Lu, J.-M. Synthesis and characterization of N-heterocyclic carbene-palladium (ii) chlorides-1-methylindazole and-1-methylpyrazole complexes and their catalytic activity toward C–N coupling of aryl chlorides. *RSC advances* **2016**, *6* (29), 24484-24490.

(236) King, A. K.; Brar, A.; Findlater, M. A tertiary phosphine oxide ligand-based recyclable system for the Suzuki–Miyaura and Negishi reactions: evidence for pseudo-homogeneous catalysis. *Catalysis Science & Technology* **2023**, *13* (2), 301-304.

(237) Prima, D. O.; Madiyeva, M.; Burykina, J. V.; Minyaev, M. E.; Boiko, D. A.; Ananikov, V. P. Evidence for “cocktail”-type catalysis in Buchwald–Hartwig reaction. A mechanistic study. *Catalysis Science & Technology* **2021**, *11* (21), 7171-7188.

(238) Ondar, E. E.; Burykina, J. V.; Ananikov, V. P. Evidence for the “cocktail” nature of platinum-catalyzed alkyne and alkene hydrosilylation reactions. *Catalysis Science & Technology* **2022**, *12* (4), 1173-1186.

(239) Fors, B. P.; Buchwald, S. L. A multiligand based Pd catalyst for C− N cross-coupling reactions. *Journal of the American Chemical Society* **2010**, *132* (45), 15914-15917.

(240) Shen, Y.; Mahjour, B.; Cernak, T. Development of copper-catalyzed deaminative esterification using high-throughput experimentation. *Communications Chemistry* **2022**, *5* (1), 83. DOI: 10.1038/s42004-022-00698-0.

(241) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28* (1), 31-36. DOI: 10.1021/ci00057a005.

(242) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, *5* (2), 107-113. DOI: 10.1021/c160017a018.

(243) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50* (5), 742-754. DOI: 10.1021/ci100050t.

(244) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **2020**, *6* (6), 1379-1390.

(245) ChemAxon. Reaction fingerprints. https://docs.chemaxon.com/display/docs/reaction-fingerprint-rf.md **2022**.

(246) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* **2002**, *42* (6), 1273-1280. DOI: 10.1021/ci010132r. From NLM Medline.

(247) Stumpfe, D.; Bajorath, J. r. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55* (7), 2932-2942. DOI: 10.1021/jm201706b.

(248) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews* **2020**, *49* (17), 6154-6168.

(249) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2021**, *2* (1), 015016.

(250) Probst, D.; Schwaller, P.; Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery* **2022**, *1* (2), 91-97.

(251) Yu, C.-G.; Matsuo, Y. Nickel-catalyzed deaminative acylation of activated aliphatic amines with aromatic amides via C–N bond activation. *Organic letters* **2020**, *22* (3), 950-955.

(252) Halpern, J.; Kemp, A. L. W. The Decarbonylation of Formic Acid by Ruthenium (II) Chloride1. *Journal of the American Chemical Society* **1966**, *88* (22), 5147-5150.

(253) Wang, D.; Zhu, N.; Chen, P.; Lin, Z.; Liu, G. Enantioselective Decarboxylative Cyanation Employing Cooperative Photoredox Catalysis and Copper Catalysis. *Journal of the American Chemical Society* **2017**, *139* (44), 15632-15635. DOI: 10.1021/jacs.7b09802.

(254) Proctor, R. S. J.; Davis, H. J.; Phipps, R. J. Catalytic enantioselective Minisci-type addition to heteroarenes. *Science* **2018**, *360* (6387), 419-422. DOI: doi:10.1126/science.aar6376.

(255) Fawcett, A.; Pradeilles, J.; Wang, Y.; Mutsuga, T.; Myers, E. L.; Aggarwal, V. K. Photoinduced decarboxylative borylation of carboxylic acids. *Science* **2017**, *357* (6348), 283-286. DOI: doi:10.1126/science.aan3679.

