University of Tartu
Institute of Computer Science
Conversion Master in IT

Anneli Kruve-Viil

Monte Carlo tree search in designing of high sensitivity derivatization reagents for mass spectrometric analysis

Master's Thesis (15 ECTS)

Supervisor: Meelis Kull, PhD

Tartu 2021

Derivatiseerivate reagentide struktuuri optimeerimine Monte Carlo meetodil

Lühikokkuvõte

Massispektromeetrilist analüüsi kasutatakse laialdaselt erinevate keemiliste ühendite tuvastamiseks ja kvantiseerimiseks. Kahjuks on mõnede ühendite tundlikkus massispektromeetrias madal. Üheks võimaluseks tundlikkust parandada on uuritavaid ühendeid derivatiseerida. Töö eesmärgiks on disainida uudse struktuuriga reagente, mis oluliselt parandaksid aminohapete massispektromeetrilise analüüsi tundlikkust. Selleks kasutati erineva laiuse ja sügavusega Monte Carlo puu otsingut koos *in-silico* ionisatsiooni efektiivsuse ennustamisega. Kõige efektiivsemaks osutus laia puu otsingu kasutamine, kus juba esimese nelja sammuga saadi kõrget tundlikkust andvaid reagente. Parimad saadud reagendid on umbes 50 korda paremad kui praegusel hetkel praktiliselt kättesaadavad reagendid.

Võtmesõnad: Puu otsing, graafid, massispektromeetria, optimeerimine

CERCS: P176

Monte Carlo tree search in designing of high sensitivity derivatization reagents for mass spectrometric analysis

Abstract

Liquid chromatography electrospray mass spectrometry is widely used to detect and quantify chemicals in various samples. Unfortunately, some compounds have very low sensitivity and require derivatization before sufficient sensitivity can be achieved. Here we apply inverse molecular design, namely Monte carlo tree search, for designing derivatization reagents from a set of functional groups to improve detection of amino acids. The most efficient proved to be search with high breadth, e.g. considering 10 to 50 functional groups in each addition. We obtain reagents which are predicted to yield 50x higher sensitivity than currently used reagents. The structures of the reagents are highly novel and inspirational for the future design.

Keywords: Monte Carlo Tree Search, inverse molecular design, mass spectrometry, ionization, derivatization

CERCS: P176

Table of Contents

Abbreviations

Arg	Arginine
ESI	Electrospray Ionisation
Glu	Glutamic acid
Gly	Glycine
LC/MS	Liquid Chromatography
MCTS	Monte Carlo Tree Search
MS	Mass Spectrometry
Phe	Phenylalanine
SMILES	Simplified Molecular-Input Line-Entry System
VAE	Variational Autoencoder

Abstract

Liquid chromatography electrospray mass spectrometry (LC/ESI/MS) is widely used to detect and quantify metabolites from biological samples, unwanted compounds from food and environment as well as chemicals in industrial processes. Unfortunately, some compounds have very low sensitivity in mass spectrometry due to poor ionization efficiency in the electrospray ionization source. In order to improve the detectability of these compounds, chemical reactions can be used to change the structure of such compounds. These chemical reactions, called derivatization, add new functional groups to the molecule and thereby change its chemical properties. Historically, derivatization reagents have been chosen based on availability. Lately, it has become of interest to develop reagents specifically suitable for LC/ESI/MS by tweaking the structure of existing reagents based on expert knowledge.

Simultaneously, computer aided design of chemical compounds has already been applied since the 80's. Drug design has been the leading power horse, though developments in energy efficient materials for solar panels and power storage devices have recently closely followed. The main techniques use either high-throughput virtual screening or optimization strategies. Most rapidly, different optimization strategies have been developed, including different evolutionary algorithms such as genetic algorithms or Monte Carlo tree search as well as machine learning based generative methods such as variational autoencoders, reinforcement learning, generative adversarial networks either alone or combined. With these methods it has become possible to in-silico synthesise molecules with desired chemical properties. Machine learning based generative methods mostly work with textual representation of the chemical compounds (so called SMILES) and yield compounds close in structure to the known chemical compounds. Evolutionary algorithms on the other hand can work on atom level or functional group (fragment) level and use either textual or graph representation of the compound structure.

The application of virtual screening, evolutionary and generative algorithms in design of derivatization reagents for LC/ESI/MS has been until recently limited due to the lack of possible in-silico ways to measure the success of designed reagents. Recently, our group has developed a random forest algorithm to predict the sensitivity of the compounds in LC/ESI/MS based on their structure only; thereby opening up a possibility for in-silico design of derivatization reagents specifically suited for LC/ESI/MS.

The aim of the current thesis is to explore the possibility to discover candidates for new derivatization reagents for the analysis of amino acids with LC/ESI/MS in positive ionization mode. Specifically, the aims are:

1) To implement a Monte Carlo tree search algorithm for in-silico generation of derivatization reagents for LC/ESI/HRMS.
2) To optimize the algorithm parameters with the aim of discovering reagents yielding highest sensitivity for amino acid analysis in ESI positive mode.
3) To compare the best candidate structures with a baseline generator as well as with the expert designed reagents from the literature.
4) To evaluate the novelty of the candidate structures in comparison to compounds for which LC/ESI/HRMS sensitivity data are available.
5) To evaluate the impact of the order of functional group addition to the ionization efficiency of the reagents.

1. Introduction

Different chemicals yield different sensitivity in the analysis with liquid chromatography electrospray ionization mass spectrometry.[1] To improve the detectability of compounds which possess low ionization efficiency in electrospray, the compounds can be derivatized with suitable chemical reagents.[2] Although trial and error in combination with expert knowledge has yielded some good derivatization reagents, inverse molecular design has potential to speed up the discovery of new reagents.

1.1. Ionization in Electrospray

Different compounds ionize with a different efficiency in electrospray ionization (ESI) source connecting liquid chromatography (LC) and mass spectrometry (MS).[1] The strongly different ionization efficiencies result from the ionization mechanism of ESI.[3] To be detected in mass spectrometry the formation of a gas phase ion occurs: compounds present in the LC effluent need to be charged and transferred into the gas phase.[4]

Measuring and quantifying ionization efficiency is an obvious method to improve the understanding of the ionization process.[1] To estimate the ionization efficiency the response or the relative response of the compound is measured and correlated with the physicochemical properties.[1,5,6] Ion evaporation rate,[7] log*P*,[5,8] hydrophobicity (carboxylic acid chain length for aliphatic compounds and number of fused rings for aromatic compounds),[9] retention times of small peptides in reversed-phase LC,[10] non-polar surface area,[11] gas-phase proton affinity,[12,13] as well as pK_a[14] have been shown to correlate with ionization efficiency.

Recently, Liigand et al.[15] in our group have proposed a random forest regression algorithm to predict the ionization efficiency of any chemical compound in electrospray. For this the compound structure is used as an input and from it different PaDEL descriptors are calculated. The PaDEL descriptors are further used as an input for the random forest regressor that calculates the predicted ionization efficiency of the given compound. The latest version[16] of the random forest regressor has been trained on almost 1500 unique chemical compounds and this dataset is also used in current work.

1.2. Derivatization

Many compounds of immense biological interest or environmental concern lack intrinsic properties that make them easy to detect with analytical methods.[2] For such compounds different chemical modifications can be done to improve their detectability. One type of modification is derivatization which generally involves reaction of specific functional groups of the analyte with a reagent. As a result of derivatization the chemical properties of the analyte are modified. For example, the product or derivate can have light absorbance at a longer wavelength, many become fluorescent or have lower boiling point, depending on the needs of the analytical technique of choice. In LC/ESI/MS the (1) retention in LC and (2) sensitivity in ESI/MS are two parameters that can be improved with derivatization.

One group of compounds that almost always require derivatization are amino acids. Amino acids are building blocks of proteins that perform numerous different tasks in our bodies; however, the concentrations of free amino acids in many samples can be low and their detection without derivatization complicated. Historically, derivatization reagents used for amino acid analysis with LC/ESI/MS have originated from the reagents developed earlier for LC with ultra-violet or fluorescence detection due to availability. Recently, Rebane et al.[17,18] have developed a number of different derivatization reagents specifically keeping in mind the factors improving ionization efficiency in ESI/MS. An overview of the structures of different reagents is available in Supporting Information.

1.3. Representation of Structure

For any computational approach the three dimensional structure of the compound that consists of atoms and bonds between atoms needs to be simplified.[19] In quantum chemical approaches the molecular structure is typically represented as coordinates of atoms in the space as *.xyz* files. Though common in quantum chemistry such data structures are inconvenient for machine learning.

Most commonly Simplified Molecular-Input Line-Entry System (SMILES) representation of the chemical structure is used for handling chemical structures in machine learning.[19] Handling SMILES representation, see Figure 1, can be dealt as with any text and machine learning tools developed for text become accessible. However, SMILES also suffers from a few limitations. Some of the most serious ones being: (1) the SMILES codes of different chemical compounds have different lengths and more importantly (2) one chemical structure

can be represented with multiple SMILE notations,[20] as well as (3) generated SMILES codes may correspond to invalid molecules.[21] To overcome the first difficulty, padding the SMILES with non-meaningful symbols is used to make SMILES representation of all compounds equally long.[22] To overcome the second and third limitation different strategies have been proposed. For example, SMILES standardization has been proposed. Recently, Bjerrum and Sattarov[20] have proposed a more creative approach by using SMILES alongside 2D representation (a picture of the structure) in training of autoencoders. In spite of these difficulties most structure generative methods have focussed on SMILES due to the similarity with text processing.

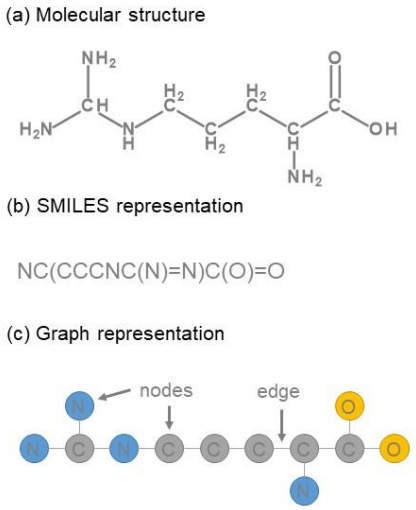


Figure 1 (a) The structure of arginine alongside (b) SMILES and (c) graph representation of it.

Molecular structure can also be represented as a graph, where atoms are denoted as nodes, and bonds as edges, see Figure 1. In contrast to SMILES the graphs generated from the molecules have high validity. Recently, many researchers have proposed that graph based machine learning methods are both more accurate as well as enable design of novel structures.[23–25]

1.4. Inverse Molecular Design

Inverse molecular design starts with a desired chemical property and searches for an ideal molecular structure possessing such property. The inputs are the desired property or properties and outputs are structures.[19] The simplest approach of inverse molecular design is high-throughput virtual screening which starts with an extensive library of structures and uses computational methods to narrow the range of structures based on desired properties.[19] The library can either be a computationally generated library or a library of existing compounds,

e.g., natural products. However, the number of structures considered by the high-throughput virtual screening is immense and it is possible that the optimal structure is not present in the starting library. To overcome these limitations different computational optimization strategies have been proposed.

Structural optimization strategies fall into two categories, evolutionary algorithms or generative models. While evolutionary algorithms have already been used since the 80's,[26] generative models are newer, many of which have emerged with the rise of neural networks.

The early evolutionary algorithms in drug design started with a known active site and aimed to in-silico synthesize a drug fitting to this site via atom or fragment based construction.[26] The fragment based construction is less computationally intensive while atom based construction allows more flexibility. Recently, Leguy et al.[24] have proposed that atom based molecular generation is able to generate chemical structures in both interpretable and highly novel manner by considering seven primary and secondary mutations on a molecular graph: append an atom, change a bond, remove an atom, substitute an atom, insert a carbon atom, cut an atom, or move a functional group. Their algorithm achieved very high accuracy in reaching target properties, though some of the generated structures were unrealistic.

In the field of generative models, variational autoencoders (VAE),[27,28] generative adversarial networks,[29] and different reinforcement learning strategies have been proposed.[30] In a seminal work in 2018 Gomez-Bombarelli et al.[28] proposed usage of three coupled functions: an encoder, a decoder, and a predictor for design of drug-like molecules. The encoder takes SMILES representation as an input, compresses this to a latent space (a continuous vectorial representation) of the molecule. This continuous representation is then further used for reconstruction of the SMILES by a decoder and a property prediction network, see Figure 2. Later the property prediction algorithm can be used to map back from the desired properties to the latent space and from there with the decoder to SMILES representation of compounds with desired properties. The lead of Gomez-Barbarelli et al.[28] has been followed by many other groups[27,31,32] who have aimed to generate chemicals with one or many desired properties.

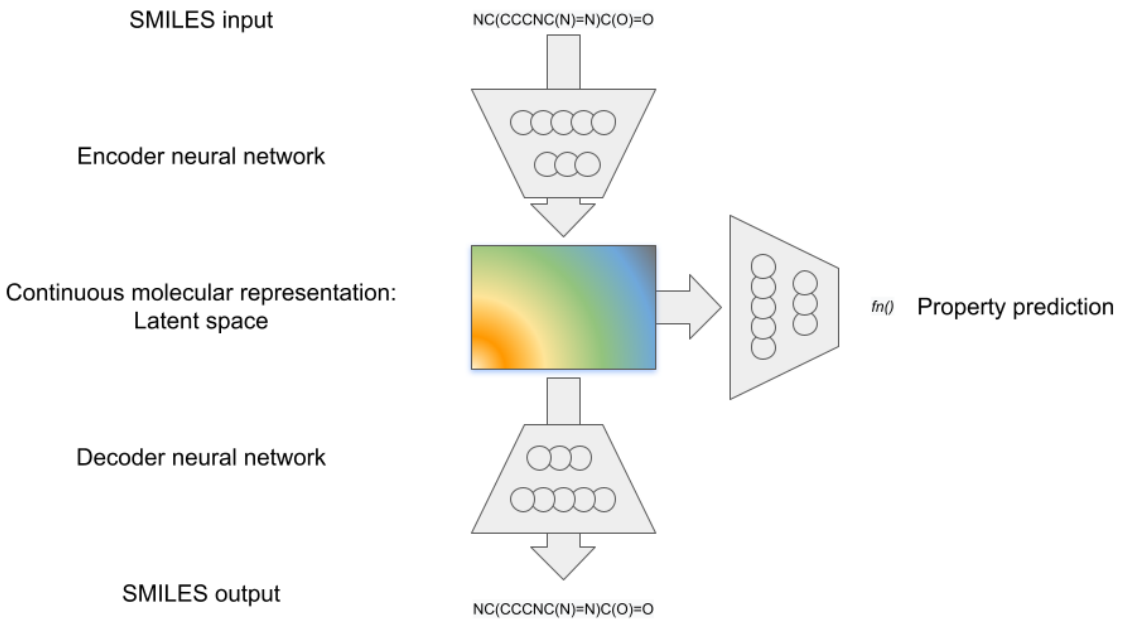


Figure 2 A schematic overview of a variational autoencoder coupled to a property prediction algorithm for generation of structures with desired properties.

VAE in combination with a property prediction layer from the latent representation has recently been also used in mass spectrometry.[33] Namely, the algorithm called DarkChem[33] was used to generate a structural library of metabolites with desired mass to charge ratio and collisional cross-section. The need for vast amounts of training data by VAE in combination with property predictions was overcome by training on different phases. On the first layer only SMILES and mass-to-charge ratio were trained on all compounds in PubChem database (53 M), followed by training on in-silico dataset, and lastly on the real data (500 structures). This was possible because mass to charge ratio can be analytically calculated to any chemical structure. Similarly, Polykovskiy et al.[32] have successfully trained an entangled conditional adversarial autoencoder for drug discovery by using a dataset with computational drug activity labels for training the autoencoder. Generally, though, datasets with large numbers of compounds with known property values are required for using VAE for structure generation. Additionally, the structures obtained with VAE are generally similar to the structures of the training data and it is, therefore, not suitable for discovery of novel structures.

Other machine learning algorithms have been proposed for in-silico design of chemicals. For example, reinforcement learning and generative adversarial networks have been used by Guimareas et al.[29] for the design of water soluble, synthesizable and drug-like molecules.

Gupta et al.[22] have, on the other hand, used generative recurrent networks for drug design, where they have found that fine-tuning for specific ligands is possible even if only five known active substances exist.

The structures obtained with generative models are more or less similar to the dataset used for training the autoencoder or network. In case of metabolite design and drug discovery this can be seen as an advantage as most biological molecules, such as metabolites, share some structural similarity. However, this simultaneously limits the discovery of truly novel structures. Additionally, a dataset is required for training, which is not accessible in case of ionization efficiency data.

Strikingly, Jensen[23] has recently shown that graph-based genetic algorithms can outperform machine learning approaches such as recurrent neural networks[34] in designing molecules with desired octanol-water partitioning coefficient, log$P$. Additionally, he observed that graph based generative models in combination with Monte Carlo tree search (MCTS) can yield results very similar to recurrent neural networks but are orders of magnitude faster (Figure 3). Kajita et al.[35] has further shown that MCTS can efficiently be combined with simulations, such as molecular dynamic simulations, to design liquids with desired viscosity.

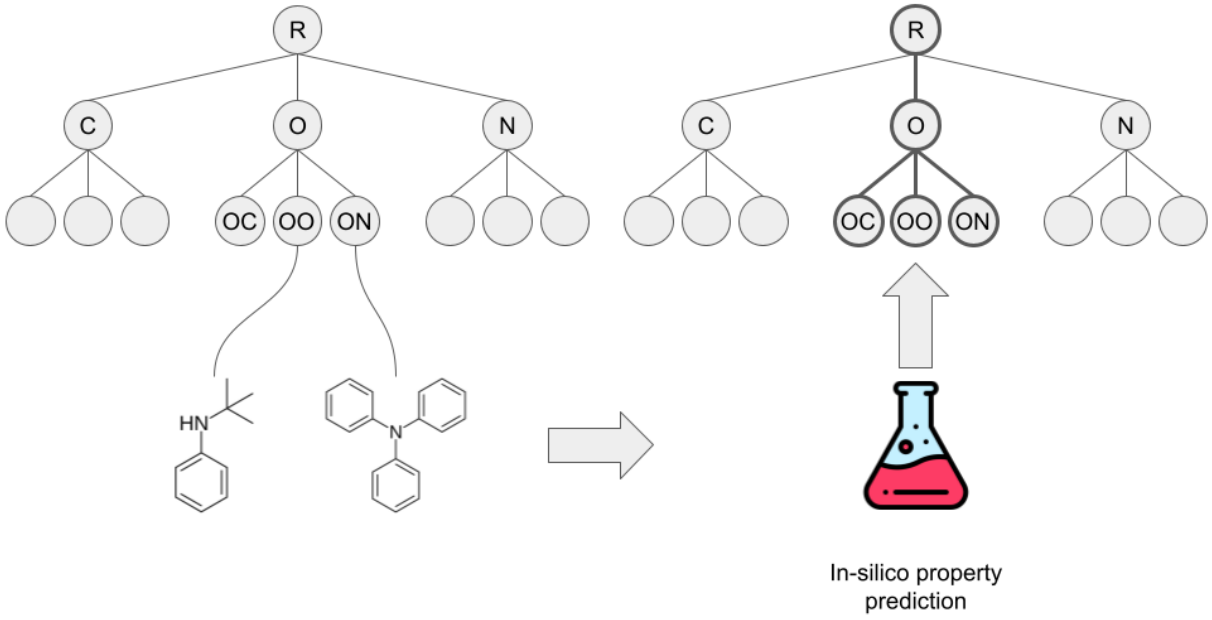


Figure 3 A schematic overview of the Monte Carlo Tree Search algorithm for inverse molecular design of chemicals with desired properties. Here the properties are predicted in-silico with an independent machine learning model or similar.

In conclusion, inverse molecular design takes advantage of many different machine learning tools and operates mainly with SMILES or graph representation of the molecules. Still, the choice of the exact algorithms depends on the studied property and datasets available for training the models.

2. Methods

The functions developed in this work alongside the datasets and generated structures are available from https://github.com/kruvelab/Derivatizating_agent_design.

