LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-703546

Multivariate Statistical Analysis of Orthogonal Mass Spectral Data for the Identification of Chemical Attribution Signatures of 3-Methylfentanyl

B. P. Mayer, C. A. Valdez, A. J. DeHope, P. E. Spackman, R. D. Sanner, H. P. Martinez, A. M. Williams

September 22, 2016

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Multivariate Statistical Analysis of Orthogonal Mass Spectral Data for the Identification of Chemical Attribution Signatures of 3-Methylfentanyl

Brian P. Mayer, Carlos A. Valdez, Alan J. DeHope,

Paul E. Spackman, Audrey M. Williams

DHS Chemical Forensics Program

November 28th, 2016

PI: Audrey Williams
Forensic Science Center
Lawrence Livermore National Laboratory
williams259@llnl.gov, 925-423-4675

LLNL-TR-703546

Abstract

Critical to many modern forensic investigations is the chemical attribution of the origin of an illegal drug. This process greatly relies on identification of compounds indicative of its clandestine or commercial production. The results of these studies can yield detailed information on method of manufacture, sophistication of the synthesis operation, starting material source, and final product. In the present work, chemical attribution signatures (CAS) associated with the synthesis of the analgesic 3-methylfentanyl, *N*-(3-methyl-1-phenethylpiperidin-4-yl)-*N*-phenylpropanamide, were investigated. Six synthesis methods were studied in an effort to identify and classify route-specific signatures. These methods were chosen to minimize the use of scheduled precursors, complicated laboratory equipment, number of overall steps, and demanding reaction conditions. Using gas and liquid chromatographies combined with mass spectrometric methods (GC-QTOF and LC-QTOF) in conjunction with inductively-coupled plasma mass spectrometry (ICP-MS), over 240 distinct compounds and elements were monitored. As seen in our previous work with CAS of fentanyl synthesis the complexity of the resultant data matrix necessitated the use of multivariate statistical analysis. Using partial least squares discriminant analysis (PLS-DA), 62 statistically significant, route-specific CAS were identified. Statistical classification models using a variety of machine learning techniques were then developed with the ability to predict the method of 3-methylfentanyl synthesis from three blind crude samples generated by synthetic chemists without prior experience with these methods.

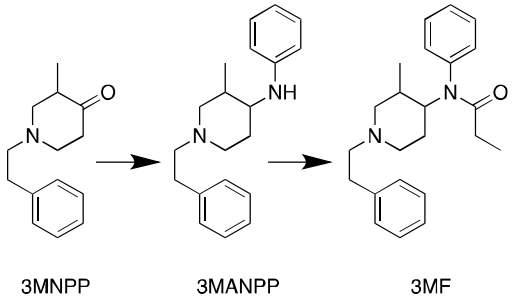
Introduction

Critical to law enforcement and related intelligence efforts to combat illicit drug abuse are analytical methods that can provide some level of attribution for drug source identification. This attribution typically relies on the chemical attribution signatures (CAS) present in a drug sample. CAS can be defined as synthesis precursors and byproducts, impurities, degradation products, and metabolites in various biological matrices. The most informative CAS are those that are persistent in the materials made via a specific route, those that remain unchanged after operational surfaces, and those that are present independent of synthesis chemist, laboratory, and facilities. Many reports have detailed the forensic application of a variety of analytical techniques for the characterization of these organic species in often complex matrices. Analytical tools generally include various forms of chromatographic separation combined with spectrometric detection schemes. Gas and high-pressure liquid chromatographies (GC and HPLC, respectively) combined with mass spectrometric (MS) detection have traditionally been the workhorses of such studies.[1] Though they each have specific merits and drawbacks, both are geared exclusively towards the identification of organic species and are generally used independently of one another. Our previous work[2a-d] has shown the power of integrating the signatures identified with these two organic detection methods into a single chemometric methodology, and explored the utility of including inorganic CAS detected by ICP-MS.[2e]

3-Methylfentanyl has a potency (often reported as an ED_{50}) reported to be roughly three orders of magnitude greater than that of morphine, making it 10-15 times stronger than its parent opioid, fentanyl, (*N*-(1-phenethylpiperidin-4-yl)-*N*-phenylpropionamide).[3]

This potency has restricted 3-methylfentanyl's usefulness as an analgesic; instead it is largely used recreationally, where mere grams of the opioid can be formulated into thousands of doses for subsequent sale. Considering the relatively low cost of production and high return on sale of the material, it is not surprising to see the marked uptick in general opioid use and number of overdose cases in both the United States and abroad.[4]

As with our previous investigation into the chemical attribution of fentanyl, the routes in the current work were selected to share various synthetic steps. The strategy was adopted to 1) reflect likely syntheses used in clandestine laboratories and 2) to test the ability of the model to discriminate between very similar reaction schemes. Also akin to our previous work, all syntheses rely on the same general pathway: the formation of an intermediate piperidinone followed by reductive amination and acylation.



Scheme 1. General synthetic strategy for the generation of 3-methylfentanyl (3MF) starting with 3-methyl-*N*-phenethylpiperidin-4-one (3MNPP) and using 3-methyl-1-phenethyl-*N*-phenylpiperidin-4-amine (3MANPP) as the intermediate.

In half of the synthetic routes studied, a commercially sourced 3-methylpiperidin-4-one was used. In the remaining three routes, however, methylation of piperidon-4-one was performed in one of two ways, namely using the reducing agents sodium hydride (NaH) or lithium diisopropylamide (LDA). In-house generation of this precursor seems

likely, as the commercial price for the methylated version is upwards of $100/g compared to $1–$10/g for 4-piperidone hydrochloride.