(256) Guo, J.-Y.; Zhang, Z.-Y.; Guan, T.; Mao, L.-W.; Ban, Q.; Zhao, K.; Loh, T.-P. Photoredox-catalyzed stereoselective alkylation of enamides with N-hydroxyphthalimide esters via decarboxylative cross-coupling reactions. *Chemical Science* **2019**, *10* (38), 8792-8798, 10.1039/C9SC03070K. DOI: 10.1039/C9SC03070K.

(257) Till, N. A.; Smith, R. T.; MacMillan, D. W. C. Decarboxylative Hydroalkylation of Alkynes. *Journal of the American Chemical Society* **2018**, *140* (17), 5701-5705. DOI: 10.1021/jacs.8b02834.

(258) Konev, M. O.; Jarvo, E. R. Decarboxylative Alkyl–Alkyl Cross-Coupling Reactions. *Angewandte Chemie International Edition* **2016**, *55* (38), 11340-11342. DOI: https://doi.org/10.1002/anie.201605593.

(259) Sandfort, F.; O'Neill, M. J.; Cornella, J.; Wimmer, L.; Baran, P. S. Alkyl−(Hetero)Aryl Bond Formation via Decarboxylative Cross-Coupling: A Systematic Analysis. *Angewandte Chemie International Edition* **2017**, *56* (12), 3319-3323. DOI: https://doi.org/10.1002/anie.201612314.

(260) Smith, J. M.; Qin, T.; Merchant, R. R.; Edwards, J. T.; Malins, L. R.; Liu, Z.; Che, G.; Shen, Z.; Shaw, S. A.; Eastgate, M. D.; Baran, P. S. Decarboxylative Alkynylation. *Angewandte Chemie International Edition* **2017**, *56* (39), 11906-11910. DOI: https://doi.org/10.1002/anie.201705107.

(261) Wang, J.; Qin, T.; Chen, T.-G.; Wimmer, L.; Edwards, J. T.; Cornella, J.; Vokits, B.; Shaw, S. A.; Baran, P. S. Nickel-Catalyzed Cross-Coupling of Redox-Active Esters with Boronic Acids. *Angewandte Chemie International Edition* **2016**, *55* (33), 9676-9679. DOI: https://doi.org/10.1002/anie.201605463.

(262) Huihui, K. M. M.; Caputo, J. A.; Melchor, Z.; Olivares, A. M.; Spiewak, A. M.; Johnson, K. A.; DiBenedetto, T. A.; Kim, S.; Ackerman, L. K. G.; Weix, D. J. Decarboxylative Cross-Electrophile Coupling of N-Hydroxyphthalimide Esters with Aryl Iodides. *Journal of the American Chemical Society* **2016**, *138* (15), 5016-5019. DOI: 10.1021/jacs.6b01533.

(263) Huang, L.; Olivares, A. M.; Weix, D. J. Reductive Decarboxylative Alkynylation of N-Hydroxyphthalimide Esters with Bromoalkynes. *Angewandte Chemie International Edition* **2017**, *56* (39), 11901-11905. DOI: https://doi.org/10.1002/anie.201706781.

(264) Pulikottil, F. T.; Pilli, R.; Suku, R. V.; Rasappan, R. Nickel-Catalyzed Cross-Coupling of Alkyl Carboxylic Acid Derivatives with Pyridinium Salts via C–N Bond Cleavage. *Organic Letters* **2020**, *22* (8), 2902-2907. DOI: 10.1021/acs.orglett.0c00554.

(265) Wang, J.; Hoerrner, M. E.; Watson, M. P.; Weix, D. J. Nickel-Catalyzed Synthesis of Dialkyl Ketones from the Coupling of N-Alkyl Pyridinium Salts with Activated Carboxylic Acids. *Angewandte Chemie International Edition* **2020**, *59* (32), 13484-13489. DOI: https://doi.org/10.1002/anie.202002271.