2.1. Obtaining the Functional Groups

In this work the graphs of the derivatization reagents are built consequently from smaller building blocks, which will be called for simplicity functional groups. Smallest functional groups considered are individual atoms, followed by two bonded atoms, and the largest are up to eleven atoms. The functional groups considered are based on the list of compounds for which ionization efficiency values have been previously measured in different sources.[15,16] This list is useful as it describes which functional groups can and should be present in molecules that can be analysed with LC/ESI/MS. For all molecules in the list, so-called PubChem fingerprints, were retrieved from the PubChem database.[36] The functional groups which were present for less than 1% of the molecules in this dataset, were removed as unlikely to have practical use. Some elements were removed due to restrictions in synthesis, such as elements Si, B, S, P, etc. Additionally, some PubChem fingerprints are originally described textually and were therefore manually drawn and added to the list of functional groups. Some functional groups used in this work are shown in Figure 4, the full list is available from the GitHub repository (SMILES_of_fingerprints.csv). Importantly, the probability of functional groups being present in the dataset was ignored in the following graph generation, as the current dataset may be biased in the representation of some functional groups due to availability of chemicals or choices of the researchers. All functional groups were stored as a list of respective SMILES representations.

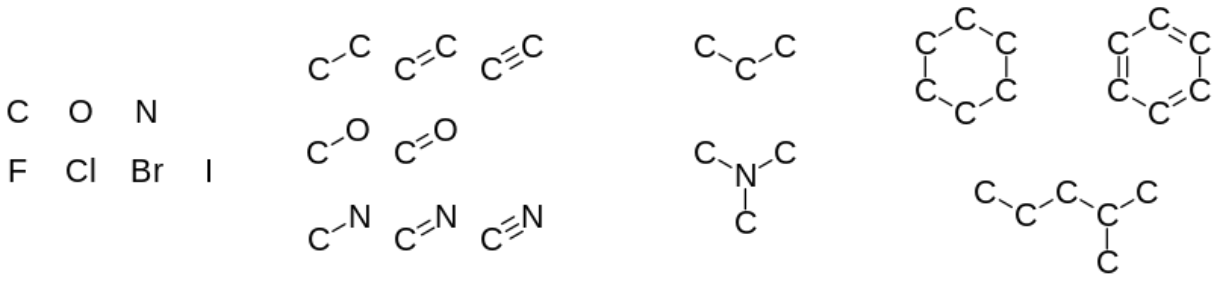


Figure 4 Some functional groups used in the current thesis: simplest functional groups contained a single atom while the more complex functional groups contained six or even more atoms.

2.2. Monte Carlo Tree Search

For generating the structures of a derivatization reagent a Monte Carlo type approach was used (code in MonteCarlo.R). The generation of the structure starts from a "seed". Here in all simulations the seed was methane, so a functional group with a single carbon atom and no edges. In spite of being the simplest possible seed it has been shown to yield compounds with desired properties previously.[24]

The MCTS was carried out with different breadths and depths, where the pseudocode is shown in Code 1. Below we will use notation B^D to denote a search with breadth of B and depth of D.

Code 1 The pseudocode for recursive generation of the reagent structures at each step of MCTS with given *depth* D and *breadth* B. The function addFuncGroup is described in delta in chapters 2.3 and the function evaluateIE is a combination of functions described in 2.4 and 2.5.

```
function MCTS(D, B):
    reagent = seed
    while MolecularWeigh(reagent) < 500 Da:
        sample B functional groups f1:fB
        best_b = None
        best_medianIE = 0
        for f_b in f1:fB
            generated = generateRecursively( addFuncGroup(reagent, f_b) ,D, B)
            listIE = []
            foreach generatedReagent in generated:
                listIE.append(evaluateIE(generatedReagent))
            if median(listIE) > best_medianIE:
                best_b = b
                best_medianIE = median(listIE)
        reagent = addFuncGroup(reagent, best_b)

function  generateRecursively(reagent, D, B):
    if D = 0:
        return [reagent]
    generated = []
    sample B functional groups f1:fB
    for f_b in f1:fB
        generated = generated + generateRecursively(addFuncGroup(reagent, f_b),
    D - 1, B)
    return generated
```

2.3. Connecting Functional Groups Into a Reagent Structure

Appending the structures with new functional groups consisted of four steps (code in connecing_graphs.R). Firstly, the SMILES representation of the reagent structure and a randomly sampled functional group were converted to graph representation with the *rcdk* and *igraph* packages. For practical use the graph representation consisted only of heavy atoms and bonds between these atoms, hydrogens were ignored. Secondly, all atoms that had a free valence available were determined in both the reagent and functional group. Next, for both reagent and functional group an atom with a free valence was randomly sampled and a bond (edge) was added between these two atoms. Lastly, the graph structure was converted back to SMILES representation for further calculations with the RDKit package via python interface (code graph_to_SMILES.R and matrix_to_smiles.py). The respective function in Code 1 is denoted with addFuncGroup.

2.4. Calculating Ionization Efficiency Values

The ionization efficiency values predicted need to account for the complete structure of the chemical ionized in electrospray ionization (code in IE_calculator.R). Therefore, all reagent structures created were additionally coupled (code in add_AH.R) with the structure of four amino acids: glycine (Gly), glutamic acid (Glu), arginine (Arg), and phenylalanine(Phe). These amino acids were chosen to represent compounds with a wide properties range. Glutamic acid and arginine are representatives of amino acids with acidic and basic side chain while glycine is a poorly ionizing amino acid and phenylalanine is known to be a well ionizing amino acid.[37] While connecting the structure of the reagent with the structures of the amino acids, additionally a reactive link was added to the structure. Here we used the methoxycarbonyl group, which is also used in FMOC-Cl reagent,[17] as a reactive link. In coupling amino acids the bond was always induced to indicate reactivity to the amino group of the amino acid.

Ionization efficiency was calculated for each of the derivatized amino acids in each of the MCTS events with a random forest regressor originally trained by Jaanus Liigand.[15] Only ionization efficiencies in electrospray positive mode were considered. Here we used a retrained version of the forest with more data and small modifications by the author of the current work.[16] The ionization efficiency values in the dataset used for training the random

forest ranged from -2 to 7 in a logarithmic scale and consisted of more than 1400 unique chemicals.

To calculate the ionization efficiency PaDEL descriptors[38] were calculated from the SMILES structure of the derivatized amino acids with a program written previously by Henri Anilo[39] and available in GitHub repository (PaDEL_descs_calculator.R). Alongside descriptors about the structure of the compound also eluent composition is required. As there is no information whether the retention in LC is available for the structures, we assumed that all compounds elute in 80/20 acetonitrile/0.1% formic acid mixture.

2.5. Ranking the Structures

In each iteration of MCTS all reagents were ranked based on the ionization efficiency values (evaluation_function.R). For each added functional group the median ionization efficiency over all finally generated reagents starting from this functional group were calculated. The reagents were ranked based on the obtained median ionization efficiency for each amino acid so that the higher rank corresponded to the higher median ionization efficiency. It was considered that improving the ionization efficiency for compounds that have intrinsically lower response is much more important than improving the response of already well ionizing compounds. Therefore, the total rank was calculated as:

$$rank = 3 \cdot rank_{Gly} + 2 \cdot rank_{Glu} + rank_{Arg} + rank_{Phe} \quad (1)$$

for all candidate functional groups, where $rank_{Gly}$, $rank_{Glu}$, $rank_{Arg}$, and $rank_{Phe}$ are the ranks for each amino acid and the coefficients have been selected based on the ionization efficiency of the underivatized amino acids. The functional group with the highest rank was chosen and added to the reagent structure.

2.6 Selection of the Final Structure

In some cases the addition of functional groups reached a plateau after some first steps and further addition did not improve or even reduce the ionization efficiency, see Figures S1 to S24. Therefore, within each search the final structure was chosen as the structure yielding highest mean ionization efficiency over all four amino acids.

3. Results

With Monte Carlo Tree Search 50 derivatization reagents were generated with 2^2, 2^5, 3^3, 10^2, and 50^1 search, where the base shows the breadth and the power shows the depth of the search. Additionally, 70 reagent structures were generated with 1^1 as a baseline method where functional groups are sampled randomly without any tree search. MCTS of 2^2 visited on average 30 structures during one search while 2^5, 3^3, 10^2, and 50^1 searches visited 232, 192, 662 and 362 structures, respectively. The average computation times were 4, 42, 22, 90, and 30 min, respectively. The most time-limiting step was the calculation of PaDEL descriptors.

Depending on the search algorithm the mean predicted ionization efficiency values, hear and below ionization efficiency, ranged between 4.1 and 4.6; 3.7 and 4.4; 3.9 and 4.5; and 4.1 and 4.7 for glycine, glutamic acid, arginine, and phenylalanine, respectively. The highest mean predicted ionization efficiencies for these MCTS were observed for reagents generated with 50^1 MCTS for all amino acids. The minimum, mean, and maximal ionization efficiency values are described in Table 1 and visualized in Figure 5. *t*-test revealed that MCTS yielded on the average higher ionization efficiency values than the baseline method with 1^1 search ($p < 0.05$). Additionally, the mean ionization efficiency values increased in the line of 2^2, 2^5, 3^3, 10^2 to 50^1. Statistically significant differences were not observed between 2^2 and 2^5 as well as 2^5 and 3^3 MCTS for any of the amino acids. MCTS of 50^1 yielded ionization efficiency values statistically significantly higher than 10^2 only for glutamic acid. The *p*-values for all *t*-tests are shown in Table S2.

Additionally, the MCTS yielded reagents that gave ionization efficiency values significantly higher than the currently available reagents, see Figure 5. An average reagent performed 1.0 units better for glycine analysis than the best commercial reagent, see Table 1 for details. This means an increase in sensitivity by a factor of 10 as ionization efficiencies are predicted in a logarithmic scale. The largest improvements, however, were observed for glutamic acid. Best reagents enabled an increase in ionization efficiency of glutamic acid by 1.7 units in comparison to the best currently available reagents, which would mean a sensitivity increase by a factor of 50.

While mean and lowest ionization efficiencies clearly improved for reagents obtained with MCTS with increasing breadth, the best reagents varied. For glutamic acid and arginine the

best reagent was obtained with 3^3 search, while for phenylalanine and glycine 10^2 and 50^1 search yielded reagents with highest ionization efficiency.

Table 1 Summary of the ionization efficiency values for the reagents generated with MCTS and for known reagents. The structures of the known reagents are brought in Table S1.

	MCTS	Gly	Glu	Arg	Phe
minimal IE	1^1	3.0	2.7	2.8	3.3
	2^2	3.3	3.1	3.1	3.5
	2^5	3.6	2.8	3.1	3.4
	3^3	3.5	2.9	3.5	3.1
	10^2	3.7	3.7	3.7	3.9
	50^1	3.8	3.6	3.9	4.1
mean IE	1^1	4.1	3.7	3.9	4.1
	2^2	4.2	3.9	4.1	4.3
	2^5	4.4	3.9	4.2	4.4
	3^3	4.4	4.1	4.3	4.5
	10^2	4.5	4.3	4.4	4.7
	50^1	4.6	4.4	4.5	4.7
maximal IE	1^1	4.8	4.8	4.8	4.9
	2^2	4.9	4.9	4.9	5.1
	2^5	4.9	4.9	4.8	5.0
	3^3	4.9	5.0	5.1	5.0
	10^2	5.0	4.9	4.9	5.1
	50^1	5.2	4.8	4.9	5.0
known reagents	AQC	2.6	2.7	3.0	3.3
	DBEMM	3.3	3.2	3.7	3.8
	DEEMM	2.4	2.0	3.1	2.7
	DNS	2.9	3.1	4.0	4.2
	EBEMM	2.7	2.9	3.2	3.5
	FMOC	2.3	3.1	3.9	3.4
	FOSF	3.6	2.9	3.4	3.5
	PrCl	2.1	2.3	2.7	2.7
	TAHS	2.4	2.4	2.9	3.0

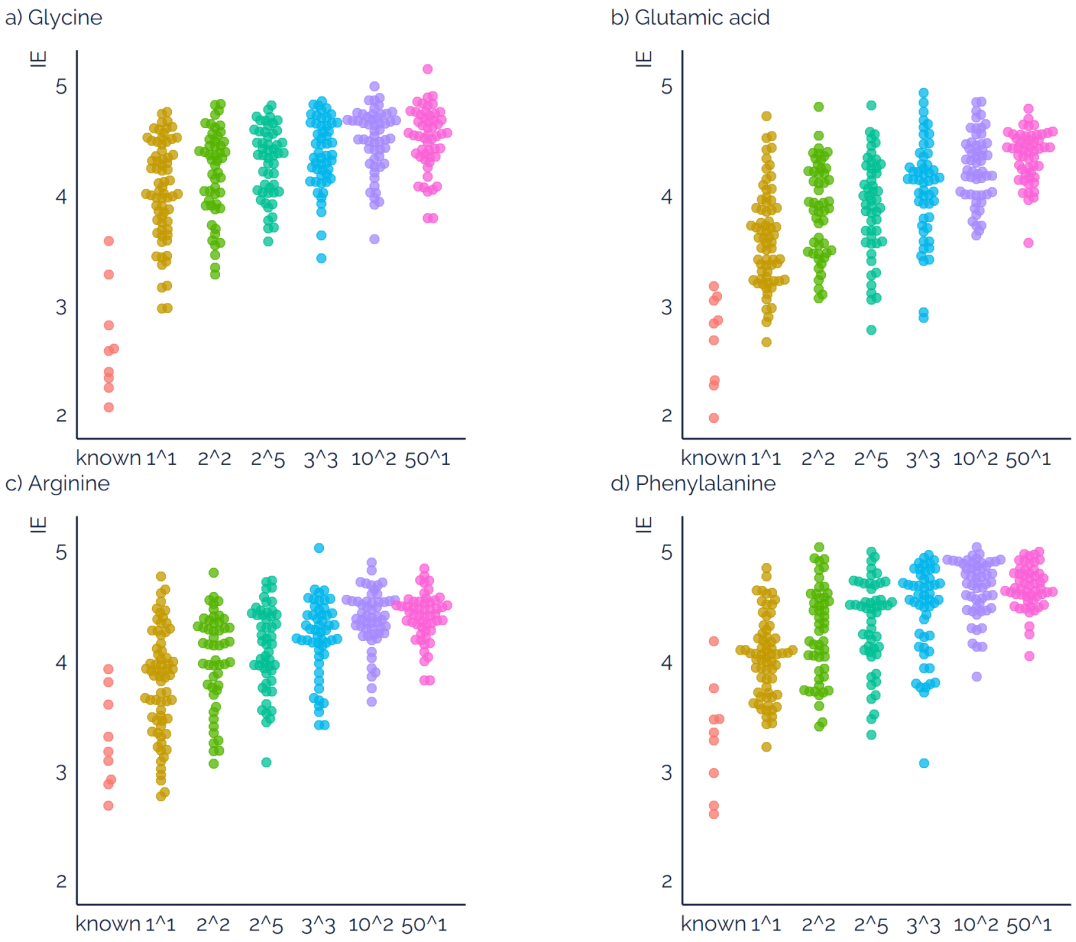


Figure 5 The comparison of ionization efficiency values for a) glycine, b) glutamic acid, c) arginine, and d) phenylalanine derivatized with known reagents (see Table S1 for structures) and reagents obtained with 1^1 baseline method, as well as MCTS of 2^2, 2^5, 3^3, 10^2, and 50^1.

Additionally, the range of ionization efficiency values for the reagents became narrower with the wider search breadth, Figure 5. This visual observation was tested with an *F*-test. It was confirmed that variability in ionization efficiency values yielded by the baseline 1^1 search was higher in comparison to 2^5, 3^3, 10^2 or 50^1. Also, 2^2 search yielded statistically significantly higher variability in comparison to 10^2 or 50^1 searches. However, further increase in the search breadth did not prove to narrow the range of ionization efficiency values statistically significantly.

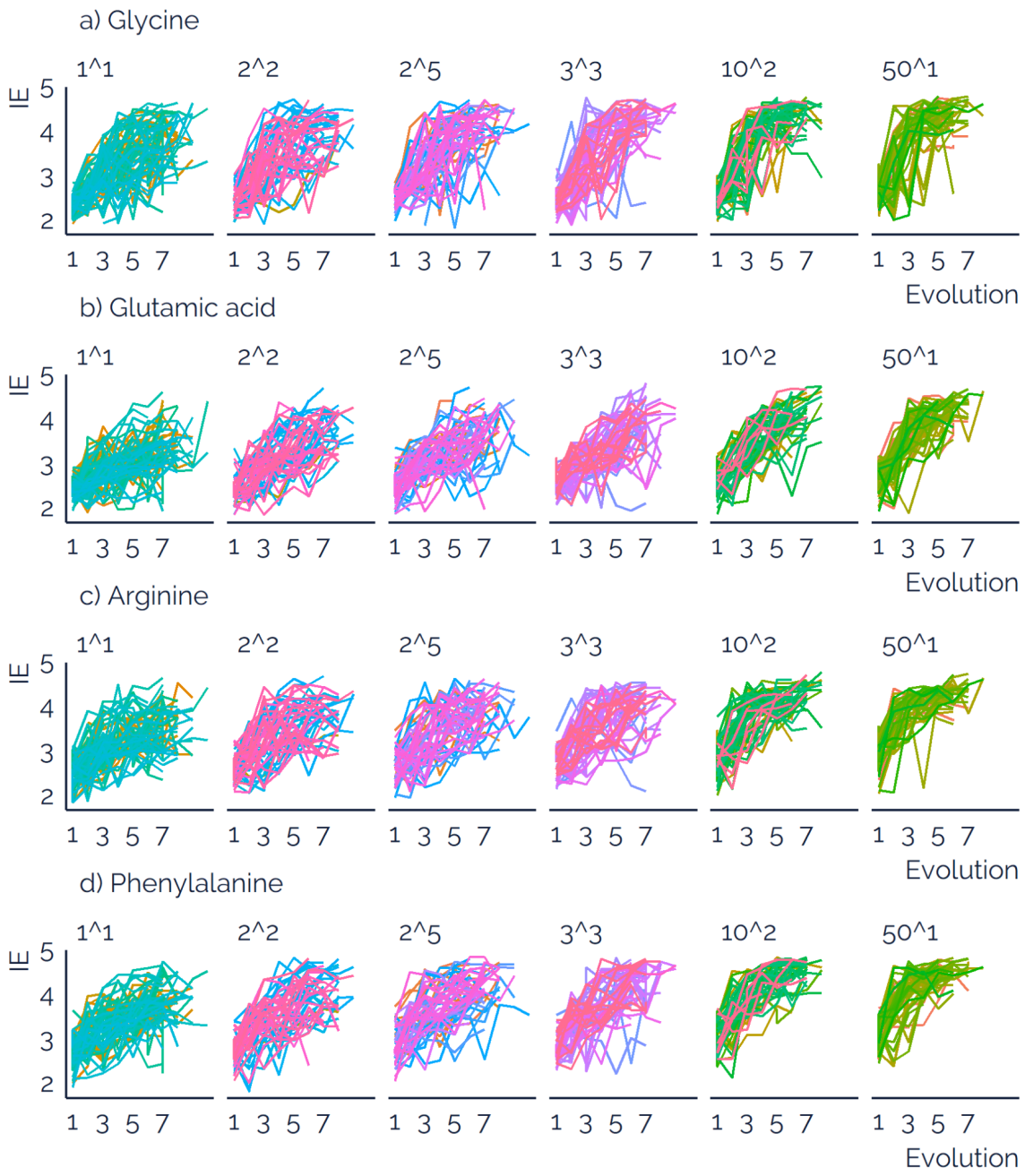


Figure 6 The change in the ionization efficiency of derivatized amino acids along the structure generation of the derivatization reagent. One line shows one search and evolution indicates the number of functional groups added.

Independent of the MCTS strategy, five to seven steps were compiled by the search before the finishing criterion, molecular weight over 500 Da, was reached. The evolution of the ionization efficiency value for each search algorithm is shown in Figure 6 and each individual search together with ionization efficiency values for all visited states can be seen in Figures S1 to S24. Visual analysis suggests that generally 10^2 and 50^1 search resulted in a reagent with high ionization efficiency with less steps. This is considered highly beneficial, as less steps also means molecules which consist of less atoms and are easier to synthesise. Therefore, we

also compared the mean ionization efficiencies for reagents obtained after addition of 4 function groups with all MCST algorithms. The 10^2 and 50^1 search yielded statistically significantly highest ionization efficiencies for all amino acids in comparison to other search breadth and depth combinations. No statistically significant differences were observed between 10^2 and 50^1 search. The *p*-values are described in Table S3.