From a forensic analytical perspective, synthetic schemes that share a number of starting materials or synthetic steps can complicate chemical attribution due to a small number of unique signatures, particularly when present at trace levels. In our FY14 CFP work, we extracted comprehensive exact mass information from LC-QTOF data in addition to using electron impact (EI) and chemical ionization (CI) GC/MS data. Inorganic material present in crude samples were detected and semi-quantified using ICP-MS. In the current work on 3MF, several instrumental and sample preparation techniques were improved upon to enhance instrumental sensitivity and the amount and quality of data. To compliment the use of LC-QTOF, GC-QTOF was used to extract high-resolution mass spectrometric data from the more volatile organic species present in crude 3MF samples. This GC-QTOF was recently acquired in the Forensic Science Center with internal funding, and gives us additional GC-MS sensitivity as well as the ability to obtain exact mass information. In addition, the crude samples were completely microwave digested before subsequent analysis by ICP-MS using a microwave digester that was also recently acquired in the FSC using internal and other sponsor funding. The impact of these strategies is discussed.

The three analytical techniques used herein provided over 200 unique synthesis-related signatures and a full panel of elemental data was acquired by ICP-MS. The continued complexity of the resultant signature data, however, demanded the use of statistical techniques to extract relevant CAS. Therefore, multivariate statistical analyses were employed to extract the main sources of variance among the six synthetic routes.

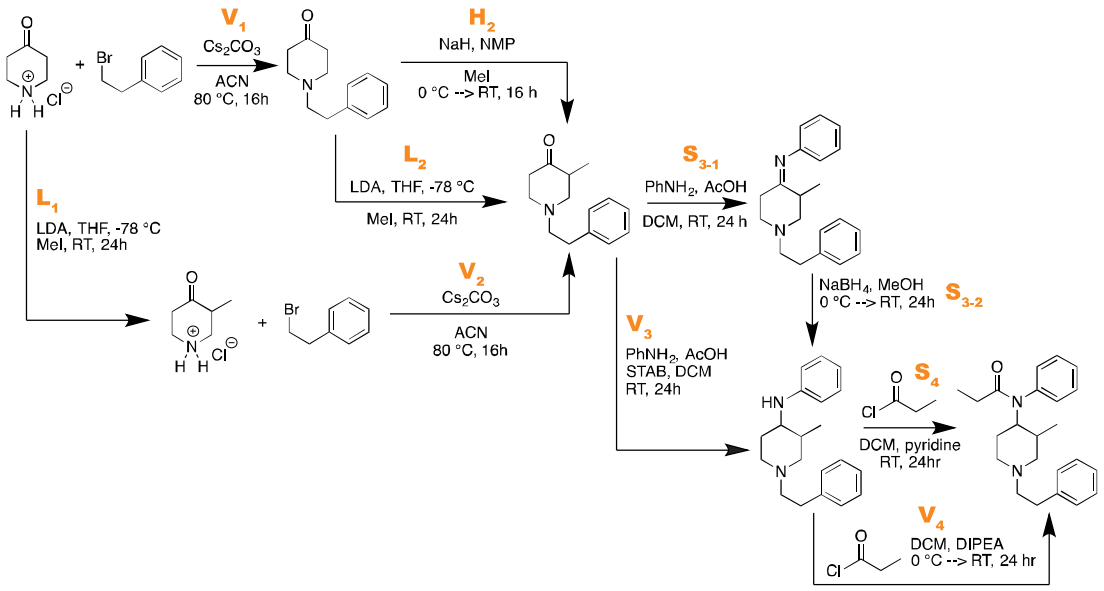
Resulting models were shown to be able to identify compounds of significance as chemical attribution signatures. Blind syntheses were then performed by two synthetic chemists not previously involved with the work. Choosing three routes known only to the two chemists, three crude samples of 3MF were generated and analyzed using the multivariate models, which classified the unknown sample data against the training data. Despite apparent dissimilarities between the training and unknown data sets, the models were able to confidently assign the correct route in most cases. For samples where classification was ambiguous, the models were able at the very least to eliminate a significant number of the routes studied.

Experimental

Synthetic Approach

The routes chosen for this study were selected to generate the 3MF in ways that closely mimic probable illicit manufacturing methods. These methods use 3MNPP as the synthetic foundation. While commercially available 3-methylpiperidin-4-one was used in three of the six routes studied, its cost pre unit weight is relatively high. In the remaining routes, then, in-house synthesis of 3-methylpiperidin-4-one was performed. Though the synthesis of fentanyl-like compounds dates back to the 1960s in work by Janssen,[5] it involves techniques believed to be too complicated or expensive for clandestine laboratories such as hydrogenations with precious metal catalysts. Therefore, the alternative routes chosen use common, easily accessible components and relatively mild reaction conditions – namely the Valdez and Siegfried methods (*vide infra*).

As with the FY14 work on fentanyl, the six chosen methods offer variations on the attachment of the different 3MF functional groups on the piperidine ring. 3-methyl-1-phenethyl-*N*-phenylpiperidine-4-amine (3MANPP), the direct precursor to 3MF, was generally formed from 3MNPP using the Valdez method (i.e. via reductive amination).[6] Only in one case ($VSS) was a two-step condensation-reduction method processed used involving sodium triacetoxyborohydride (STAB) (i.e. the Siegfried method[7]). Similarly, 3MF was generally formed using propionyl chloride and *N*,*N*-diisopropylethylamine (DIPEA, Valdez method), though in two cases pyridine used as the base (Siegfried route). To summarize: 1) when commercially available 3-methylpiperidon-4-one was used, variations using some combination of Valdez and Siegfried routes from 3MNPP to 3MF were used; and 2) when 3-methylpiperidon-4-one was synthesized in-house, only the Valdez methods for 3MF synthesis were used. In this way, signatures relating to hydride- or LDA-mediated methylation of 3-methylpiperidone could be isolated from signatures associated with 3MANPP and 3MF synthesis using the Siegfried route instead of the Valdez method. The interconnectedness of the routes is shown in Scheme 2, and a summary table of routes is given in Table 1. Samples were generated *in triplicate*, and given the six routes, a total of 18 test samples of crude 3MF product were available for analysis.