(266) Moragas, T.; Gaydou, M.; Martin, R. Nickel-Catalyzed Carboxylation of Benzylic C−N Bonds with CO2. *Angewandte Chemie International Edition* **2016**, *55* (16), 5053-5057. DOI: https://doi.org/10.1002/anie.201600697.

(267) Das, S.; Li, Y.; Junge, K.; Beller, M. Synthesis of ethers from esters via Fe-catalyzed hydrosilylation. *Chemical Communications* **2012**, *48* (87), 10742-10744, 10.1039/C2CC32142D. DOI: 10.1039/C2CC32142D.

(268) Gevorgyan, V.; Rubin, M.; Liu, J.-X.; Yamamoto, Y. A Direct Reduction of Aliphatic Aldehyde, Acyl Chloride, Ester, and Carboxylic Functions into a Methyl Group. *The Journal of Organic Chemistry* **2001**, *66* (5), 1672-1675. DOI: 10.1021/jo001258a.

(269) Yato, M.; Homma, K.; Ishida, A. Reduction of carboxylic esters to ethers with triethyl silane in the combined use of titanium tetrachloride and trimethylsilyl trifluoromethanesulfonate. *Tetrahedron* **2001**, *57* (25), 5353-5359. DOI: https://doi.org/10.1016/S0040-4020(01)00420-3.

(270) Todtz, S. R.; Schneider, C. W.; Malakar, T.; Anderson, C.; Koska, H.; Zimmerman, P. M.; Devery, J. J., III. Controlling Catalyst Behavior in Lewis Acid-Catalyzed Carbonyl-Olefin Metathesis. *Journal of the American Chemical Society* **2023**, *145* (24), 13069-13080. DOI: 10.1021/jacs.3c01508.

(271) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1* (4), 045024. DOI: 10.1088/2632-2153/aba947.

(272) Le, T.; Winter, R.; Noé, F.; Clevert, D.-A. Neuraldecipher–reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chemical Science* **2020**, *11* (38), 10378-10389.

(273) Kekule, A. *Sur la constitution des substances aromatiques*; Bulletin de la Société Chimique de Paris, 1865.

(274) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28* (1), 31-36.

(275) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns.

(276) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminform.* **2013**, *5* (1), 7. DOI: 10.1186/1758-2946-5-7.

(277) O'Boyle, N.; Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. **2018**.

(278) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25* (2), 64-73. DOI: 10.1021/ci00046a002.

(279) Boobier, S.; Hose, D. R. J.; John Blacker, A.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications* **2020**, *11* (1). DOI: 10.1038/s41467-020-19594-z.

(280) Pereira, F.; Aires-de-Sousa, J. Machine learning for the prediction of molecular dipole moments obtained by density functional theory. *J. Cheminform.* **2018**, *10* (1), 43. DOI: 10.1186/s13321-018-0296-5.

(281) Kumbar, M. Musical chemistry: Integrating chemistry and music. *Journal of chemical education* **2007**, *84* (12), 1933.

(282) Tóth, V.; Parkkonen, L. Autoencoding sensory substitution. **2019**. DOI: 10.13140/RG.2.2.10576.87048.

(283) Baum, Z. J.; Yu, X.; Ayala, P. Y.; Zhao, Y.; Watkins, S. P.; Zhou, Q. Artificial intelligence in chemistry: current trends and future directions. *Journal of Chemical Information and Modeling* **2021**, *61* (7), 3197-3212.

(284) Briot, J.-P.; Hadjeres, G.; Pachet, F.-D. Deep Learning Techniques for Music Generation. *Computational Synthesis and Creative Systems* **2020**. DOI: 10.1007/978-3-319-70163-9.

(285) Hermann, T.; Hunt, A.; Neuhoff, J. G. *The sonification handbook*; Logos Verlag Berlin, 2011.