One of the objectives of the current thesis was to obtain reagents with novel chemical structures. To evaluate the novelty, we compared the properties of the reagents generated by MCTS with known derivatization reagents and a full list of compounds, for which ionization efficiency values in ESI/MS have been measured and used for training the random forest algorithm used in this work. For this a PCA analysis of all of the calculated PaDEL descriptors of these compounds was conducted. First 10 principal components explained only 20% of the total variation in the PaDEL descriptors, which is expected as 860 descriptors are considered in the comparison. It was observed that the reagents created with MCTS formed a group which was somewhat separated from the compounds with known ionization efficiency values in PC1 vs PC2 plot, see Figure 7. Also, the reagents generated with MCTS were separated from currently available reagents, see Figure 7c. However, the groups overlap in the comparison of next principal components. While comparing the ionization efficiency and position on the PCA plot, Figure 7a, for the known reagents it is clearly seen that the reagents created with MCTS extend in the direction of high ionization efficiency values.

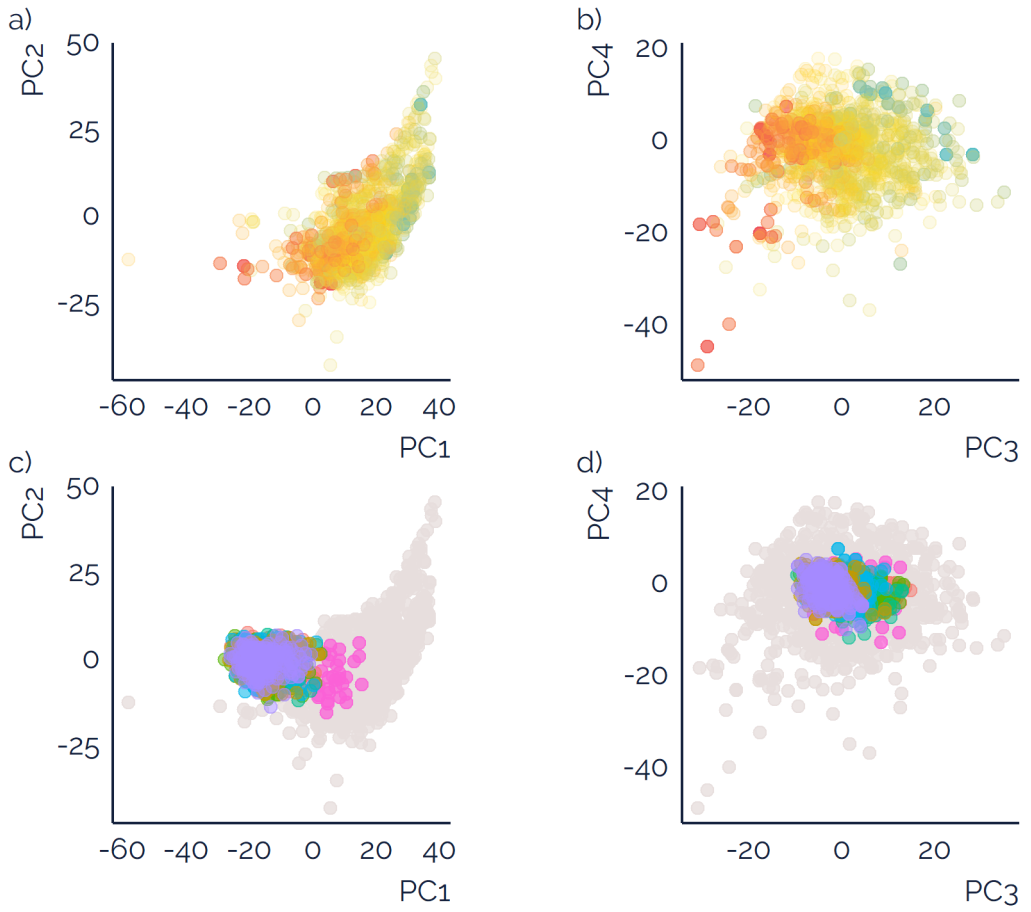


Figure 7 PCA plot for comparing the PaDEL descriptors for compounds for which ionization efficiency values have been measured. a) and b) PCA plot for compounds for which ionization efficiency values have been previously measured and used for training the random forest model. Color shows the ionization efficiency, red indicates high values while blue indicates low values. c) and d) PCA plot overlaying the compounds with known ionization efficiency values (grey dots), known derivatization reagents (pinkish purple dots) and reagents generated through MCTS.

Last but not least, it was of interest to see how much impact does the order of adding functional groups have on the obtained ionization efficiencies. To test this, we selected two cases of reagent generation from the previous MCTS. One of the cases was the so-called "successful case", where the final reagent yielded one of the highest ionization efficiencies while the second case yielded one of the lowest ionization efficiencies in the original search. Both cases consisted of 7 added functional groups. All added functional groups were extracted and their addition order was shuffled 20 times. The ionization efficiency was calculated for each step and the final structure, shown in Figure 8 and 9. For the successful case, Figure 8, the ionization efficiency generally steadily increases with addition on the next functional groups independent of the order. However, for the second set of functional groups,

Figure 9, both dramatic increase and decrease in ionization efficiency could be observed with each added functional group.

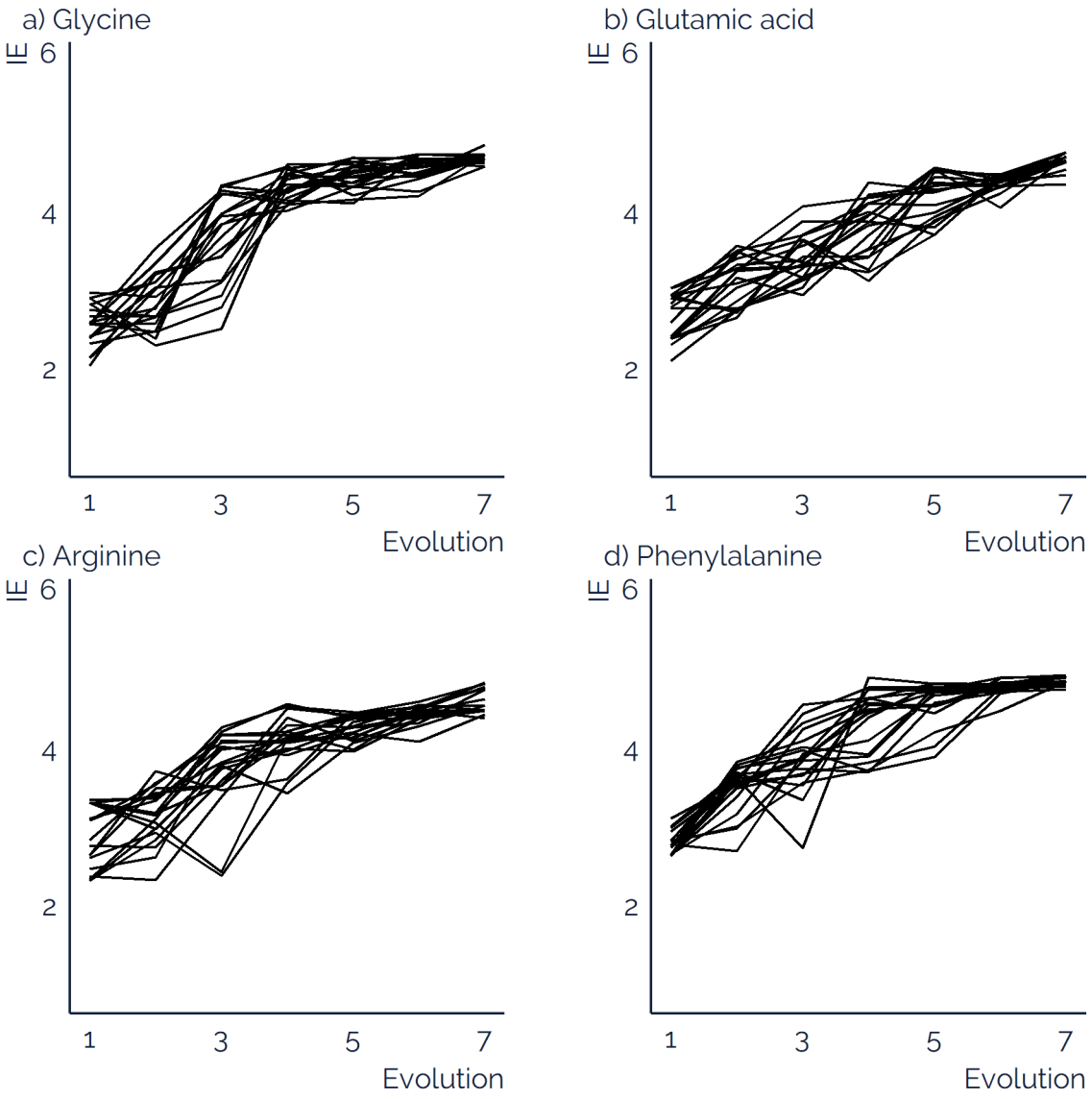


Figure 8 The impact of the order of adding functional groups to the reagent structure for a successful generation.

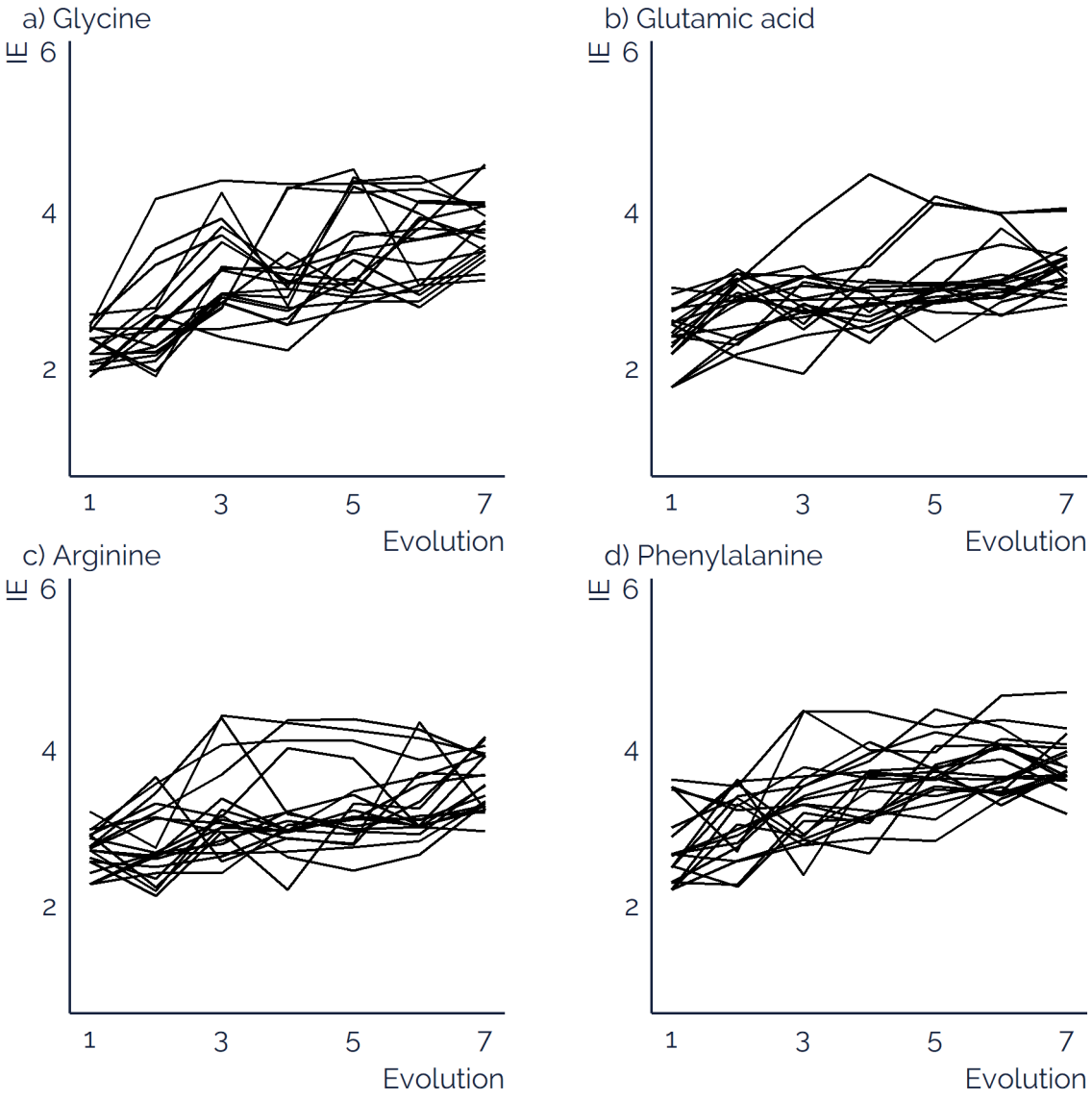


Figure 9 The impact of the order of adding functional groups to the reagent structure for an unsuccessful generation.

Generally, the variation in ionization efficiency values for the final structures are much larger for the unsuccessful case (Figure 9). Also, the distribution of ionization efficiencies of the full structure have a very narrow distribution in comparison to the "unsuccessful case". Interestingly, it can be seen that the impact of the order somewhat depends on the amino acid considered. For example, in Figure 8 the ionization efficiency of glycine and arginine reaches a plateau already after adding five first functional groups independent of the order. However, glutamic acid does not follow this trend.

4. Discussion

4.1. Implementation and Optimization of Search Algorithm

The comparison of derivatization reagent structures and obtained ionization efficiency values across different search breadths and depths clearly showed that 10^2 and 50^1 searches were more effective than MCTS with lower breadth. These search algorithms yielded structures which on the average had higher ionization efficiency, reagents with higher ionization efficiencies were obtained with less steps, and the distribution of the ionization efficiencies was narrower. This results from the fact that 10^2 and 50^1 searches consider more possible functional groups in each step and are, therefore, more likely to find a functional group that truely is improving ionization efficiency.

Interestingly, no statistically significant differences were observed between 10^2 and 50^1 search. In the 10^2 search, each of the 10 functional groups is also coupled with 10 other functional groups and the final ranking of the functional groups is based on the median ionization efficiency after these two consequent couplings. Therefore, the rank is averaged over ten next possible "moves". In 50^1 search, only the impact of the added functional group itself is considered while selecting the best addition. The fact that these two searches gave indistinguishable results facilitates the importance of considering many possible choices of functional groups over using a search with a large depth. To further confirm this observation largening the depth of the search, e.g. 50^2, would be necessary. Here we carried out a single 50^2 search for which the results are visualized in Figure 10. This single search yielded results comparable to the mean results of 10^2 and 50^1 search. However, more tests are required in the future for testing statistical significance.

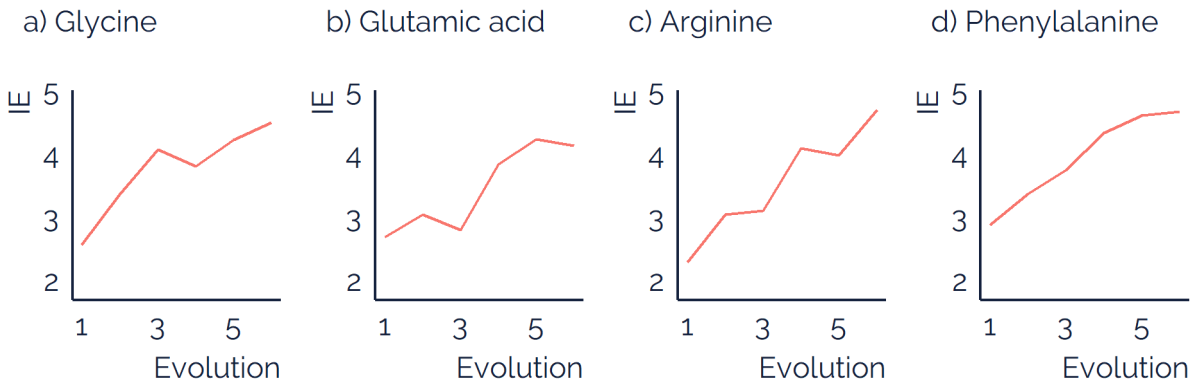


FIgure 10 The change in the ionization efficiency of derivatized amino acids along the structure generation of the derivatization reagent with 50^2 search.

The analysis of the structures of the obtained reagent revealed that most of the final structures were too complex for practical applicability. However, from the analysis of evaluation of ionization efficiency values with each step in 10^2 and 50^1 searches (Figure S17 to S24) it was observed that many structures yielded high and very high ionization efficiency values already with a few steps. This suggests that it is either possible to prune the obtained reagent structures to improve the synthetic accessibility of these structures or to lower the hard stopping criteria, e.g., to 300 Da, for structure generations.

To evaluate the possibility of pruning we evaluated the structures obtained with the first four steps. The best pruned structures gave 0.7 to 1.4 units higher ionization efficiencies than commercial reagents, depending on the amino acid. Also, many pruned structures were synthetically accessible based on expert knowledge. The verification of such structures will require chemical testing in the future.

4.2. Performance of the Search Algorithm in Comparison to Commercial Reagents

Structures generated with MCTS yielded ionization efficiency values significantly higher than the ionization efficiency values for currently existing commercial reagents. The improvement was largest for glycine and glutamic acid. These two amino acids had a higher impact on the ranking of functional groups in MCTS, equation 1. This shows that using coefficients in the ranking is an efficient way to improve performance for amino acids which are known to have intrinsically low ionization efficiency.

Importantly, the predicted ionization efficiency values obtained with the reagents generated by MCTS were more similar for all considered amino acids. For example, the reagent yielding highest mean ionization efficiency showed variation of less than 0.2 (1.6x) units between different amino acids. On the other hand, the commercial reagents possess high differences in ionization efficiency depending on the amino acid considered. The differences for commercial reagents range from 0.6 (4x) to 1.6 (40x) units. This suggests that the generation of derivatization reagents with MCTS is successful in evenly enhancing ionization efficiency of different amino acids and may propose novel reagents or be used as a starting point for the expert design of the reagents.

4.3. Novelty of the Reagents

The analysis of the PCA plot revealed that on the first principal component the reagents yielded by MCTS were somewhat separated from known reagents as well as from most of the compounds for which experimentally measured ionization efficiency values are available. This suggests that the structures obtained with MCTS are novel as they possess PaDEL descriptors different from known compounds.

A detailed analysis of the structures revealed that generated reagents indeed are very uncommon in their structure. Most structures are highly unsymmetrical, contain many longer alkyl moieties and some branching. As an example, a structural step-by-step evolution of the reagent yielding on the average the highest ionization efficiency is shown in Figure 11. Additionally, structures of three successful reagents are shown in Figure 12. Unfortunately, most such structures are extremely hard if not impossible to be synthesised in larger scales. Fortunately, it was observed (paragraph 4.1) that many such structures can be pruned and still yield reagents with high ionization efficiency.

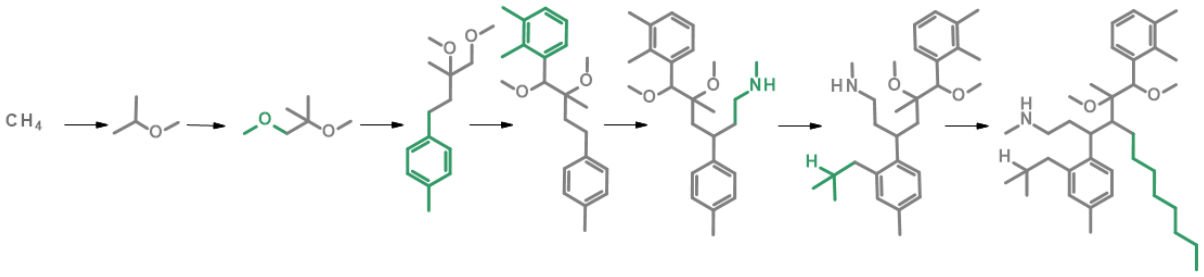


Figure 11 The evolution of the best obtained reagent structure in seven steps. The functional group added in each step is shown in green.