Scheme 2. Overall synthetic pathway to 3MF. Routes are coded via four letters, "XXXX." The orange subscripts refer to the position in the code. "L" = LDA-mediated methylation; "H" = sodium hydride-mediated methylation; "V" = Valdez method[6]; "S" = Siegfried method.[7] Use of a "$" in the first position indicates commercial 3-methylpiperidin-4-one was used. For example, LVVV describes a route that uses LDA to methylate piperidinone HCl, Cs_2CO_3 to for 3MNPP, STAB and AcOH for reductive amination, and alkylation in the presence of DIPEA.

Table 1. Breakdown of the six 3MF synthetic routes. For the top four routes, the methylation is completed as the first step either through LDA mediation or by using the commercial source. For the bottom two routes, NPP is made first with subsequent methylation with LDA or sodium hydride reagents. Commonalities among routes are highlighted through colored fields; unique steps are left white.

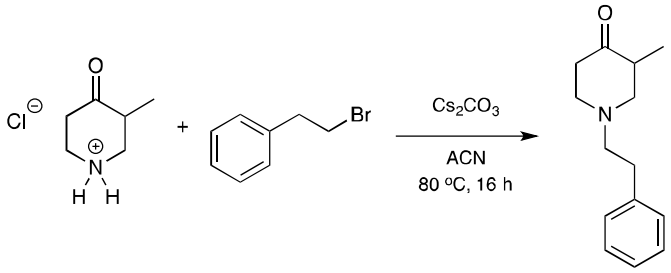
Route Code	3-methyl-piperidon-4-one	3MNPP	3MANPP	3MF
$VVV	Commercial source	Cs_2CO_3, ACN, 80°C, 16 h	AcOH, STAB, DCM, RT, 24 h	DCM, DIPEA, 0°C→RT, 24h
$VVS	Commercial source	Cs_2CO_3, ACN, 80°C, 16 h	AcOH, STAB, DCM, RT, 24 h	DCM, pyridine, RT, 24 h
$VSS	Commercial source	Cs_2CO_3, ACN, 80°C, 16 h	AcOH, DCM, RT, 24 hr	DCM, pyridine, RT, 24 h
LVVV	LDA, THF, -78°C, MeI, RT, 24h	Cs_2CO_3, ACN, 80°C, 16 h	AcOH, STAB, DCM, RT, 24 h	DCM, DIPEA, 0°C→RT, 24 h
Route Code	**NPP**	**3MNPP**	**3MANPP**	**3MF**
VHVV	Cs_2CO_3, ACN, 80°C, 16 h	NaH, NMP, MeI, 0°C→RT, 16 h	AcOH, STAB, DCM, RT, 24 h	DCM, DIPEA, 0°C→RT, 24 h
VLVV	Cs_2CO_3, ACN, 80°C, 16 h	LDA, THF, -78°C, MeI, RT, 24h	AcOH, STAB, DCM, RT, 24 h	DCM, DIPEA, 0°C→RT, 24 h

Materials

Unless otherwise stated, all reagents and solvents were obtained from commercial suppliers (Sigma-Aldrich (St. Louis, MO), Alfa Aesar (Ward Hill, MA), J.T. Baker (Avantor Performance Materials, Center Valley, PA), Fisher Chemical (Fairlawn, NJ) and used as received.

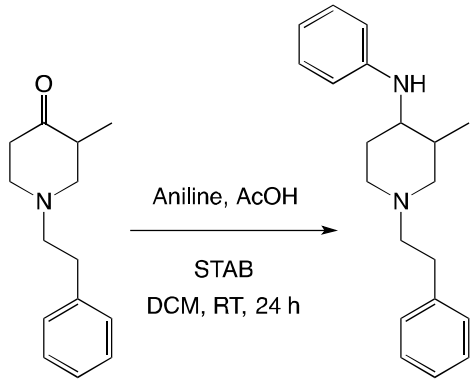
Synthesis

$VVV

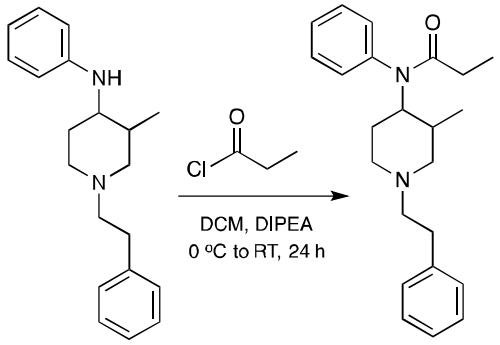


3-Methyl-4-piperidinone HCl (200 mg, 1.34 mmol) was taken up in acetonitrile (10 mL) in a 20 mL scintillation vial equipped with a stir bar. To the solution, cesium carbonate (1.1 g, 3.34 mmol, 2.5 equiv.) was added followed by the dropwise addition of phenylethyl bromide (618 mg, 3.34 mmol). The mixture was heated to 80 °C overnight. The following day, the mixture was partitioned (DCM[a]//H_2O) and the organic phase washed with brine (1 x 10 mL), dried over anhydrous sodium sulfate and evaporated *in vacuo* at 50 °C to yield a light brown oil.