(286) Muller-Eberstein, M.; van Noord, N. Translating Visual Art Into Music. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* **2019**. DOI: 10.1109/iccvw.2019.00378.

(287) Making music from spider webs.

(288) Garrido, N.; Pitto-Barry, A.; Soldevila-Barreda, J. J.; Lupan, A.; Boyes, L. C.; Martin, W. H. C.; Barry, N. P. E. The Sound of Chemistry: Translating Infrared Wavenumbers into Musical Notes. *Journal of Chemical Education* **2020**, *97* (3), 703-709. DOI: 10.1021/acs.jchemed.9b00775.

(289) A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. DOI: 10.1021/acsnano.9b02180.s001.

(290) otnemrasordep. GitHub - otnemrasordep/sonification-bela: Final Assignment for ECS7012P: Music and Audio Programming.

(291) Milazzo, M.; Buehler, M. J. Designing and fabricating materials from fire using sonification and deep learning. *Iscience* **2021**, *24* (8), 102873.

(292) Tay, N. W.; Liu, F.; Wang, C.; Zhang, H.; Zhang, P.; Chen, Y. Z. Protein music of enhanced musicality by music style guided exploration of diverse amino acid properties. *Heliyon* **2021**, *7* (9), e07933.

(293) Su, I.; Hattwick, I.; Southworth, C.; Ziporyn, E.; Bisshop, A.; Mühlethaler, R.; Saraceno, T.; Buehler, M. J. Interactive exploration of a hierarchical spider web structure with sound. *Journal on Multimodal User Interfaces* **2022**, 1-15.

(294) Meijer, D. K. F.; Jerman, I.; Melkikh, A. V.; Sbitnev, V. I. Biophysics of consciousness: A scale-invariant acoustic information code of a superfluid quantum space guides the mental attribute of the universe. *Rhythmic Oscillations in Proteins to Human Cognition* **2021**, 213-361.

(295) Mitchell, T. J.; Jones, A. J.; O'Connor, M. B.; Wonnacott, M. D.; Glowacki, D. R.; Hyde, J. Towards molecular musical instruments: interactive sonifications of 17-alanine, graphene and carbon nanotubes. 2020, pp 214-221.

(296) Groß, M. Die Musik der Proteine. *Nachrichten aus der Chemie* **2019**, *67* (10), 98-98.

(297) Franjou, S. L.; Milazzo, M.; Yu, C.-H.; Buehler, M. J. Sounds interesting: Can sonification help us design new proteins? *Expert Review of Proteomics* **2019**, *16* (11-12), 875-879.

(298) Borgonovo, A.; Haus, G. Sound synthesis by means of two-variable functions: experimental criteria and results. *Computer Music Journal* **1986**, *10* (3), 57-71.

(299) Garcia-Ruiz, M. A.; Gutierrez-Pulido, J. R. An overview of auditory display to assist comprehension of molecular information. *Interacting with Computers* **2006**, *18* (4), 853-868.

(300) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1* (4), 337-341. DOI: 10.1016/j.ddtec.2004.11.007.

(301) Ning, K. *Most used keys on Spotify*. 2020. https://forum.bassbuzz.com/t/most-used-keys-on-spotify/5886 (accessed.

(302) Chung, N. C.; Miasojedow, B.; Startek, M.; Gambin, A. Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics* **2019**, *20* (Suppl 15), 644. DOI: 10.1186/s12859-019-3118-5.

(303) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **2014**, *57* (8), 3186-3204.

(304) Roberts, A.; Engel, J.; Raffel, C.; Hawthorne, C.; Eck, D. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. **2018**.

(305) Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A. Automated design of ligands to polypharmacological profiles. *Nature* **2012**, *492* (7428), 215-220.

(306) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci* **2018**, *4* (1), 120-131. DOI: 10.1021/acscentsci.7b00512.

(307) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and their Compositionality. **2013**.

(308) Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. 2014.