Also some specific functionalities can be outlined. Interestingly, many successful structures contain an ether bond which is not likely to show up in most chemical structures designed by experts. It would be an interesting future study to verify if ether groups are truly beneficial in the reagent structure or can be substituted with alkyl groups for improved synthetic accessibility. Notably, all successfully generated reagents contained at least one amine group, either a secondary or tertiary. This is also chemically logical, as amines are basic groups that generally improve the ionization efficiency of a compound. Underivatized amino acids contain an amine group but after derivatization this group is turned into a very weakly basic amide

group and the addition of strongly basic sites to the structure through derivatization is likely improving ionization efficiency.

As a result, the PCA analysis and expert evaluation reveal that most generated structures are indeed novel. This assures that using graphs with Monte Carlo search elements is indeed able to widen the range of imaginable structures.

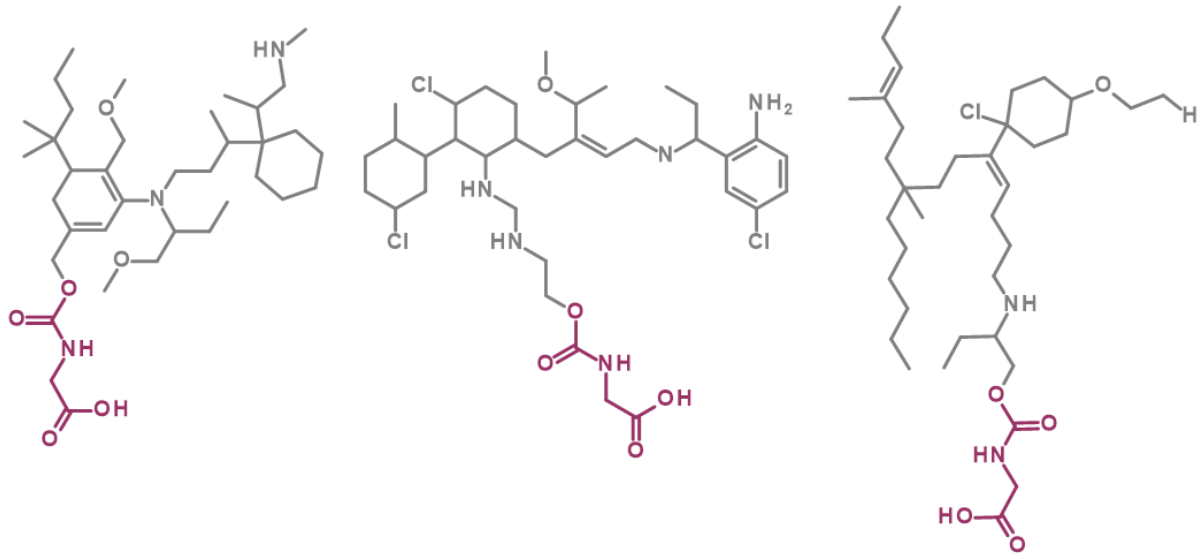


Figure 12 The structures of the three reagents yielded from the MCTS. The purple part indicates amino acid and reactive group.

4.4. Order of Adding Functional Groups

One of the research questions in this work was also, how important it is to consider the order of adding the functional groups to the structure. For example, if an individual success score could be annotated to each functional group independent of the structure to which this functionality is added and independent of the next functional groups that can be added, it would enable scoring of the functionalities before the MCTS starts and considering only the so called successful functional groups. Importantly, here the addition order captures both the actual order in which the functional groups are used to append the reagent structure as well as the relative position of these functional groups as most functional groups can be connected with the base structure via multiple different bonds.

However, the two considered simulations, Figure 8 and 9, showed that the impact of functional groups strongly depends on the structure to which it is added as well as the amino acid that is being considered. In some cases ionization efficiency increased with the addition

of a single functional group by one unit but adding the next functional group decreased the ionization efficiency again by one unit, or vice versa.

We also compared the functional group frequency for the dataset with measured ionization efficiency values to reveal if any functional groups which increase ionization efficiency can be pinpointed. We compared the functional groups that were present in at least 1% of all structures for both the full dataset and compounds yielding ionization efficiencies in the top quartile. We observed that the functional groups highly overlapped. Only 22 out of 368 functional groups were present in the full dataset but were lacking in the top quartile. This supports the earlier finding that ionization efficiency does not solely depend on the functional groups but also the way they are coupled with each other. Therefore, we evaluate that here prescoring of the functional groups is likely to be inefficient.

Conclusion

The aim of this work was to propose new derivatization reagents for amino acid analysis with LC/ESI/MS based on an inverse molecular design with Monte Carlo tree search. During the thesis a full set of functions enabling tree search was written, including converting SMILES based functional groups to graphs, generating the reagent structures from these graphs, computing the ionization efficiency from necessary PaDEL descriptors for the reagents, and ranking the functional groups.

The Monte Carlo tree search was tested with six different realizations, 1^1, 2^2, 2^5, 3^3, 10^2, and 50^1. Generally, the mean ionization efficiency values increased with the increasing breadth of search. Searches with larger breadth also yielded reagents with very high efficiency with less steps. The most efficient proved to be 10^2 and 50^1 search, which yielded many structures with ionization efficiency values well above the values reachable with currently available reagents. Based on the calculated ionization efficiency values the best structures can improve analysis sensitivity by a factor of 50.

We additionally observed that yielded structures are very different from the currently available reagents, both based on principal component analysis as well as on visual analysis. This indicates that tree search based on functional groups is able to yield highly novel structures, which are particularly important in designing compounds and materials with completely novel properties for applications in analytical chemistry. For example, reaching novel structures is essential in development of many new reagents, mobile phase components, and stationary phase materials in chromatographic separation. Even more so, Monte Carlo tree search can widen the currently available databases of chemical compounds, which are widely used in structural identification in analytical chemistry.

Future Perspective

The results of the thesis suggest that graph generative methods with Monte Carlo tree search can yield derivatization reagents that are better than any current reagent. Still, algorithms developed here are based on a few chemical and algorithmic simplifications which shall be further evaluated in the future.

Chemical simplifications are accounting for the solvent composition, ionization mode, and reactive group of the designed reagents. Firstly, current ionization efficiency predictions assume the same solvent composition for all amino acids and reagents. In reality the solvent composition depends on the time at which the molecule elutes from the LC column. This can be accounted for by combining a retention time prediction alongside ionization efficiency prediction into the algorithm. Secondly, measurements in mass spectrometry can also be carried out in negative ionization mode. Therefore, it would be advantageous to optimise the reagents in parallel for analysis both in positive and negative mode. Thirdly, all reagents developed here use the same reactive group, methoxycarbonyl group, to react with the amino acids. In practice many different reactive groups can be used and it would be of interest to add this as one layer to the search algorithm.

Algorithmic simplifications are mostly related to the rollout policy and ranking of the functional groups. For example, in the current algorithm the reagent structures are ranked solely based on the ionization efficiency. However, some of the structures are very complicated and almost impossible to synthesise. Here two strategies can be used to overcome these limitations. Either to prune the structures, which was tested in this work, or to adjust the ranking function to account for the complexity of the structures. For example, either molecular mass or synthetic accessibility score could be added to the ranking to facilitate design of derivatization reagents with simpler structures.

All in all, the development of reagents with Monte Carlo tree search has many promising avenues to speed up the development of derivatization reagents. Coupling of current algorithms with ionization efficiency predictions in negative mode and retention time predictions are feasible extensions for the near future alongside adjustment of the scoring function.

References

(1) Oss, M.; Kruve, A.; Herodes, K.; Leito, I. Electrospray Ionization Efficiency Scale of Organic Compounds. *Anal. Chem.* **2010**, *82* (7), 2865–2872. https://doi.org/10.1021/ac902856t.

(2) Santa, T. Derivatization Reagents in Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry. *Biomed. Chromatogr.* **2011**, *25* (1–2), 1–10. https://doi.org/10.1002/bmc.1548.

(3) Cech, N. B.; Enke, C. G. Practical Implications of Some Recent Studies in Electrospray Ionization Fundamentals. *Mass Spectrom. Rev.* **2001**, *20* (6), 362–387.

(4) Enke, C. G. A Predictive Model for Matrix and Analyte Effects in Electrospray Ionization of Singly-Charged Ionic Analytes. *Anal. Chem.* **1997**, *69* (23), 4885–4893. https://doi.org/10.1021/ac970095w.

(5) Chalcraft, K. R.; Lee, R.; Mills, C.; Britz-McKibbin, P. Virtual Quantification of Metabolites by Capillary Electrophoresis-Electrospray Ionization-Mass Spectrometry: Predicting Ionization Efficiency Without Chemical Standards. *Anal. Chem.* **2009**, *81* (7), 2506–2515. https://doi.org/10.1021/ac802272u.

(6) Kruve, A.; Kaupmees, K.; Liigand, J.; Leito, I. Negative Electrospray Ionization via Deprotonation: Predicting the Ionization Efficiency. *Anal. Chem.* **2014**, *86* (10), 4822–4830. https://doi.org/10.1021/ac404066v.

(7) Kebarle, P.; Tang, L. FROM IONS IN SOLUTION TO IONS IN THE GAS PHASE. *Anal. Chem.* **1993**, *65* (22), 972A-986A. https://doi.org/10.1021/ac00070a715.

(8) Henriksen, T.; Juhler, R. K.; Svensmark, B.; Cech, N. B. The Relative Influences of Acidity and Polarity on Responsiveness of Small Organic Molecules to Analysis with Negative Ion Electrospray Ionization Mass Spectrometry (ESI-MS). *J. Am. Soc. Mass Spectrom.* **2005**, *16* (4), 446–455. https://doi.org/10.1016/j.jasms.2004.11.021.

(9) Huffman, B. A.; Poltash, M. L.; Hughey, C. A. Effect of Polar Protic and Polar Aprotic Solvents on Negative-Ion Electrospray Ionization and Chromatographic Separation of Small Acidic Molecules. *Anal. Chem.* **2012**, *84* (22), 9942–9950. https://doi.org/10.1021/ac302397b.

(10) Cech, N. B.; Enke, C. G. Relating Electrospray Ionization Response to Nonpolar Character of Small Peptides. *Anal. Chem.* **2000**, *72* (13), 2717–2723.

https://doi.org/10.1021/ac9914869.

(11) Ghosh, B.; Jones, A. D. Dependence of Negative-Mode Electrospray Ionization Response Factors on Mobile Phase Composition and Molecular Structure for Newly-Authenticated Neutral Acylsucrose Metabolites. *The Analyst* **2015**, *140* (19), 6522–6531. https://doi.org/10.1039/C4AN02124J.

(12) Amad, M. H.; Cech, N. B.; Jackson, G. S.; Enke, C. G. Importance of Gas-Phase Proton Affinities in Determining the Electrospray Ionization Response for Analytes and Solvents. *J. Mass Spectrom.* **2000**, *35* (7), 784–789. https://doi.org/10.1002/1096-9888(200007)35:7<784::AID-JMS17>3.0.CO;2-Q.

(13) Alymatiri, C. M.; Kouskoura, M. G.; Markopoulou, C. K. Decoding the Signal Response of Steroids in Electrospray Ionization Mode (ESI-MS). *Anal Methods* **2015**, *7* (24), 10433–10444. https://doi.org/10.1039/C5AY02839F.

(14) Ehrmann, B. M.; Henriksen, T.; Cech, N. B. Relative Importance of Basicity in the Gas Phase and in Solution for Determining Selectivity in Electrospray Ionization Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (5), 719–728. https://doi.org/10.1016/j.jasms.2008.01.003.

(15) Liigand, J.; Wang, T.; Kellogg, J.; Smedsgaard, J.; Cech, N.; Kruve, A. Quantification for Non-Targeted LC/MS Screening without Standard Substances. *Sci. Rep.* **2020**, *10* (1), 5808. https://doi.org/10.1038/s41598-020-62573-z.

(16) Kruve, A.; Kiefer, K.; Hollender, J. Benchmarking of the Quantification Approaches for the Non-Targeted Screening of Micropollutants and Their Transformation Products in Groundwater. *Anal. Bioanal. Chem.* **2021**. https://doi.org/10.1007/s00216-020-03109-2.

(17) Rebane, R.; Oldekop, M.-L.; Herodes, K. Comparison of Amino Acid Derivatization Reagents for LC–ESI-MS Analysis. Introducing a Novel Phosphazene-Based Derivatization Reagent. *J. Chromatogr. B* **2012**, *904*, 99–106. https://doi.org/10.1016/j.jchromb.2012.07.029.

(18) Rebane, R.; Rodima, T.; Kütt, A.; Herodes, K. Development of Amino Acid Derivatization Reagents for Liquid Chromatography Electrospray Ionization Mass Spectrometric Analysis and Ionization Efficiency Measurements. *J. Chromatogr. A* **2015**, *1390*, 62–70. https://doi.org/10.1016/j.chroma.2015.02.050.

(19) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine

Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(20) Bjerrum, E.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8* (4), 131. https://doi.org/10.3390/biom8040131.

(21) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep Learning for Molecular Design—a Review of the State of the Art. *Mol. Syst. Des. Eng.* **2019**, *4* (4), 828–849. https://doi.org/10.1039/C9ME00039A.

(22) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for *De Novo* Drug Design. *Mol. Inform.* **2018**, *37* (1–2), 1700111. https://doi.org/10.1002/minf.201700111.

(23) Jensen, J. H. A Graph-Based Genetic Algorithm and Generative Model/Monte Carlo Tree Search for the Exploration of Chemical Space. *Chem. Sci.* **2019**, *10* (12), 3567–3572. https://doi.org/10.1039/C8SC05372C.

(24) Leguy, J.; Cauchy, T.; Glavatskikh, M.; Duval, B.; Da Mota, B. EvoMol: A Flexible and Interpretable Evolutionary Algorithm for Unbiased de Novo Molecular Generation. *J. Cheminformatics* **2020**, *12* (1), 55. https://doi.org/10.1186/s13321-020-00458-z.

(25) Lim, J.; Hwang, S.-Y.; Moon, S.; Kim, S.; Kim, W. Y. Scaffold-Based Molecular Design with a Graph Generative Model. *Chem. Sci.* **2020**, *11* (4), 1153–1164. https://doi.org/10.1039/C9SC04503A.

(26) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for de Novo Drug Design – A Survey. *Appl. Soft Comput.* **2015**, *27*, 543–552. https://doi.org/10.1016/j.asoc.2014.09.042.

(27) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminformatics* **2018**, *10* (1), 31. https://doi.org/10.1186/s13321-018-0286-7.

(28) 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. https://doi.org/10.1021/acscentsci.7b00572.

(29) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik,

A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *ArXiv170510843 Cs Stat* **2018**.

(30) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4* (7), eaap7885. https://doi.org/10.1126/sciadv.aap7885.

(31) Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; Renslow, R. S. ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries. *Anal. Chem.* **2019**, *91* (7), 4346–4356. https://doi.org/10.1021/acs.analchem.8b04567.

(32) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. *Mol. Pharm.* **2018**, *15* (10), 4398–4405. https://doi.org/10.1021/acs.molpharmaceut.8b00839.

(33) Colby, S. M.; Nuñez, J. R.; Hodas, N. O.; Corley, C. D.; Renslow, R. R. Deep Learning to Generate *in Silico* Chemical Property Libraries and Candidate Molecules for Small Molecule Identification in Complex Samples. *Anal. Chem.* **2020**, *92* (2), 1720–1729. https://doi.org/10.1021/acs.analchem.9b02348.

(34) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: An Efficient Python Library for *de Novo* Molecular Generation. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 972–976. https://doi.org/10.1080/14686996.2017.1401424.

(35) Kajita, S.; Kinjo, T.; Nishi, T. Autonomous Molecular Design by Monte-Carlo Tree Search and Rapid Evaluations Using Molecular Dynamics Simulations. *Commun. Phys.* **2020**, *3* (1), 77. https://doi.org/10.1038/s42005-020-0338-y.

(36) PubChem Database https://pubchem.ncbi.nlm.nih.gov/.

(37) Gornischeff, A.; Liigand, J.; Rebane, R. A Systematic Approach toward Comparing Electrospray Ionization Efficiencies of Derivatized and Non-Derivatized Amino Acids and Biogenic Amines. *J. Mass Spectrom.* **2018**. https://doi.org/10.1002/jms.4272.

(38) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32* (7), 1466–1474. https://doi.org/10.1002/jcc.21707.

(39) Annilo, H. Keemilsie Analüüsi Tulemsute Kvantiseerimine Tarkvarateenusena Quantem Analytics OÜ Näitel.

Licence

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

1. Mina, Anneli Kruve-Viil, annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Monte Carlo tree search in designing of high sensitivity derivatization reagents for mass spectrometric analysis, mille juhendaja on Dr. Meelis Kull, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

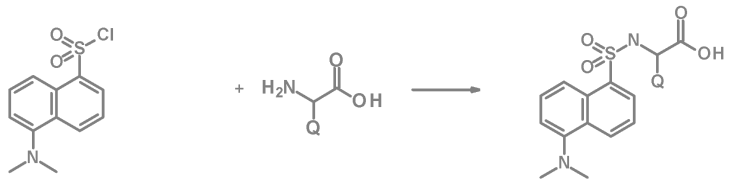
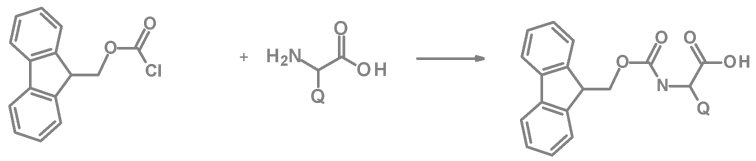
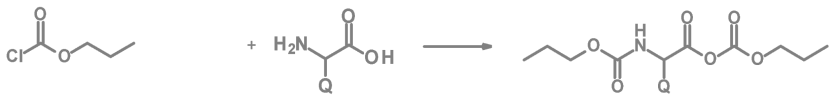
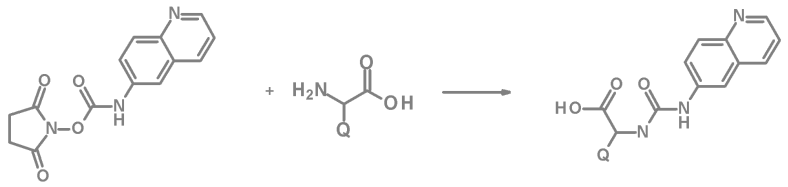
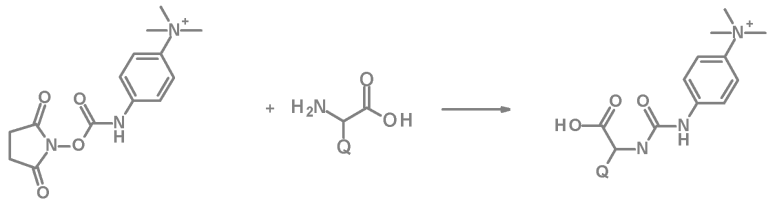
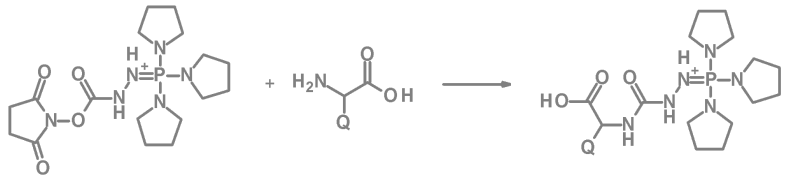
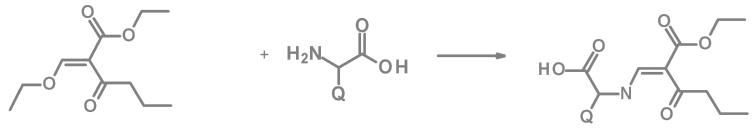
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Anneli Kruve-Viil

14.05.2021

Supplementary Information

Table S1 Overview of the derivatization reactions and reagents that have been previously used and designed for amino acid analysis with LC/ESI/MS.