[a] DCM = dichloromethane



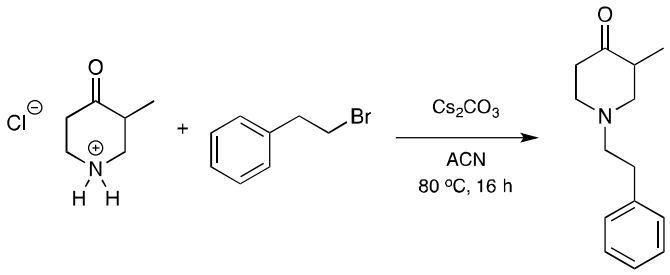
The oil residue from the previous step (131 mg, ~ 0.60 mmol) was taken up in DCM (5 mL) in a 20 mL scintillation vial equipped with a stir bar. This mixture was added to a mixture of aniline (57 µL) and acetic acid (80% in water, 45 µL) in DCM (10 mL) in another 20 mL scintillation vial. The mixture was then treated with sodium triacetoxyborohydride (STAB, 127 mg) and stirred overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.



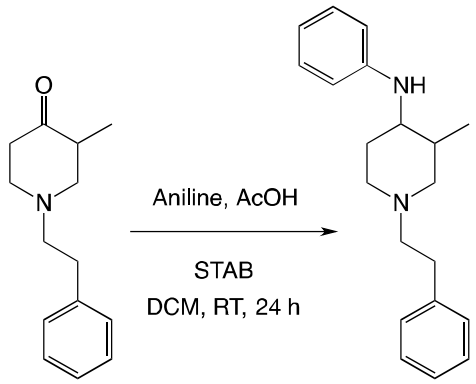
The above mixture (88 mg, ~ 0.30 mmol) was taken up in DCM (3 mL) in a 20 mL scintillation vial equipped with a stir bar. The mixture was cooled to 0 °C and treated sequentially with diisopropylethylamine (105 µL, 0.60 mmol) and propionyl chloride (53 µL, 0.60 mmol). The mixture was allowed to reach RT, and stirred overnight at this temperature. The next day, the mixture was partitioned (DCM//H_2O) and the organic

layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.

$VVS

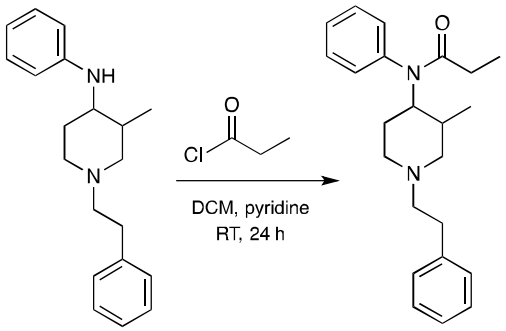


3-Methyl-4-piperidinone HCl (200 mg, 1.34 mmol) was taken up in acetonitrile (10 mL) in a 20 mL scintillation vial equipped with a stir bar. To the solution, cesium carbonate (1.1 g, 3.34 mmol, 2.5 equiv.) was added followed by the dropwise addition of phenylethyl bromide (618 mg, 3.34 mmol). The mixture was heated to 80 °C overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic phase washed with brine (1 x 10 mL), dried over anhydrous sodium sulfate and evaporated in vacuo at 50 °C to yield a light brown oil.



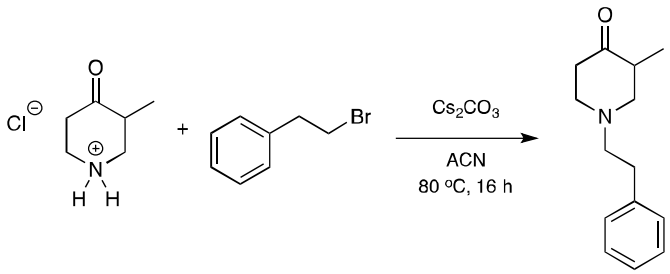
The oil residue from the previous step (131 mg, ~ 0.60 mmol) was taken up in DCM (5 mL) in a 20 mL scintillation vial equipped with a stir bar. This mixture was added to a mixture of aniline (57 µL) and acetic acid (80% in water, 45 µL) in DCM (10

mL) in another 20 mL scintillation vial. The mixture was then treated with sodium triacetoxyborohydride (127 mg) and stirred overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.



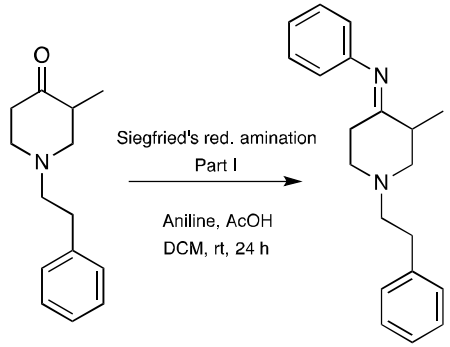
The above residue (283 mg, ~ 0.96 mmol) was taken up in DCM (8 mL) in a 20 mL scintillation vial equipped with a stir bar. The yellow solution was treated sequentially with pyridine (139 □L, 1.73 mmol) and propionyl chloride (151 □L). The resulting mixture was vigorously stirred at RT overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated in the rotavap to yield a light brown oil.

$VSS

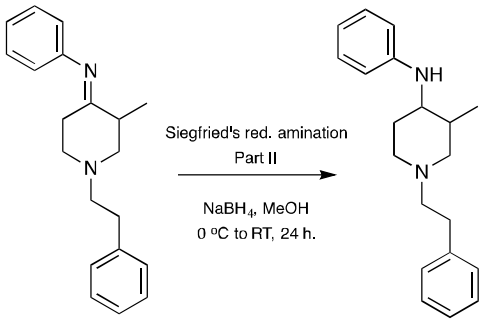


3-Methyl-4-piperidinone HCl (200 mg, 1.34 mmol) was taken up in acetonitrile (10 mL) in a 20 mL scintillation vial equipped with a stir bar. To the solution, cesium

carbonate (1.1 g, 3.34 mmol, 2.5 equiv.) was added followed by the dropwise addition of phenylethyl bromide (618 mg, 3.34 mmol). The mixture was heated to 80 °C overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic phase washed with brine (1 x 10 mL), dried over anhydrous sodium sulfate and evaporated in vacuo at 50 °C to yield a light brown oil.