(309) Bravenec, A. D.; Ward, K. D. Interactive Python Notebooks for Physical Chemistry. *Journal of Chemical Education* **2022**, *100* (2), 933–940. DOI: 10.1021/acs.jchemed.2c00665.

(310) Lafuente, D.; Cohen, B.; Fiorini, G.; García, A. A.; Bringas, M.; Morzan, E.; Onna, D. A Gentle introduction to machine learning for chemists: an undergraduate workshop using python notebooks for visualization, data processing, analysis, and modeling. *Journal of Chemical Education* **2021**, *98* (9), 2892-2898. DOI: 10.1021/acs.jchemed.1c00142.

(311) van Staveren, M. Integrating Python into a Physical Chemistry Lab. *Journal of Chemical Education* **2022**, *99* (7), 2604-2609. DOI: 10.1021/acs.jchemed.2c00193.

(312) Weiss, C. J. A creative commons textbook for teaching scientific computing to chemistry students with python and Jupyter notebooks. *Journal of Chemical Education* **2020**, *98* (2), 489-494. DOI: 10.1021/acs.jchemed.0c01071.

(313) Menke, E. J. Series of Jupyter Notebooks Using Python for an Analytical Chemistry Course. *Journal of Chemical Education* **2020**, *97* (10), 3899-3903. DOI: 10.1021/acs.jchemed.9b01131.

(314) Gupta, Y. M.; Kirana, S. N.; Homchan, S.; Tanasarnpaiboon, S. Teaching Python programming for bioinformatics with Jupyter notebook in the Post-COVID-19 era. *Biochemistry and Molecular Biology Education n/a* (n/a). DOI: 10.1002/bmb.21746.

(315) Bisong, E.; Bisong, E. Google colaboratory. *Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners* **2019**, 59-64. DOI: 10.1007/978-1-4842-4470-8_7.

(316) Kluyver, T.; Ragan-Kelley, B.; Pérez, F.; Granger, B. E.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J. B.; Grout, J.; Corlay, S. *Jupyter Notebooks-a publishing format for reproducible computational workflows*; 2016. DOI: 10.3233/978-1-61499-649-1-87.

(317) McKinney, W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference* **2010**, *445*, 51–56. DOI: 10.25080/Majora-92bf1922-00a.

(318) Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; Oliphant, T. E. Array programming with NumPy. *Nature* **2020**, *585* (7825), 357-362. DOI: 10.1038/s41586-020-2649-2.

(319) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **2020**, *17* (3), 261-272. DOI: 10.1038/s41592-019-0686-2.

(320) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **2007**, *9* (3), 90-95. DOI: 10.1109/MCSE.2007.55.

(321) Matplotlib. *Matplotlib Figure Documentation*. https://matplotlib.org/stable/api/figure_api.html#matplotlib.figure.Figure (accessed 2023 January).

(322) Matplotlib. *Matplotlib Scatter Documentation*. https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.scatter.html (accessed 2023 January).

(323) Wishart, D.; Knox, C.; Guo, A.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. Drugbank: a comprehensive resource for in silico drug discovery and exploration. **2006**, *34 (Database issue):D668-72. 16381955.* DOI: 10.1093/nar/gkj067.

(324) Achdout, H.; Aimon, A.; Bar-David, E.; Barr, H.; Ben-Shmuel, A.; Bennett, J.; Bilenko, V. A.; Boby, M. L.; Borden, B.; Bowman, G. R. Open Science Discovery of Oral Non-Covalent SARS-CoV-2 Main Protease Inhibitor Therapeutics (preprint). **2020**. DOI: 10.1101/2020.10.29.339317.

(325) Hill, A. P.; Young, R. J. Getting physical in drug discovery: a contemporary perspective on solubility and hydrophobicity. *Drug discovery today* **2010**, *15* (15-16), 648-655. DOI: 10.1016/j.drudis.2010.05.016