Reagent	**Reaction**
DNS	
FMOC	
PrCl	
AQC	
TAHS	
FOSF	
DEEMM	

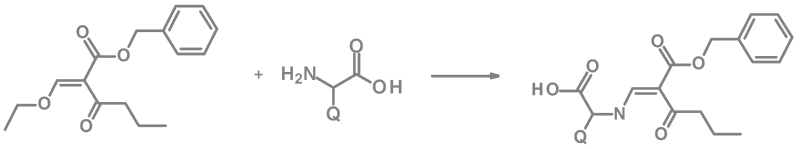
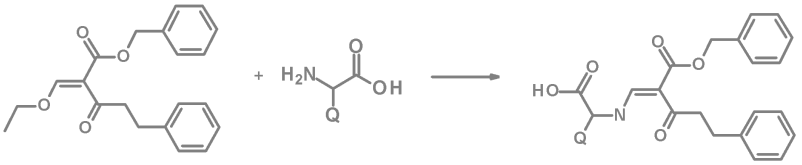
EDEMM	
DBEMM	

Table S2 *p*-values for comparison of the mean ionization efficiency values predicted for amino acids derivatized with reagents obtained with different MCTS algorithms. Values above 0.05 are colored grey while statistically significant values are shown in scientific style and in black color.

Glycine					
	2^2	2^5	3^3	10^2	50^1
1^1	0.070	5.03E-04	4.81E-06	8.23E-10	2.53E+10
2^2		0.136	8.75E-03	2.44E-05	1.03E-05
2^5			0.188	1.23E-03	5.29E-04
3^3				0.0527	2.93E-02
10^2					0.804
Glutamic acid					
1^1	2.30E-03	1.20E-03	3.13E-06	2.29E-15	2.20E-15
2^2		0.760	2.40E-02	1.10E-06	2.67E-10
2^5			0.058	1.18E-05	7.72E-09
3^3				1.45E-02	5.99E-05
10^2					4.71E-02
Argnine					
1^1	9.72E-03	1.09E-04	2.67E-07	8.11E-14	5.53E-15
2^2		0.224	1.28E-02	9.16E-07	1.39E-07
2^5			0.184	6.09E-05	9.31E-06
3^3				4.49E-03	9.40E-04
10^2					0.665
Phenylalanine					
1^1	2.84E-03	2.96E-05	8.07E-08	2.20E-16	2.20E-16
2^2		0.284	2.07E-02	6.77E-08	1.60E-08
2^5			0.193	6.75E-06	1.72E-06
3^3				2.10E-03	8.74E-04
10^2					0.890

Table S3 *p*-values for comparison of the mean ionization efficiency values predicted for amino acids derivatized with reagents obtained with different MCTS algorithms after addition of four functional groups. *p*-values above 0.05 are colored grey while statistically significant values are shown in scientific style and in black color.

Glycine					
	2^2	2^5	3^3	10^2	50^1
1^1	1.19E-02	7.94E-03	6.09E-03	5.61E-15	6.25E-15
2^2		0.920	0.850	2.66E-06	1.11E-06
2^5			0.929	3.06E-06	1.28E-06
3^3				4.70E-06	1.96E-06
10^2					0.644
Glutamic acid					
1^1	1.21E-03	1.64E-05	1.44E-05	2.20E-16	2.20E-16
2^2		0.503	0.435	9.71E-07	1.40E-13
2^5			0.893	2.44E-06	3.71E-14
3^3				6.83E-06	2.04E-13
10^2					2.45E-04
Argnine					
1^1	3.63E-03	3.77E-04	6.82E-07	2.20E-16	2.20E-16
2^2		0.480	0.070	6.44E-07	1.04E-09
2^5			0.301	3.45E-05	1.25E-07
3^3				6.23E-04	1.85E-06
10^2					0.095
Phenylalanine					
1^1	0.463	1.19E-03	9.22E-05	2.20E-16	2.20E-16
2^2		3.48E-02	7.36E-03	1.00E-11	1.14E-13
2^5			0.547	1.52E-07	1.37E-09
3^3				2.12E-06	2.07E-08
10^2					0.315

Figure S1 Visualization of each of the baseline 1[1] search events for glycine.

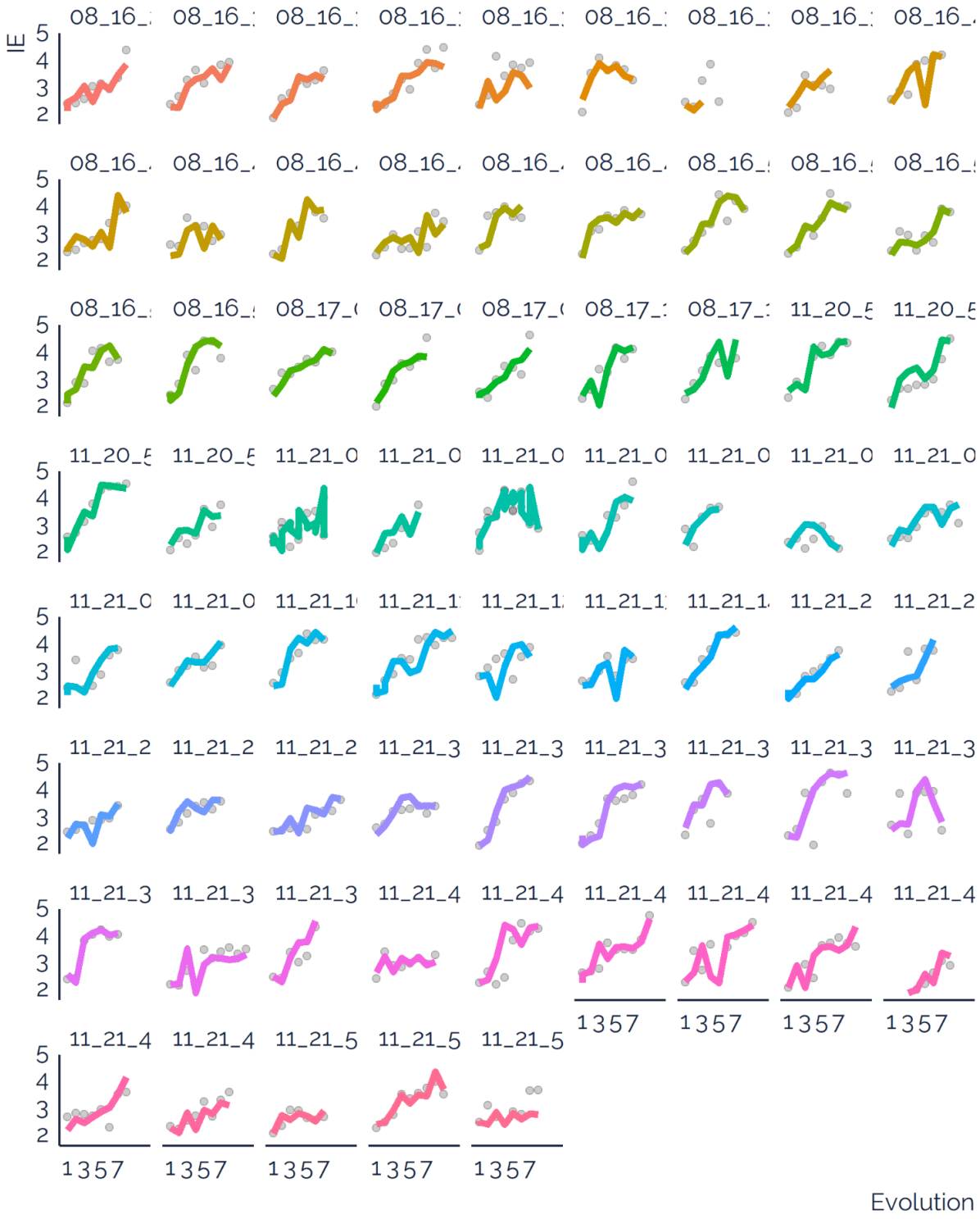


Figure S2 Visualization of each of the baseline 1[1] search events for glutamic acid.

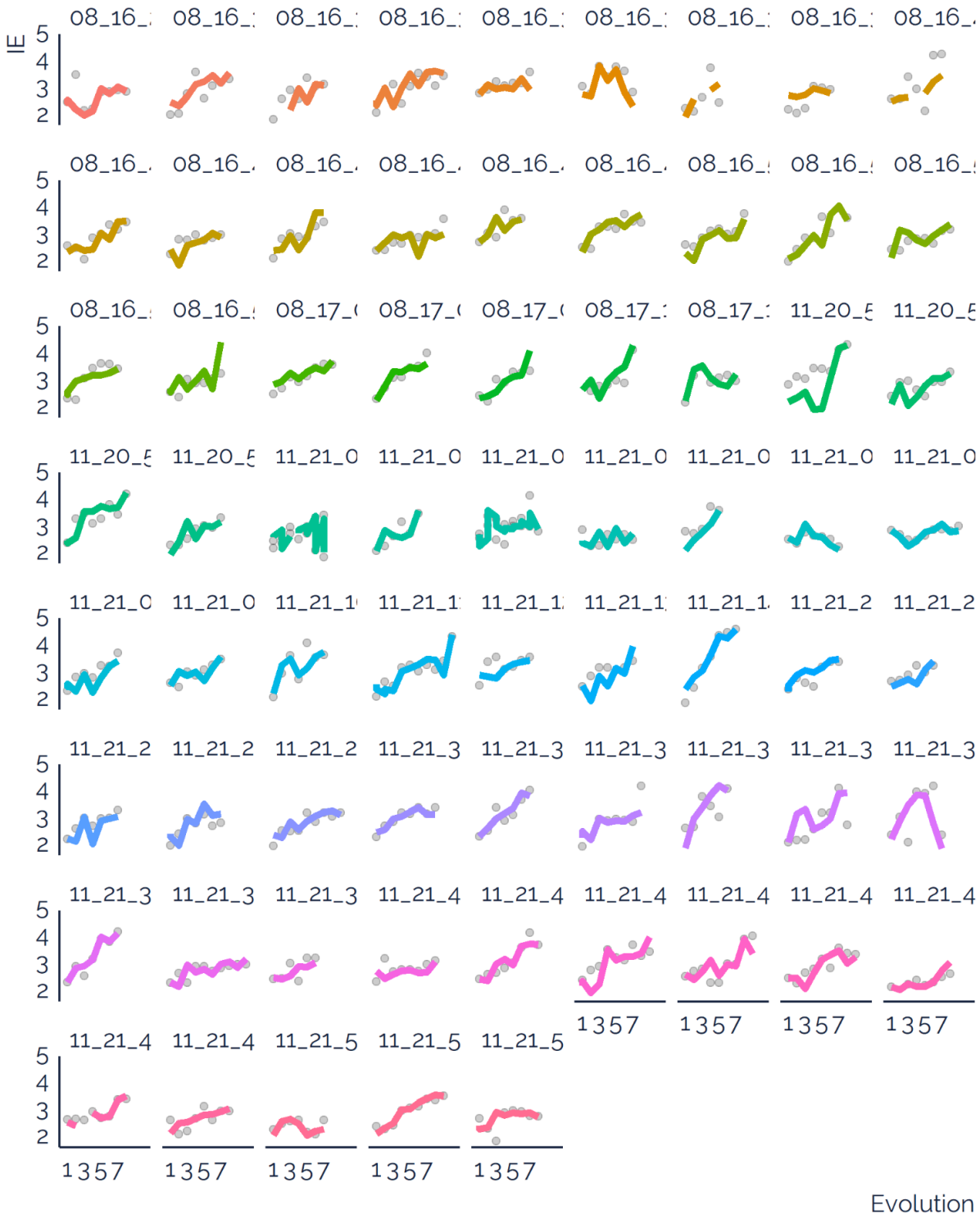


Figure S3 Visualization of each of the baseline 1[1] search events for arginine.

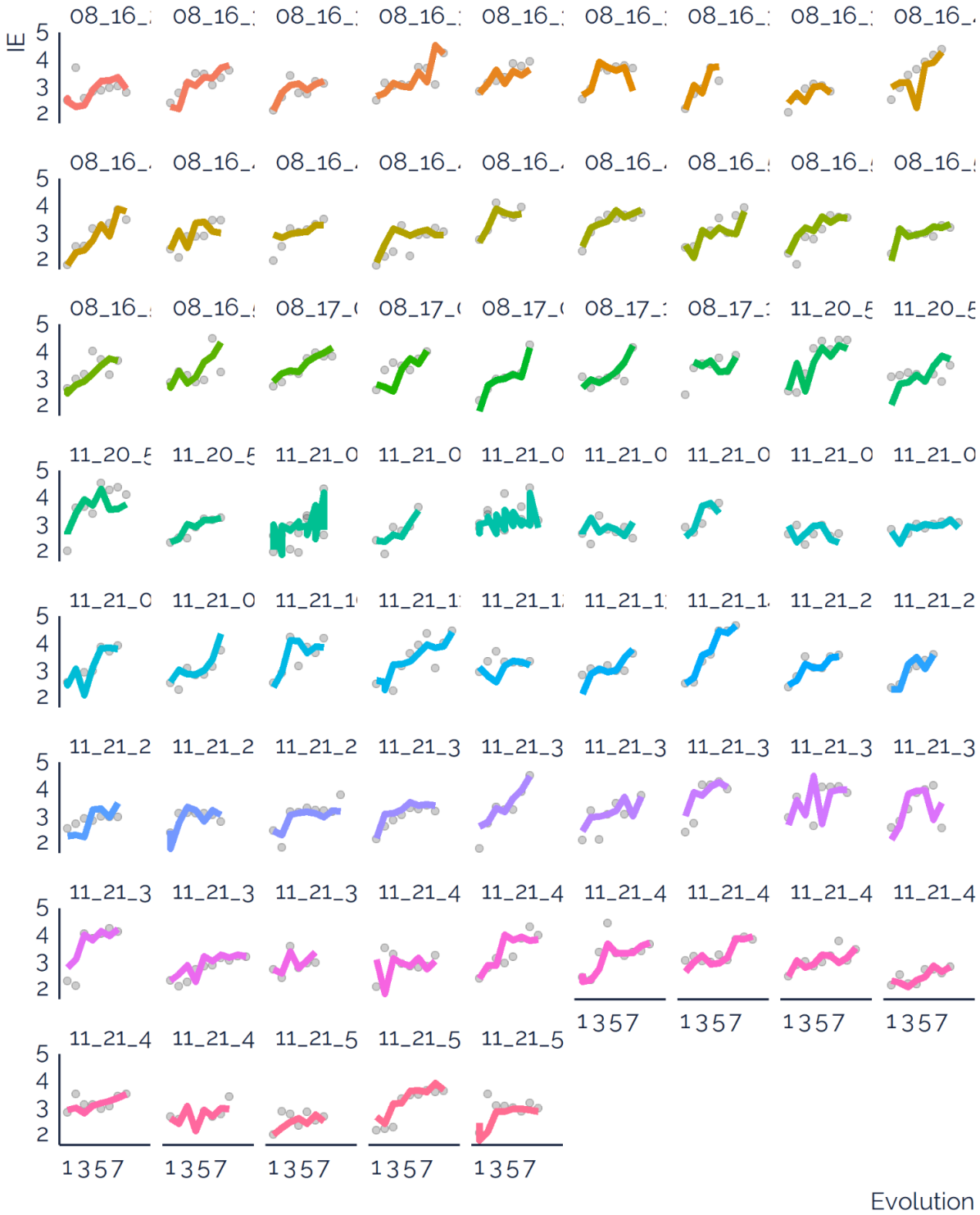


Figure S4 Visualization of each of the baseline 1^1 search events for phenylalanine.

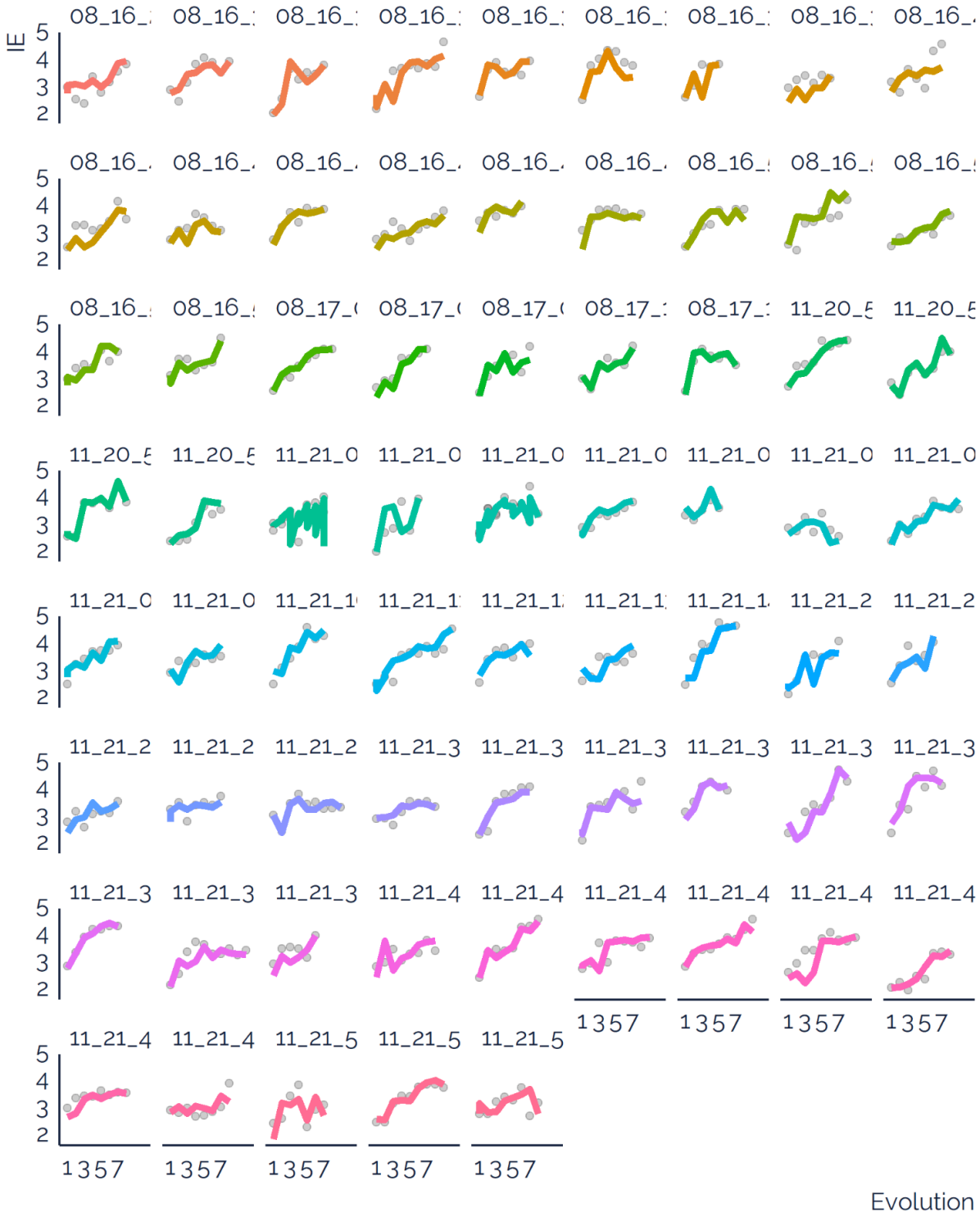


Figure S5 Visualization of each of the 2^2 search events for glycine.