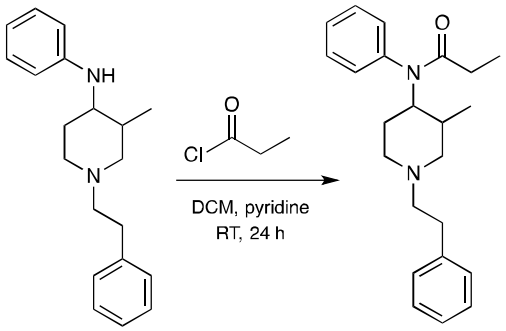


The reaction mixture obtained above (258 mgs of oil) was taken up in DCM (5 mL) and added to a stirred solution of aniline (111 µL) and acetic acid (80% in water, 90 µL) in DCM (10 mL) in a 20 mL scintillation vial equipped with a stir bar. The mixture was vigorously stirred overnight at RT. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over sodium sulfate and evaporated to dryness to furnish a yellow oil (net weight = 461 mg).



The imine mixture obtained above (461 mg, ~ 1.58 mmol) was dissolved in MeOH (3 mL) in a 20 mL scintillation vial equipped with a stir bar and cooled to 0 °C

with an ice bath. To the mixture, sodium borohydride (175 mg, 4.74 mmol, 3.0 equiv.) was added in small portions and the resulting suspension (yellow) was stirred at RT overnight. The next day, the vial was diluted with water (6 mL) and partitioned (DCM//H_2O). The organic phase was dried over anhydrous sodium sulfate and evaporated *in vacuo* to yield a yellow oily residue.

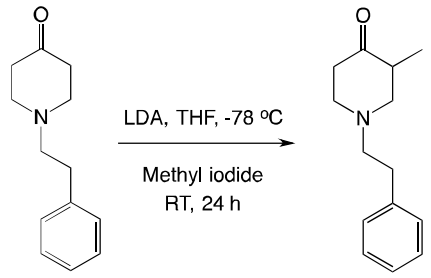


The above residue (283 mg, ~ 0.96 mmol) was taken up in DCM (8 mL) in a 20 mL scintillation vial equipped with a stir bar. The yellow solution was treated sequentially with pyridine (139 μL, 1.73 mmol) and propionyl chloride (151 μL). The resulting mixture was vigorously stirred at RT overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated in the rotavap to yield a light brown.

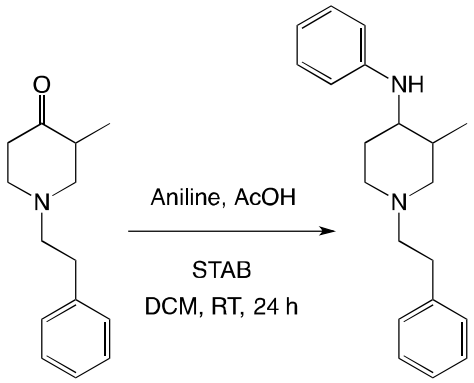
VLVV

2.9 g (19 mmol) of pulverized piperidinone HCl was loaded into a round flask with a stir bar. 60 mL of acetonitrile was added and the slurry set to stir. 13.5 g (41 mmol) pulverized cesium carbonate was added in portions. 3.1 g (17 mmol) 2-bromoethylbenzene was added and a reflux condenser was attached. The mixture was refluxed at 60 °C for 5 hours and cooled to RT. The mixture was filtered into a separatory funnel and partitioned (DCM//H_2O). The organic phase was isolated and washed with

brine (3 x 20 mL) and saturated sodium bicarbonate (3 x 20 mL). The solution was dried with sodium sulfate, filtered, and evaporated yielding 2.75 g of NPP as a yellow oil.

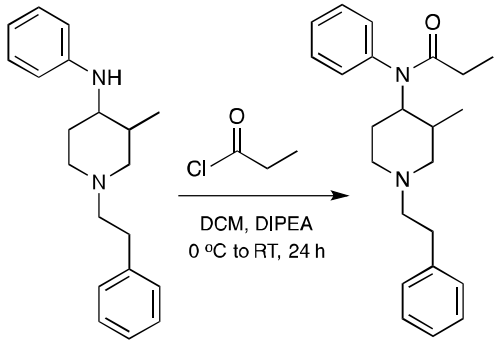


The resultant material (150 mg, < 0.74 mmol) was dissolved in anhydrous THF (5 mL) and the orange suspension was cooled to -78 °C. LDA (95 mg, 0.88 mmol) was added in small portions to the mixture and this was stirred for 10 minutes at -78 °C. Methyl iodide (55 µL, 0.88 mmol) was added using a pipette and the resulting mixture was allowed to slowly warm up to RT and stirring was continued overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish an amber-yellow oil.



The oil residue from the previous step (131 mg, ~ 0.60 mmol) was taken up in DCM (5 mL) in a 20 mL scintillation vial equipped with a stir bar. This mixture was added to a mixture of aniline (57 µL) and acetic acid (80% in water, 45 µL) in DCM (10 mL) in another 20 mL scintillation vial. The mixture was then treated with sodium

triacetoxyborohydride (127 mg) and stirred overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.