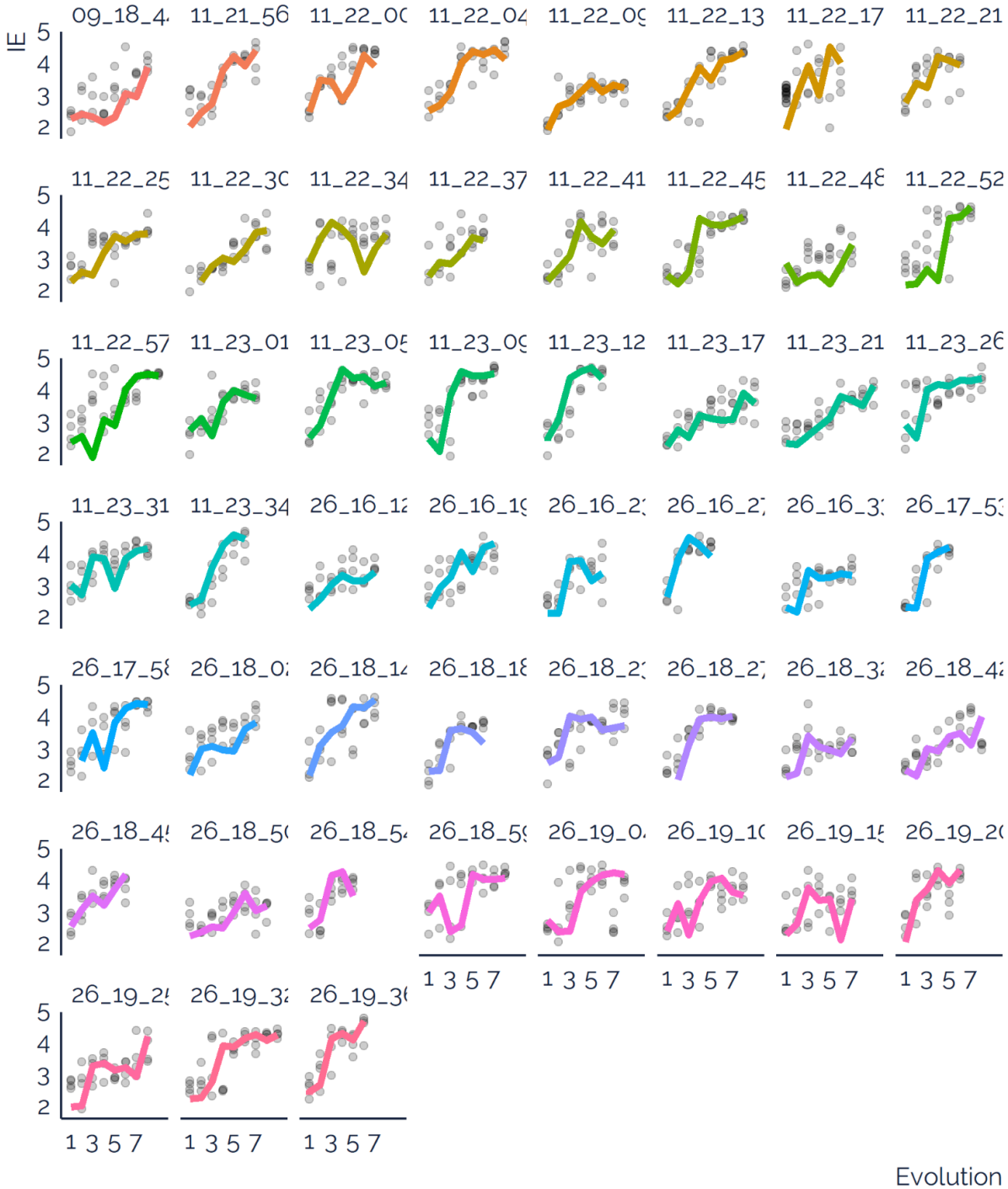


Figure S6 Visualization of each of the 2^2 search events for glutamic acid.

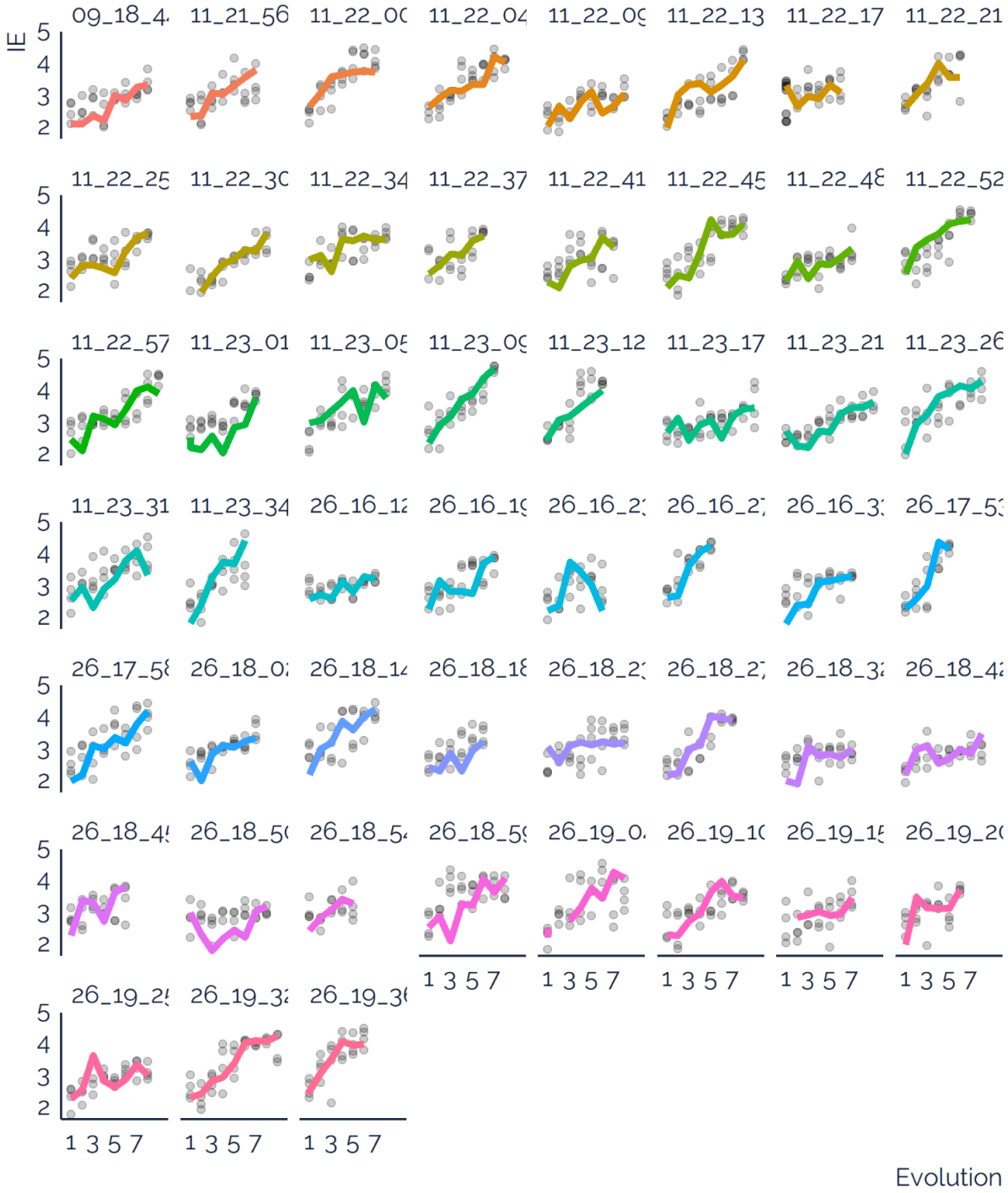


Figure S7 Visualization of each of the 2^2 search events for arginine.

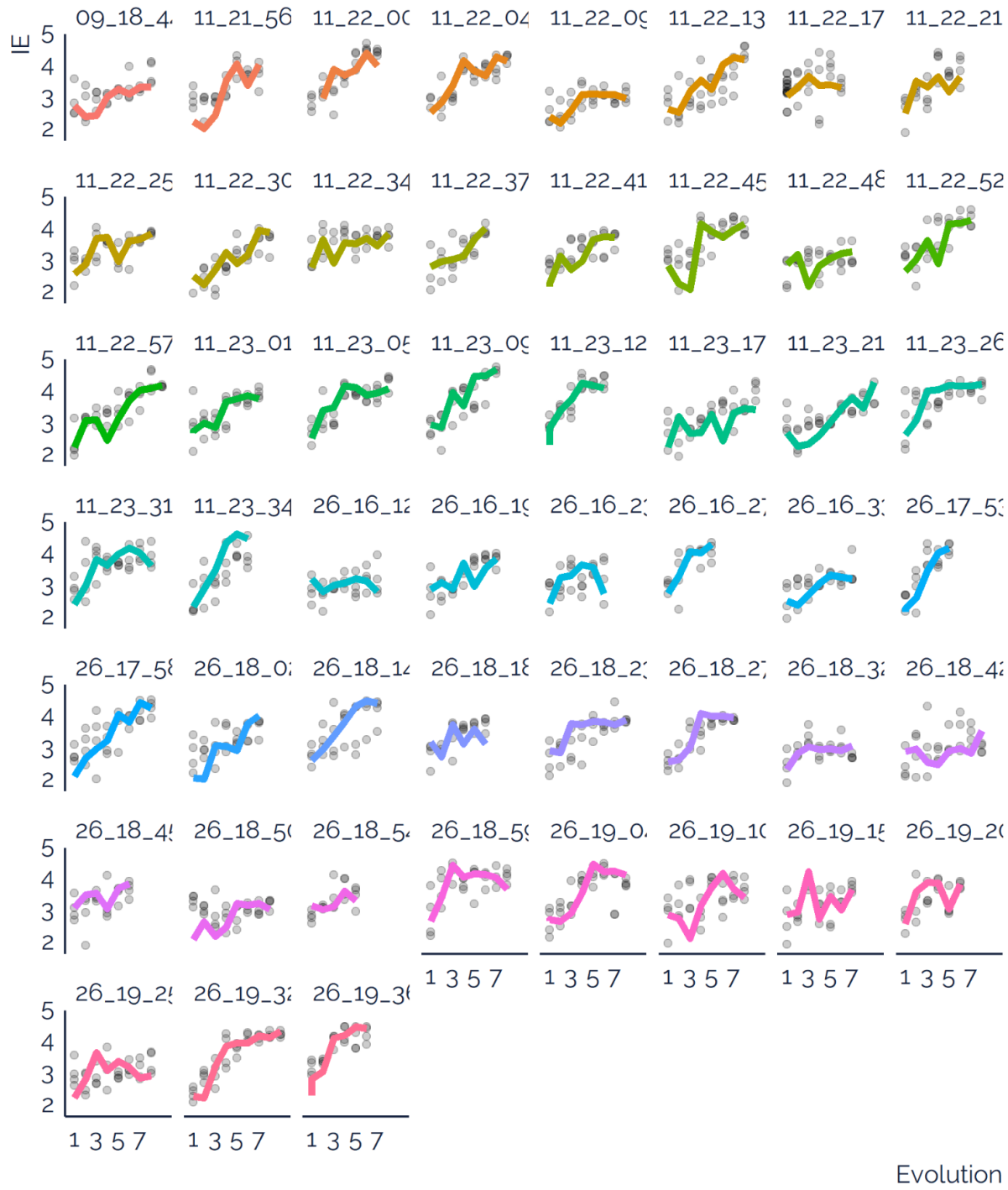


Figure S8 Visualization of each of the 2^2 search events for phenylalanine.

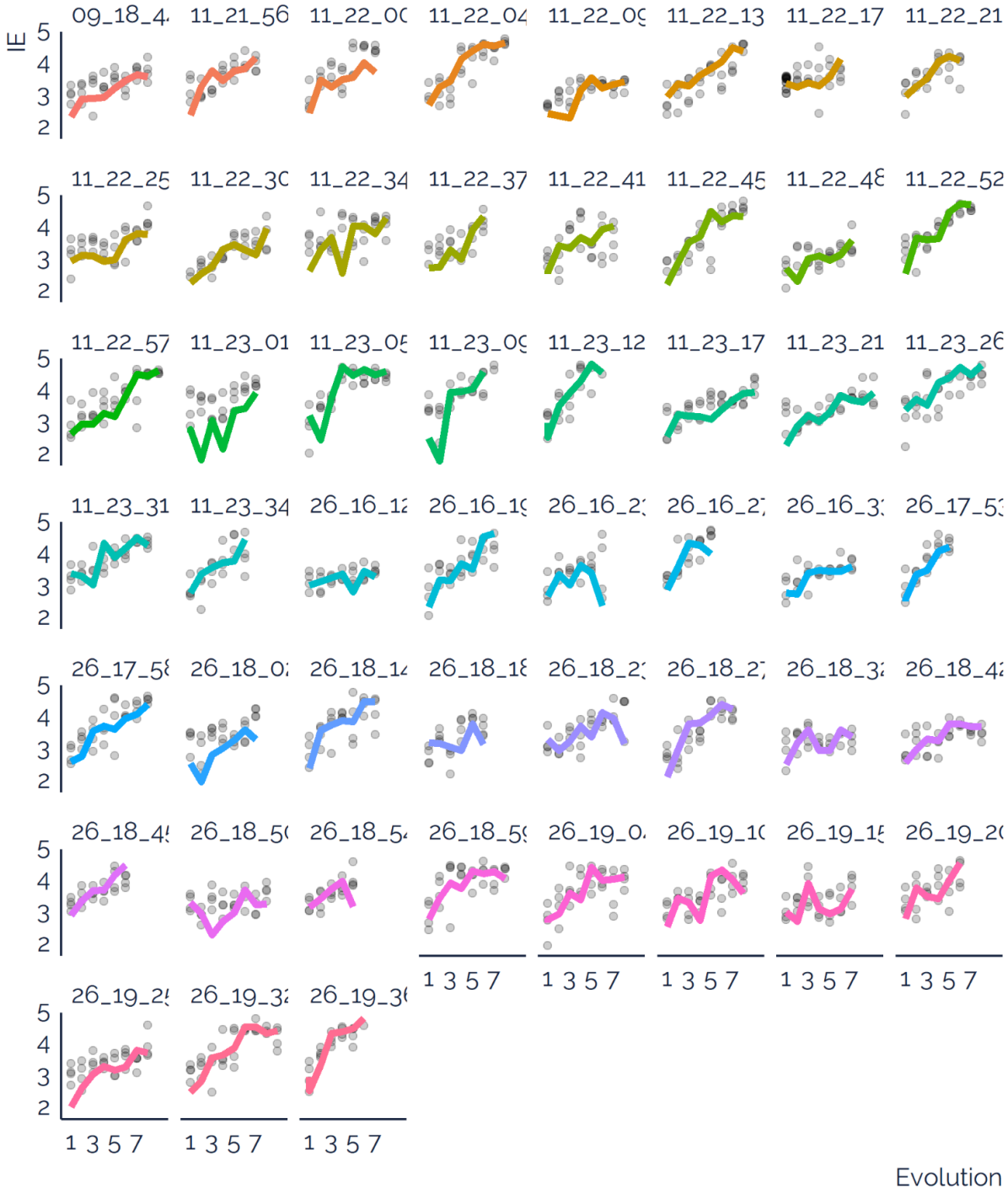


Figure S9 Visualization of each of the 2^5 search events for glycine.

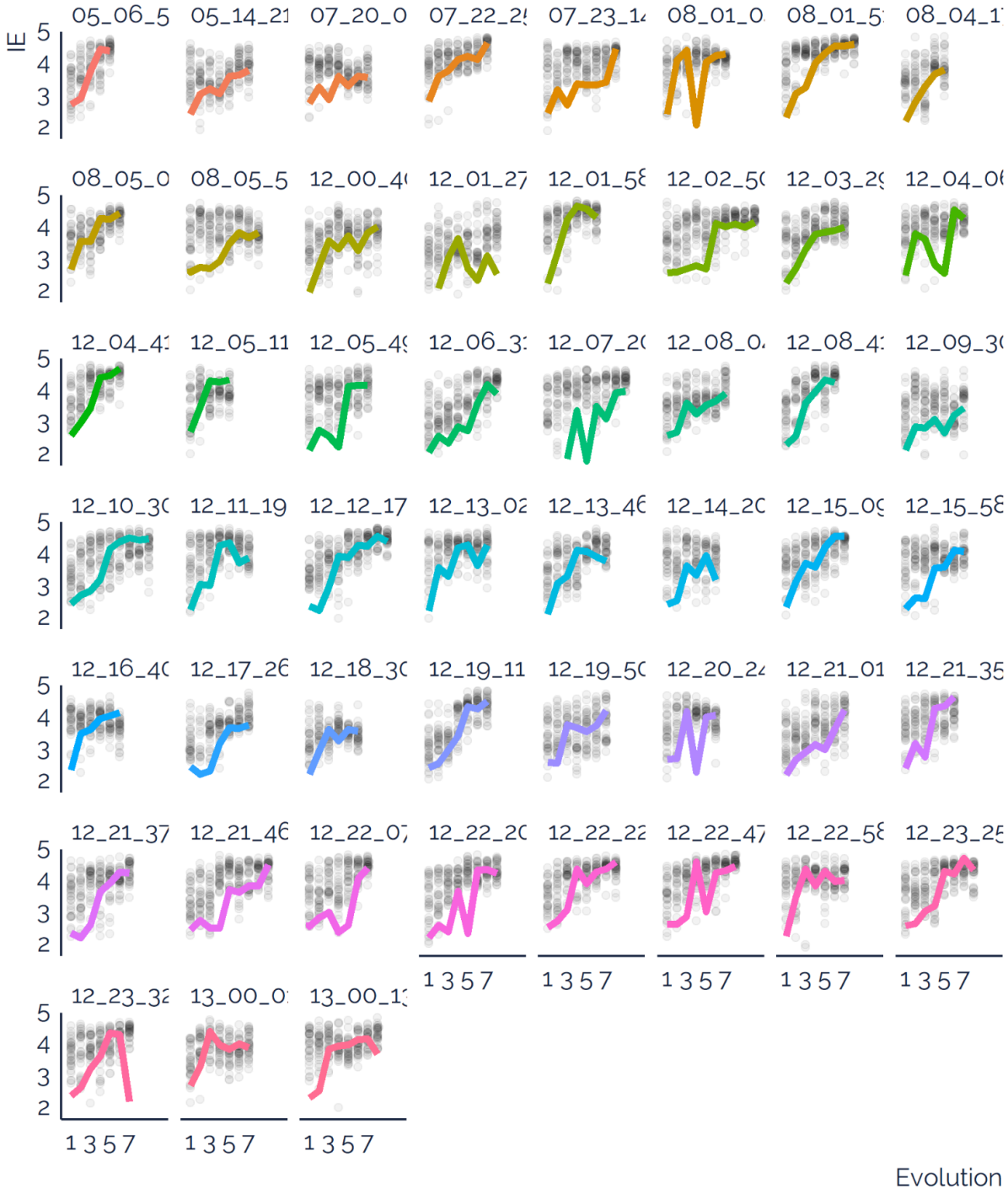


Figure S10 Visualization of each of the 2^5 search events for glutamic acid.

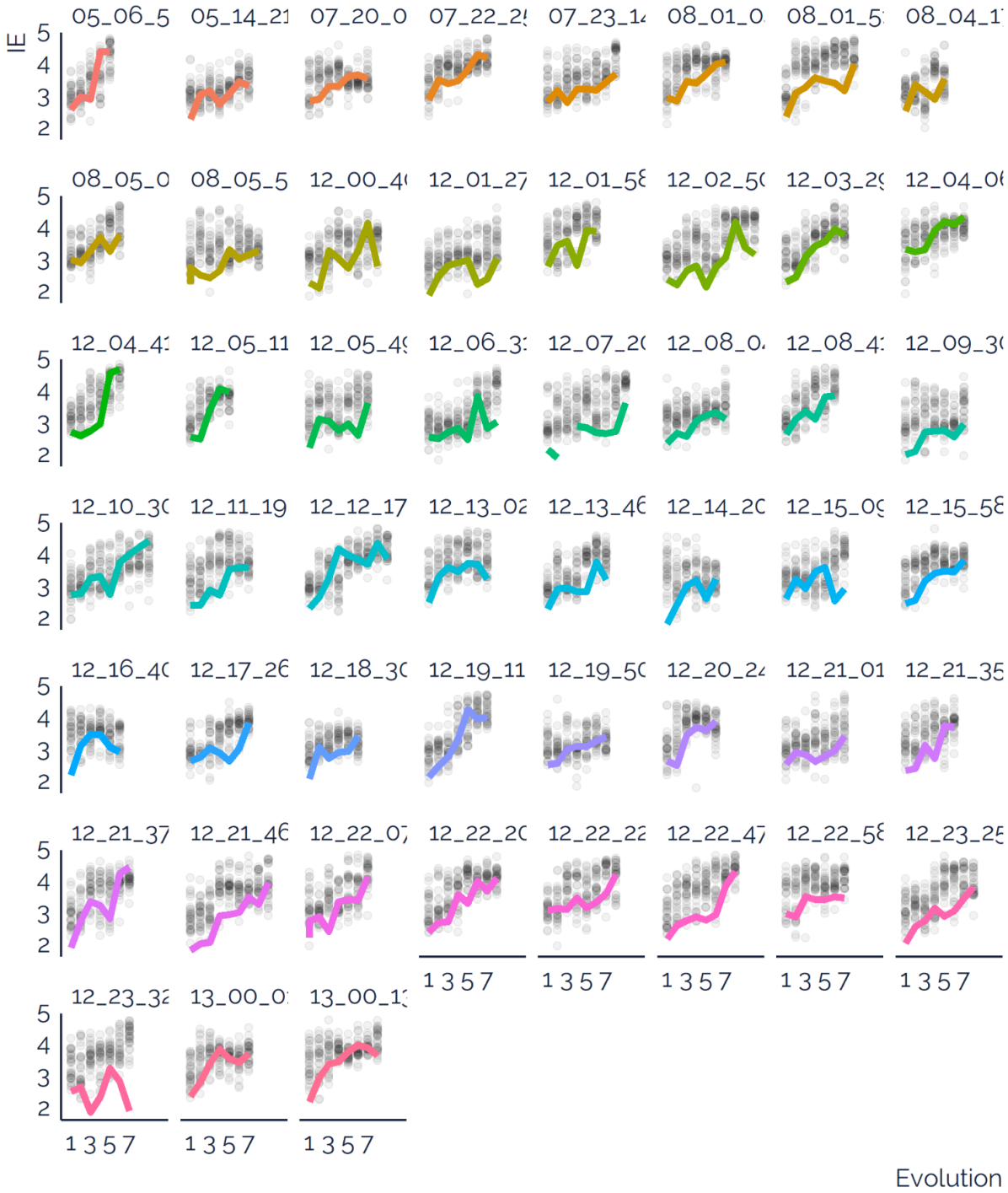


Figure S11 Visualization of each of the 2^5 search events for arginine.