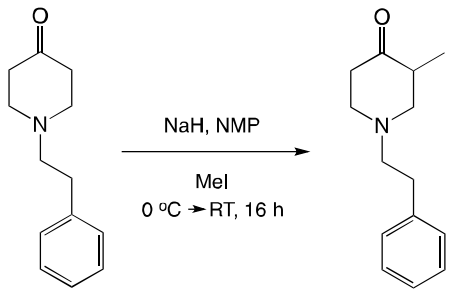


The above mixture (88 mg, ~ 0.30 mmol) was taken up in DCM (3 mL) in a 20 mL scintillation vial equipped with a stir bar. The mixture was cooled to 0 °C and treated sequentially with diisopropylethylamine (105 µL, 0.60 mmol) and propionyl chloride (53 µL, 0.60 mmol). The mixture was allowed to reach RT, and stirred overnight at this temperature. The next day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.

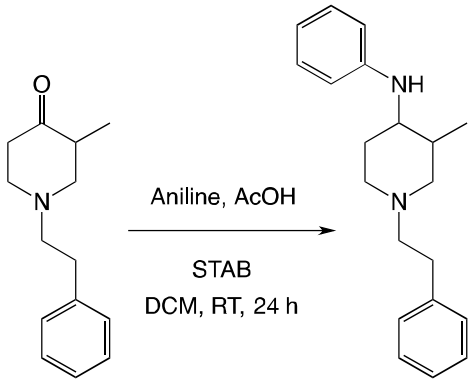
VHVV

2.9 g (19 mmol) of pulverized piperidinone HCl was loaded into a round flask with a stir bar. 60 mL of acetonitrile was added and the slurry set to stir. 13.5 g (41 mmol) pulverized cesium carbonate was added in portions. 3.1 g (17 mmol) 2-bromoethylbenzene was added and a reflux condenser was attached. The mixture was refluxed at 60°C for 5 hours and cooled to room temperature. The mixture was filtered into a separatory funnel and partitioned (DCM//H_2O). The organic phase was isolated and

washed with brine (3 x 20 mL) and saturated sodium bicarbonate (3 x 20 mL). The solution was dried with sodium sulfate, filtered, and evaporated yielding 2.75 g of NPP as a yellow oil

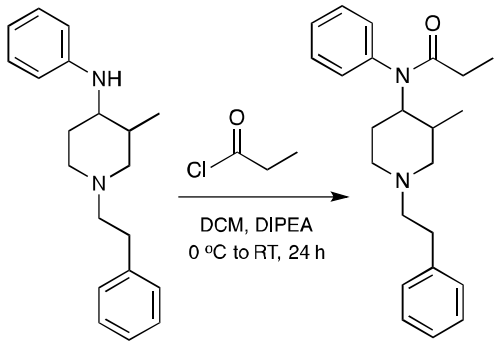


The above yellow oil (160 mg, ~ 0.79 mmol) was dissolved in NMP (3 mL) in a 20 mL scintillation vial equipped with a stir bar and cooled to ~ 0 °C using an ice bath. Sodium hydride (60% dispersion in oil, 34 mg, 1.1 equiv.) was added in one portion and evolution of a gas (H_2) was observed. The resulting orange suspension was treated with methyl iodide (54 µL, 1.1 equiv. to FSC2-33) and the mixture stirred at RT overnight. The following day, the dark orange mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish an amber-yellow oil.



The oil residue from above (100 mg, ~ < 0.46 mmol) was taken up in DCM (5 mL) in a 20 mL scintillation vial equipped with a stir bar. This mixture was added to a

mixture of aniline (46 µL, 2.5 equiv.) and acetic acid (80% in water, 115 µL) in DCM (10 mL) in another 20 mL scintillation vial. The mixture was then treated with sodium triacetoxyborohydride (58 mg) and stirred overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated *in vacuo* to dryness to furnish a yellow oil.



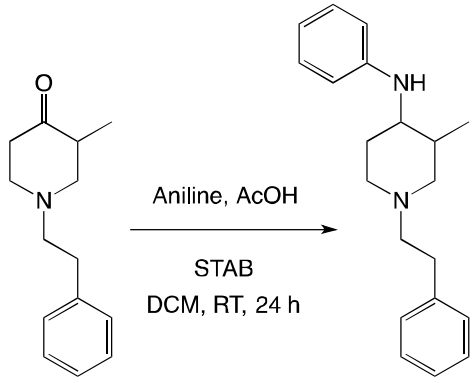
The above oily residue (78 mg, ~ < 0.27 mmol) was taken up in DCM (3 mL) in a 20 mL scintillation vial equipped with a stir bar. The mixture was cooled to 0 °C and treated sequentially with diisopropylethylamine (93 µL, 0.53 mmol) and propionyl chloride (47 µL, 0.53 mmol). The mixture was allowed to reach RT, and stirred overnight at this temperature. The next day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated *in vacuo* to dryness to furnish a dark yellow oil.

LVVV

4.75g of 1-Boc-4-piperidone was dissolved in 100 mL of dry THF and cooled with a -78 °C dry ice/acetone bath under a flow of argon. 25 mL of 1.0 M LDA in THF was added dropwise and stirred for 1 hour at low temperature. 1.8 mL of methyl iodide was added and the mixture was brought to RT and stirred overnight. Aqueous workup

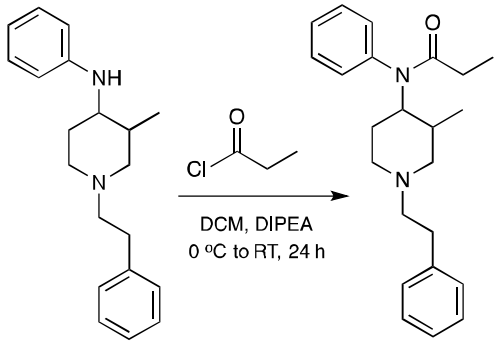
afforded 3.7 g of a yellow oil which was dissolved in 25mL of dry dichloromethane, cooled to 5 °C, and treated with 20 mL of 2.0 M HCl in diethyl ether, dropwise. Precipitation begins shortly after the addition of the acid. The mixture was left to stir overnight. The solid was isolated by decantation, washed with dichloromethane, and dried yielding roughly 2 g of 3-methylpiperidone HCl salt.