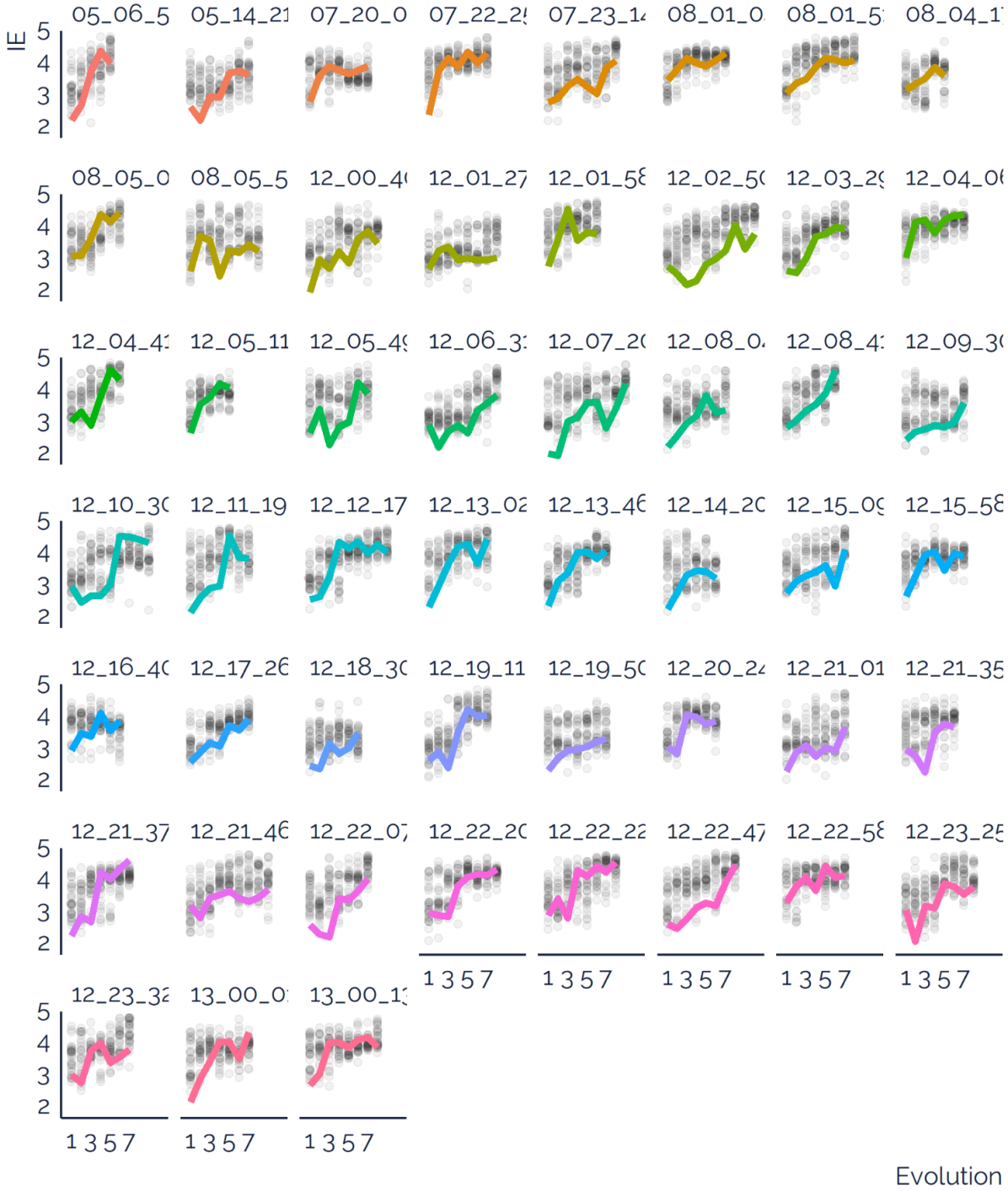


Figure S12 Visualization of each of the 2^5 search events for phenylalanine.

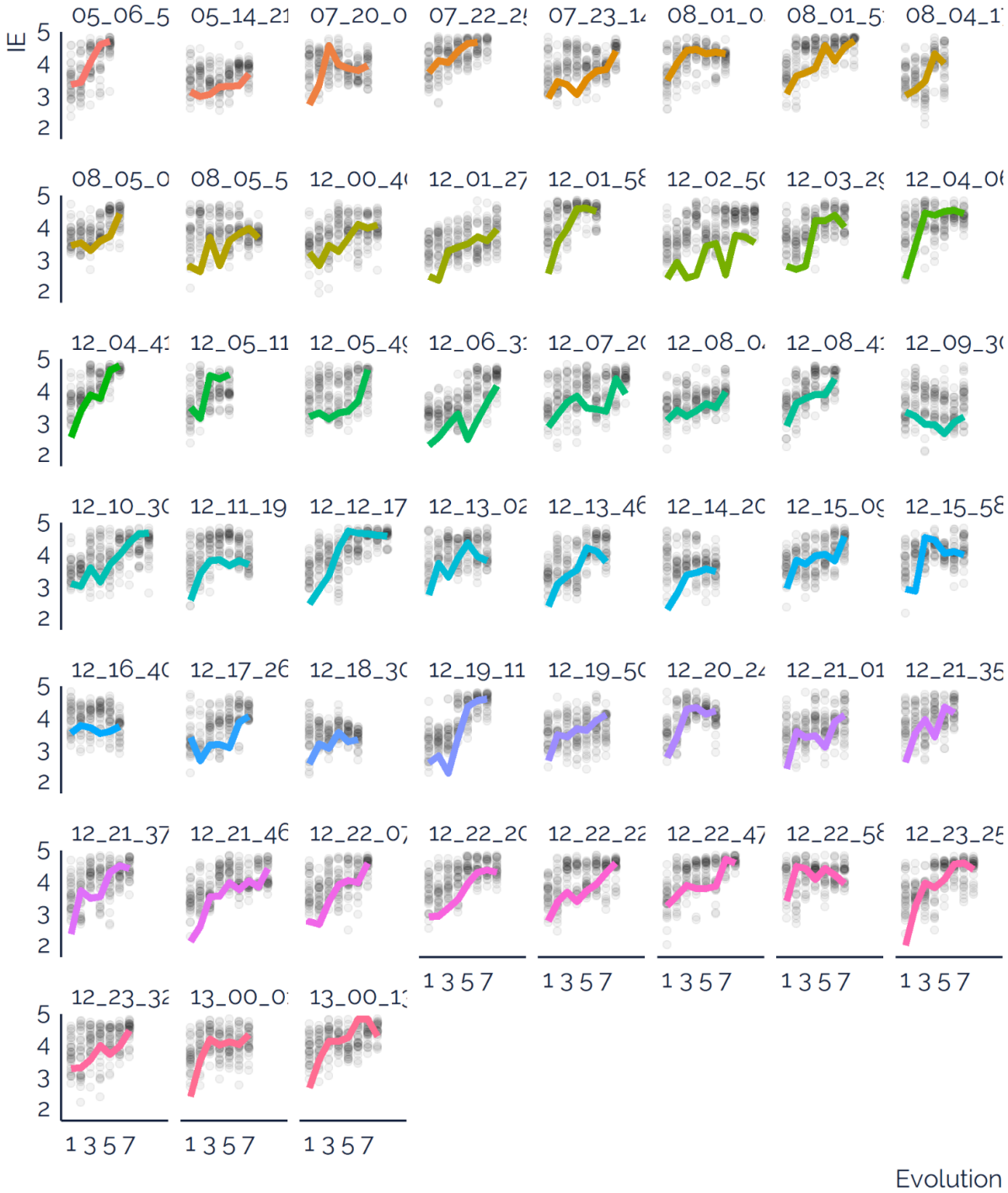


Figure S13 Visualization of each of the 3^3 search events for glycine.

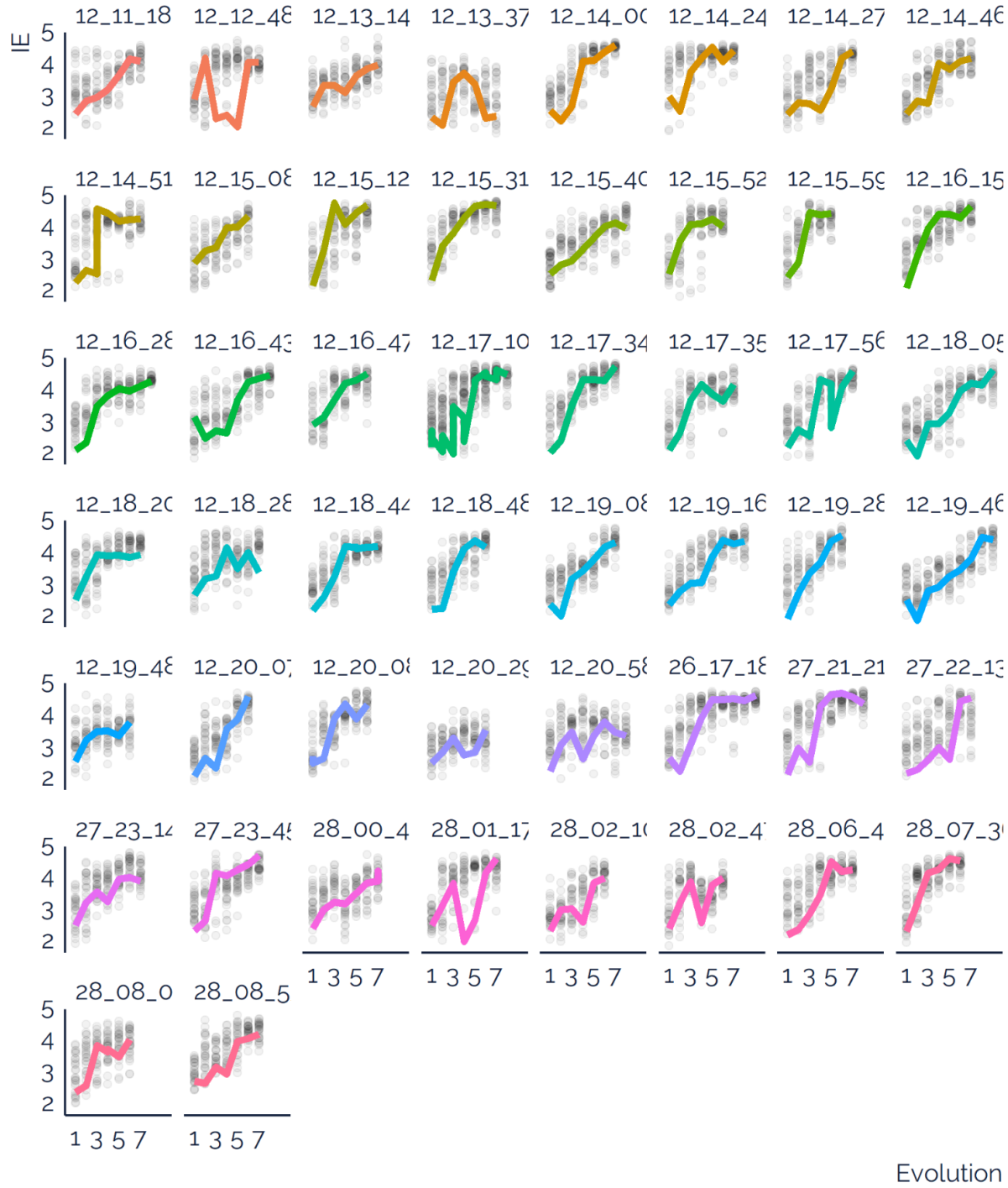


Figure S14 Visualization of each of the 3^3 search events for glutamic acid.

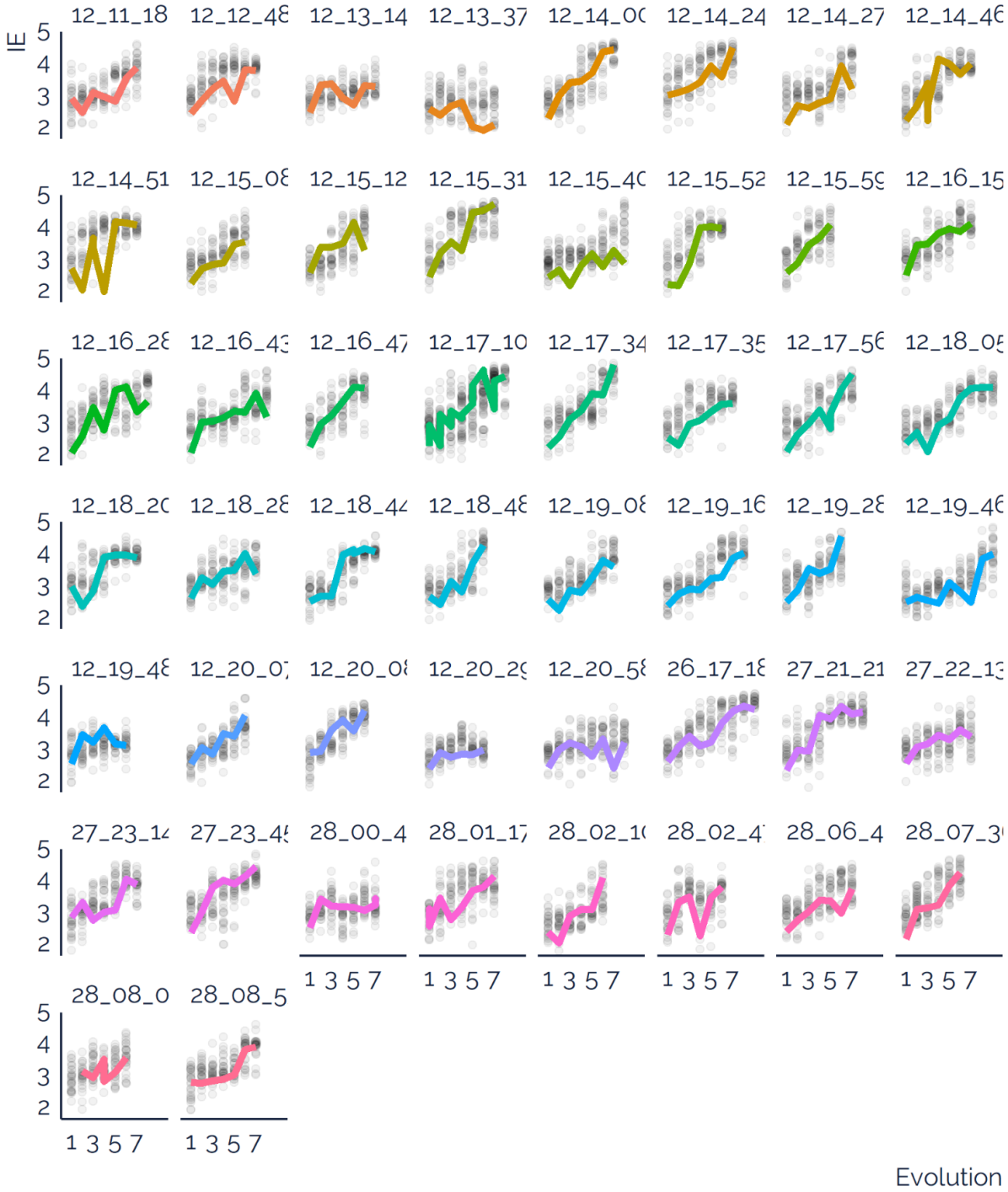


Figure S15 Visualization of each of the 3^3 search events for arginine.

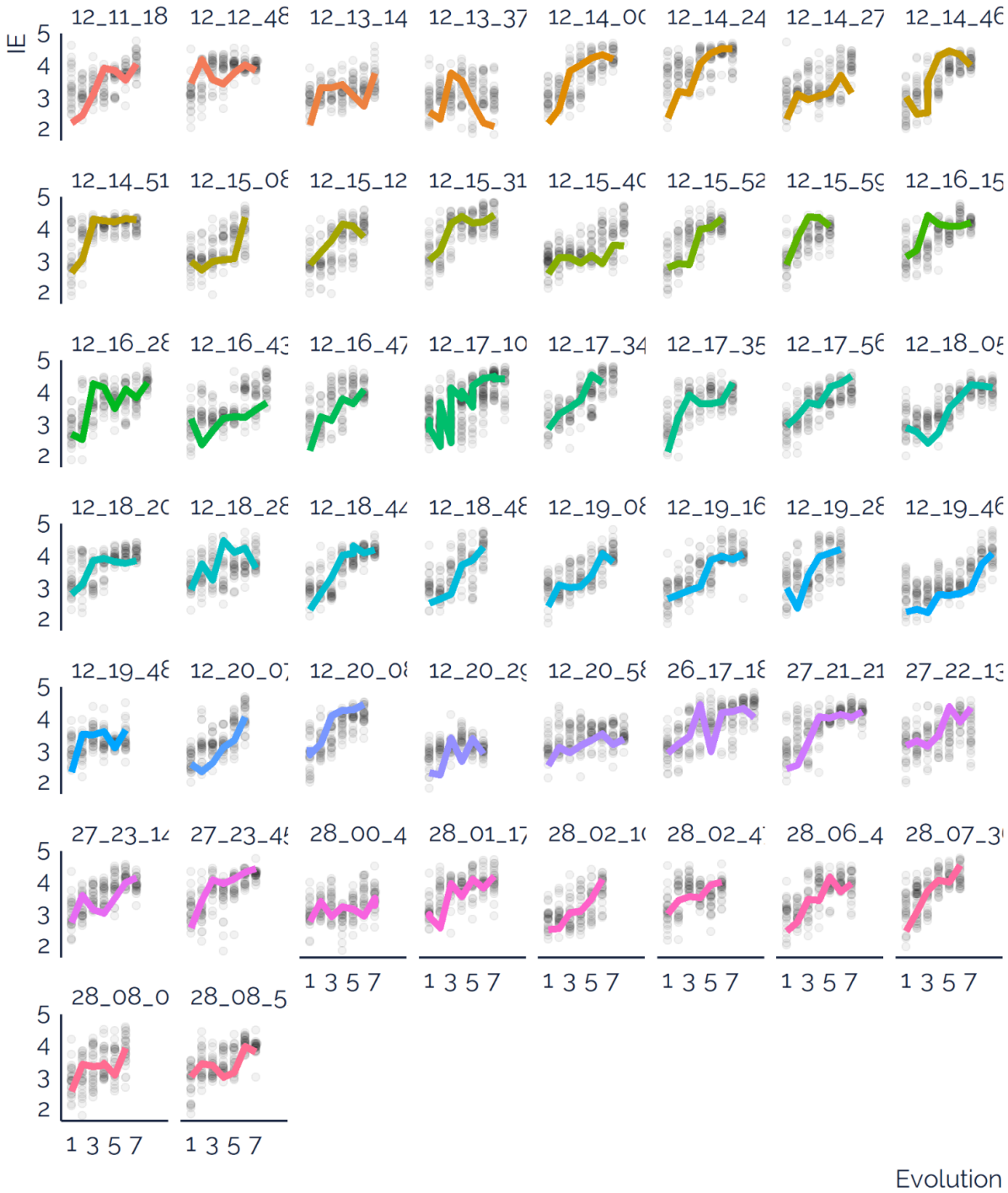


Figure S16 Visualization of each of the 3^3 search events for phenylalanine.