3-Methyl-4-piperidinone HCl (200 mg, 1.34 mmol) was taken up in acetonitrile (10 mL) in a 20 mL scintillation vial equipped with a stir bar. To the solution, cesium carbonate (1.1 g, 3.34 mmol, 2.5 equiv.) was added followed by the dropwise addition of phenylethyl bromide (618 mg, 3.34 mmol). The mixture was heated to 80 °C overnight. The following day, the mixture was partitioned (DCM//H_2O) and the organic phase washed with brine (1 x 10 mL), dried over anhydrous sodium sulfate and evaporated *in vacuo* at 50 °C to yield a light brown oil.



The oil residue from the previous step (131 mg, ~ < 0.60 mmol) was taken up in DCM (5 mL) in a 20 mL scintillation vial equipped with a stir bar. This mixture was added to a mixture of aniline (57 μL) and acetic acid (80% in water, 45 μL) in DCM (10 mL) in another 20 mL scintillation vial. The mixture was then treated with sodium triacetoxyborohydride (127 mg) and stirred overnight. The following day, the mixture

was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.



The above mixture (88 mg, ~ < 0.30 mmol) was taken up in DCM (3 mL) in a 20 mL scintillation vial equipped with a stir bar. The mixture was cooled to 0 °C and treated sequentially with diisopropylethylamine (105 µL, 0.60 mmol) and propionyl chloride (53 µL, 0.60 mmol). The mixture was allowed to reach RT, and stirred overnight at this temperature. The next day, the mixture was partitioned (DCM//H_2O) and the organic layer was dried over anhydrous sodium sulfate and evaporated to dryness to furnish a yellow oil.

Instrumental Analysis

LC-QTOF

A quantitative weight of each crude fentanyl product was transferred to a 4 mL glass vial and diluted to 10 mg/mL in acetonitrile. From each solution, dilutions were performed to yield a final series of 20 µg/mL solutions in 75/25 water/acetonitrile. An Agilent 1260 LC equipped with an Atlantis T3 reverse phase column (C18, 150 mm x 2.1 mm, 3 µm particle size, Waters, Milford, MA) was used for the separation of all compounds. Time-of-flight mass spectrometric detection was performed in positive ion

mode with a Bruker micrOTOF-Q III (Bruker Daltonics, Billerica, MA) equipped with an electrospray ionization (ESI) source and operated in Auto MS/MS mode (m/z 50-1000). Three precursor ions were monitored at a given time (m/z 50-450) with active exclusion after three spectra. MS was performed with a capillary voltage of 4500 V, a dry gas flow rate of 4 L/min at 180 °C, quadrupolar ion and collision energies of 4.0 eV and 8.0 eV, respectively, and a spectral acquisition rate of 2 Hz.

The mobile phase consisted of water with 0.1% formic acid (A) and acetonitrile with 0.1% formic acid (B); flowrate of 0.25 mL/min. The gradient profile started with 95% A for 2 min, ramped to 20% A over 7 minutes, held for 14.75 minutes, ramped quickly back to 95% over 0.25 min, and held for 8 min for column regeneration. Five microliters of the liquid sample were introduced via an autosampler (Agilent B1329B) to the injection port. The detector was tuned and calibrated using the 20 μL injection loop of a 6-port valve on the MS using Agilent's ESI-L Low Concentration Tuning Mix (G1969-85000). Compounds relevant to each synthetic route were identified based on computer-aided identification of MS/MS peaks using Bruker's Compass for otofSeries 1.5 software. Detailed analysis of each sample was done manually with peak areas calculated by manually integrating the extracted ion chromatogram (EIC) of the base peak. After route-specific compounds were identified for each route, a target table was created and searched against all 18 samples.

GC-QTOF

A quantitative weight of each crude fentanyl product was transferred to a 2 mL glass vial and diluted to 10 mg/mL using dichloromethane. Dilutions were then

performed in dichloromethane to yield a series of 20 μg/mL solutions. An Agilent Technologies (Santa Clara, CA) 7890B GC equipped with an Agilent HP-5MS column (5%-Phenyl-methylpolysiloxane, 30 m x 0.25 mm x 0.25 μm) was used for the chromatographic separation. A carrier gas of helium (99.999%, Praxair, Inc., Danbury CT) was used, and the GC was operated in constant flow mode (1.0 mL/min). One microliter of the liquid sample was introduced via an autosampler (Agilent 7890 series) to the injection port held at 250°C, splitless injection. The oven temperature was held at 40°C for 3 minutes, then ramped at 8°C/min to 300°C and held for 3 minutes. Detection was performed with an Agilent 7200 Accurate-Mass Q-TOF MS detector and operated using EI (70 eV) and an emission current at 10 μA. The system was operated in scan mode (m/z 29-600, 8 spectra/s) with the source and quadrupole mass analyzer held at 230°C and 150°C, respectively. A solvent delay of 3 minutes was used. The detector was auto-tuned and was mass calibrated using a perfluorotributylamine (PFTBA) solution before running each sample. Compounds were identified based on spectral comparison to the NIST Mass Spectral Library (v 2014) as well as manual comparison to mass spectra published in the literature. Peak areas were calculated from extracted ion chromatograms of the base peak identified for each compound using manual integration in the MassHunter software (v B.07.00).

ICP-MS

Sample preparation prior to ICP-MS analysis was performed via microwave digestion of the crude samples in acidic solution. 10 mL of 3 M nitric acid were added 100 mg of crude material \which was then sonicated to homogenize the solution as much

as possible and quantitatively transferred into microwave digestion vessels with an additional 2 mL of 3 M nitric acid. The vessels were installed on the rotary stage and inserted into a MARS 6 microwave digester (CEM Corp., Matthews, NC). Samples were then processed using a standard pharmaceutical digestion protocol supplied by the instrument's manufacturer.