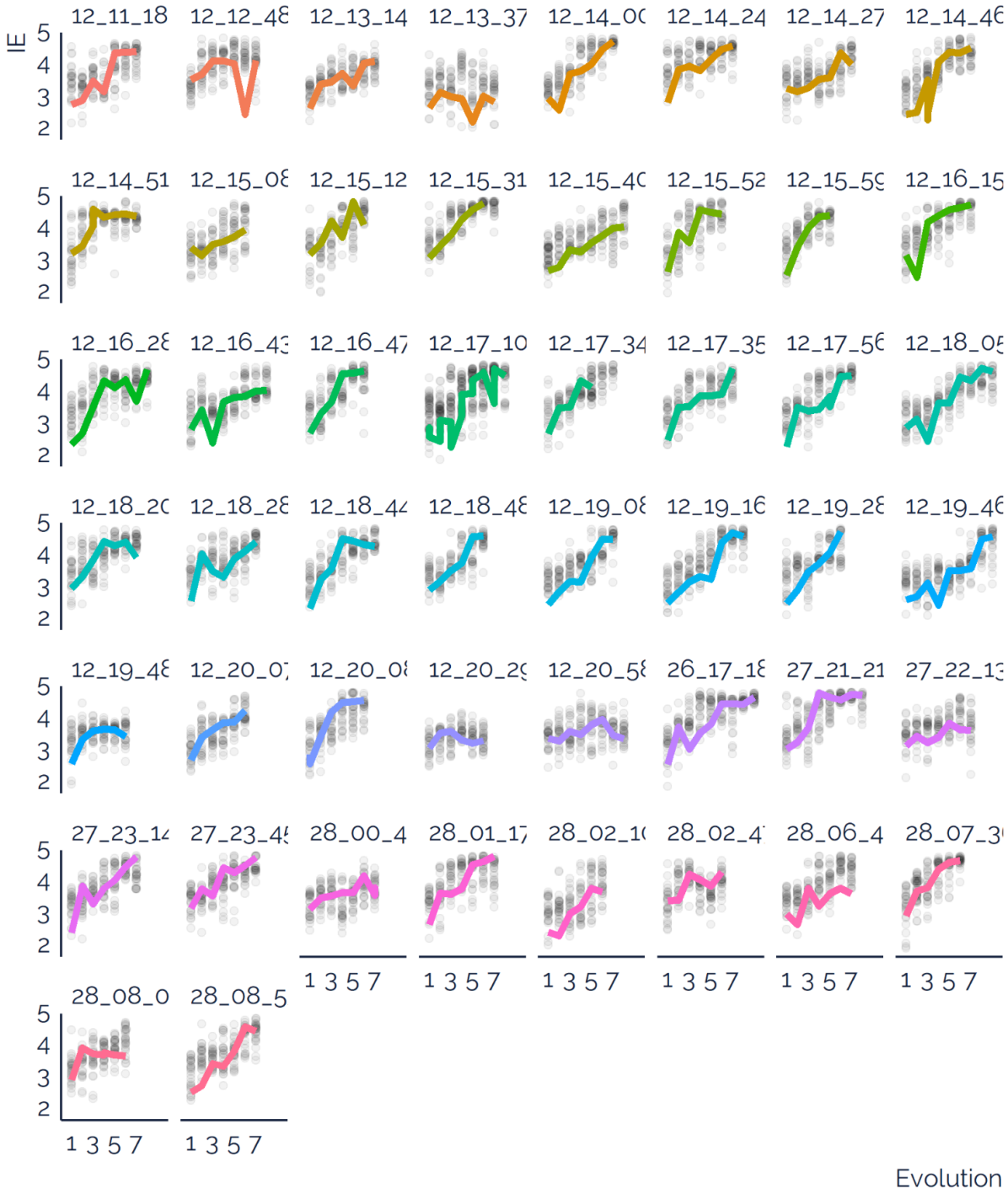


Figure S17 Visualization of each of the 10^2 search events for glycine.

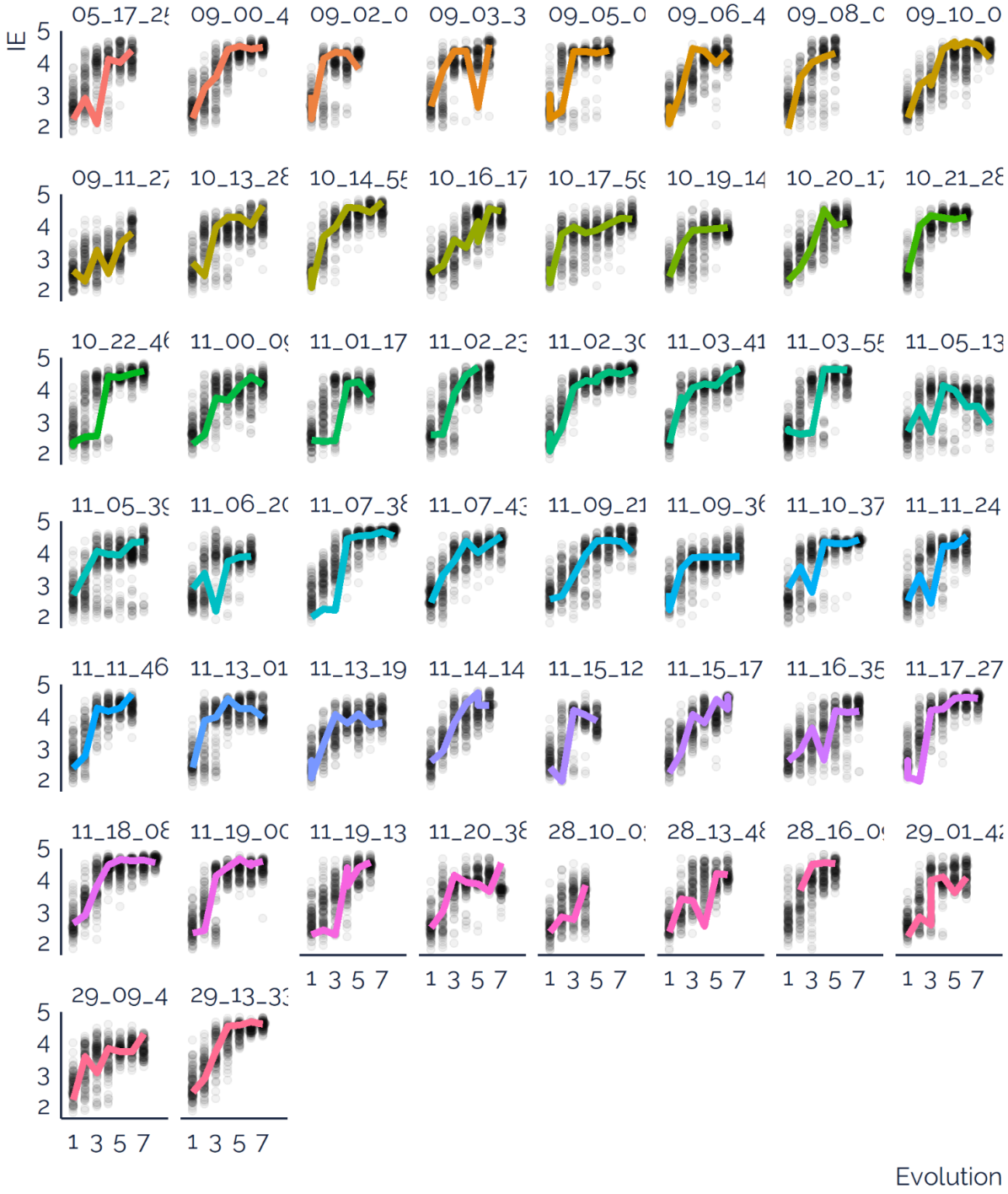


Figure S18 Visualization of each of the 10^2 search events for glutamic acid.

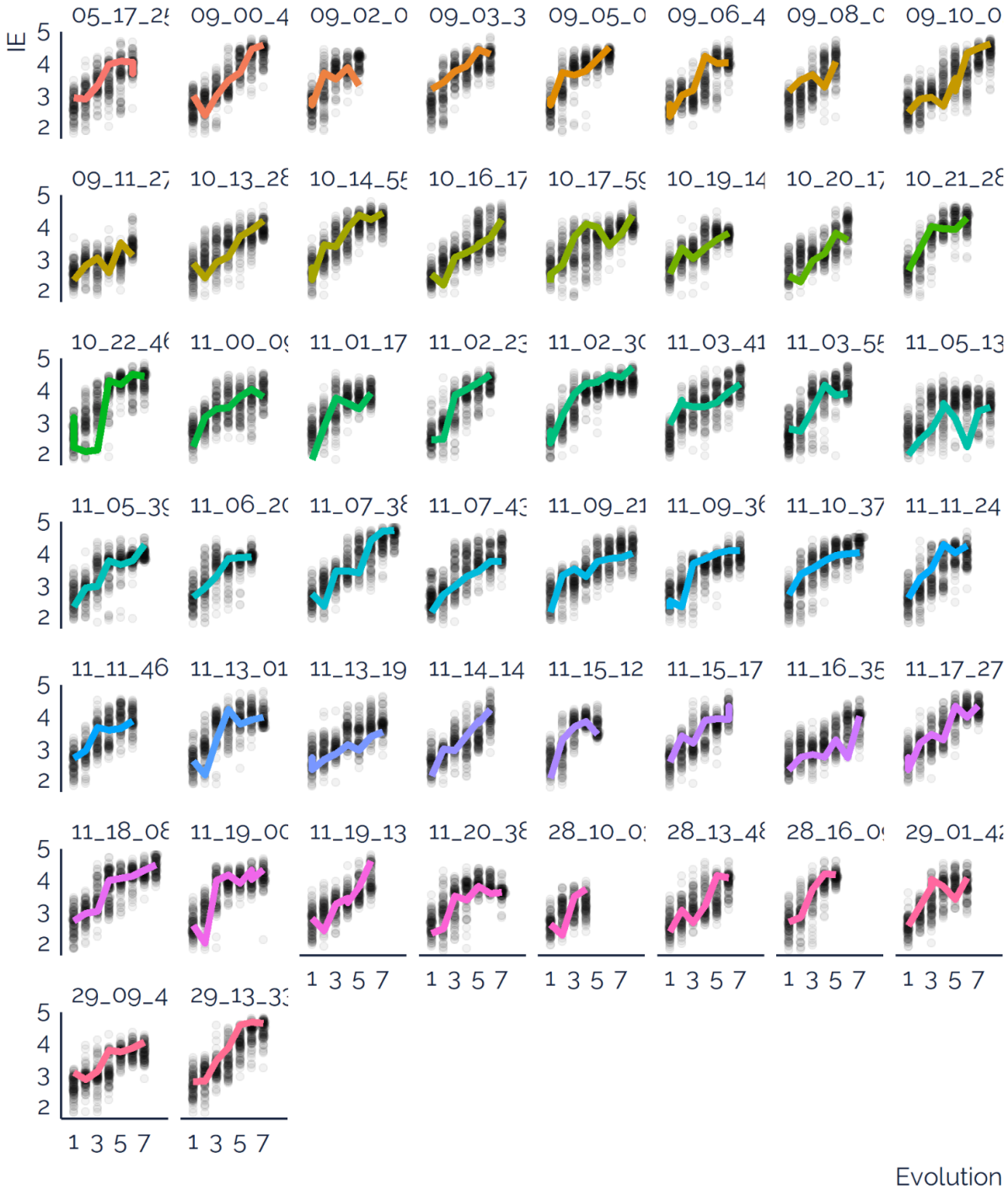


Figure S19 Visualization of each of the 10^2 search events for arginine.

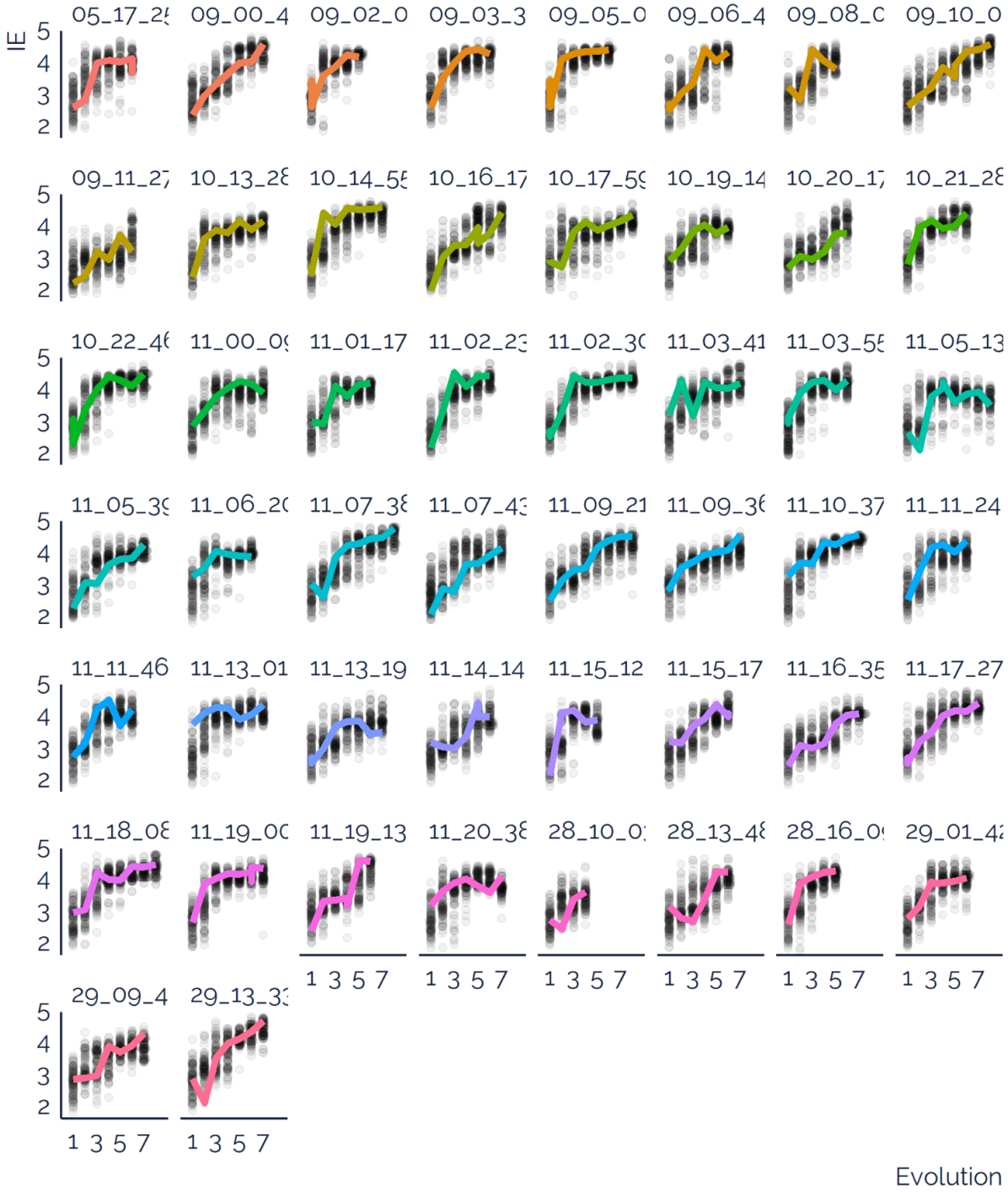


Figure S20 Visualization of each of the 10^2 search events for phenylalanine.

Figure S21 Visualization of each of the 50[1] search events for glycine.

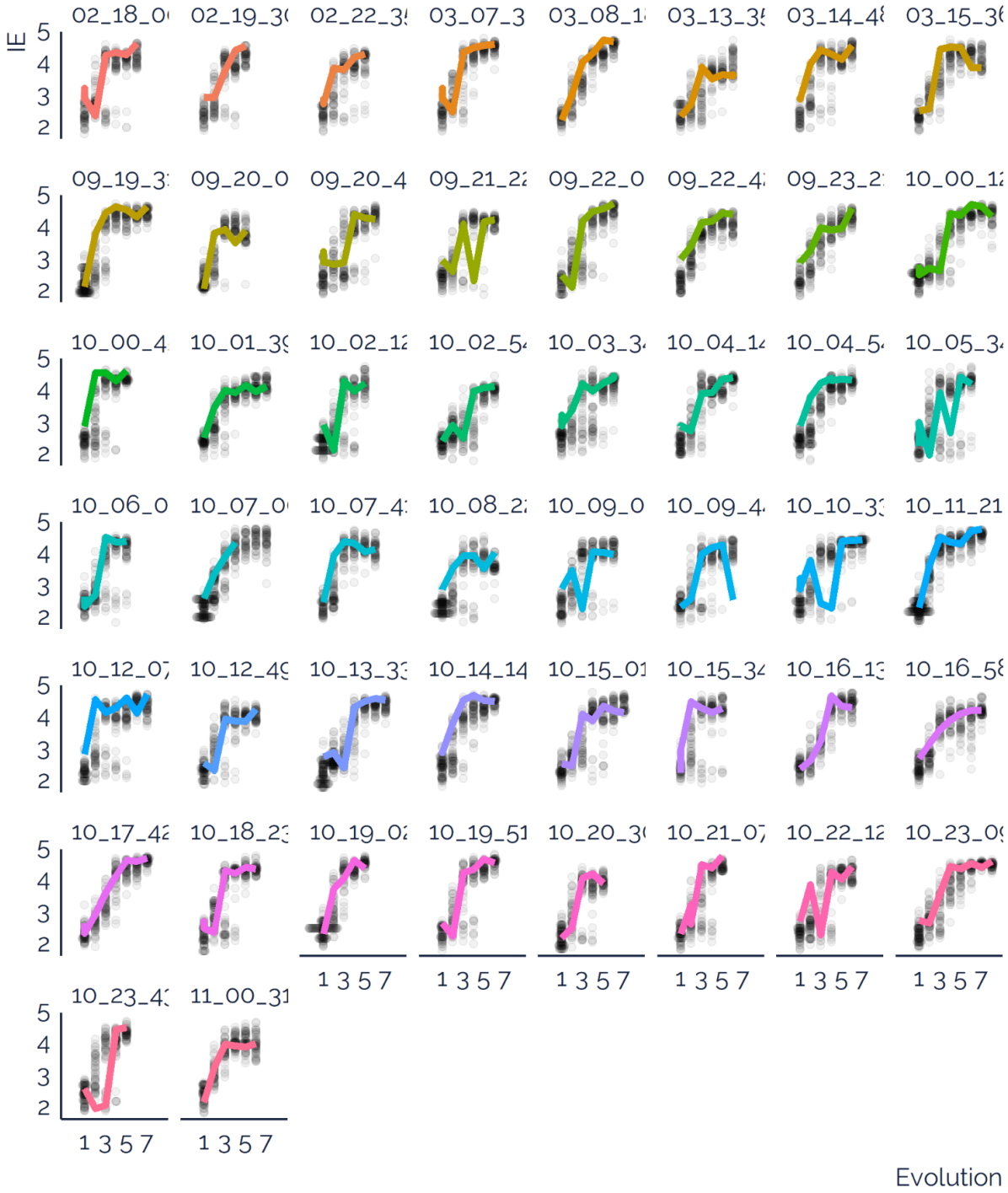


Figure S22 Visualization of each of the 50[1] search events for glutamic acid.

Figure S23 Visualization of each of the 50^1 search events for arginine.

Figure S24 Visualization of each of the 50^2 search events for phenylalanine.