Elemental analysis was performed using an Agilent Technologies (Santa Clara, CA) 8800 Triple Quadrupolar ICP-MS (ICP-QQQ). Initially, a semi-quantitative scan was performed to determine elements of interest within the sample sets. Down-selected analytes were then measured quantitatively with the following parameters: carrier gas (0.65 L/min), nebulizer pump (0.50 rps), spray chamber temperature (15 °C), and dilution gas (0.40 L/min). Argon was used as plasma, carrier, and dilution gas. In the collision cell, a helium flow of 4.0 mL/min was used. The measurements were performed as three replicates, with 50 sweeps per replicate. Integration time per mass was held at 0.10 sec. The rinse time was set to 30 sec at 0.3 rps of the nebulizer pump, followed by 10 sec at 0.3 rps. Sample introduction was performed using an ASX-500 autosampler (Cetac, Omaha, NE). Peak area determination and quantitation were performed using MassHunter software (4.1 C.01.01).

Data Analysis

Partial Least Squares Discriminant Analysis (PLS-DA), a supervised technique that facilitates classification, was performed using Solo (V8.0, Eigenvector Research Inc., Wenatchee, WA). This software was also used to perform the Support Vector Machines analyses described in detail below. Mathematica 9.0 (Wolfram, Champaign, IL, USA)

was used for two other machine learning techniques, namely logisitic regression and neural network analyses. The data sets fed into all algorithms were identical regardless of software package and included peak areas (LC-QTOF and GC-QTOF) and accurate concentrations (ICP-MS).

Results and Discussion

Signature Identification

From the exhaustive list of signatures revealed by LC (given in Appendix A), we investigated the extent to which MS/MS spectra could be used to posit structures associated with each peak's exact mass. From MS/MS fragment exact masses, substructural units of the parent molecule can be used to reconstruct its original identity. Table 2 shows the results of proposed structures using this methodology. Note that it was outside the scope of work to confirm these assignments with commercial standards or in-house synthesis, but these proposed structures are the most consistent with available data and most likely reaction chemistries.

Early eluting compounds largely correspond to the bases (pyridine or DIPEA) used for the final acylation step, which converts 3MANPP to 3MF. Note that signature 3 is most likely a synthesis impurity in DIPEA. Compounds 11-43 in Table 2 are intermediates associated with incomplete reductive amination of 3MNPP. For example, the ketone group of unreacted NPP or 3MNPP is reduced to a hydroxyl group (compounds 11, 15, 16). LDA- or NaH-mediated methylation creates the intended 3MNPP (compound 13) but also creates compounds with higher degrees of methylation

(compounds 25, 29).[b] The reduced NPP and 3MNPP can react in the final step with propionyl chloride to create the ester moeties of compounds 40 and 43. Because the conversion of 3MNPP to 3MANPP was apparently the lowest yielding step (see below), one would expect to see a significant amount of fentanyl. Indeed, this expectation was observed (compound 49). Direct reaction between excess aniline and propionyl chloride and to a lesser extent AcOH is responsible for compounds 65 and 66, respectively.

The majority of early eluting species from GC data correspond to highly volatile reagents and their impurities such as pyridine, aniline, and DIPEA. Also observed were unreacted 2-bromoethyl benzene and its likely impurities, 2-chloroethyl benzene and benzeneacetaldehyde. Compounds of moderate volatility correspond to many of those detected by LC-QTOF. For example, *N*-phenyl propanamide (LC signature 66), 3-methyl-1-phenethylpiperidin-4-one (LC signature 13), and dimethyl-1-phenethylpiperidin-4-one[c] (LC signatures 25, 29) were detected by both techniques. Despite their relatively low volatility and their large retention times (32-33 min), fentanyl and two 3-methylfentanyl peaks (corresponding to the 3*R* (*cis*) and 3*S* (*trans*) enantiomers) were also detected by GC-QTOF.

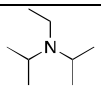
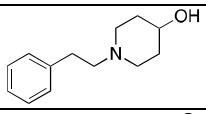
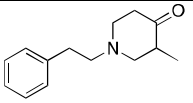
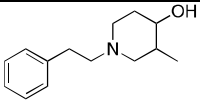
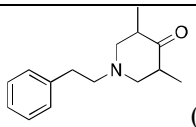
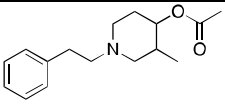
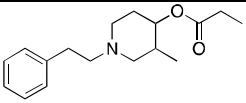
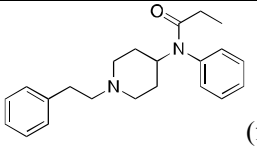
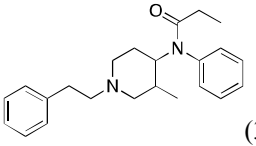
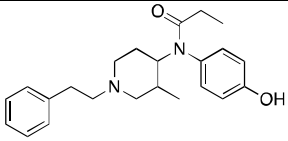
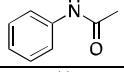
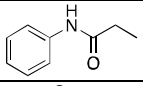
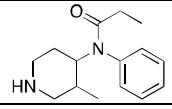
It is important to mention that while we are able to assign identities/structures to the above compounds, the vast majority of those discussed above are *not* ultimately considered CAS by statistical modeling. Most of those compounds (as determined via LC) can only be represented as formulae or exact masses.

---

[b] Negligible amounts of compounds with higher degrees of methylation were detected in crude samples that started with commercially available product. Detection of these compounds strongly suggests in-house synthesis of the starting material, 3-methylpiperidin-4-one, for the routes chosen for the present study.

[c] Indeterminate methyl group substitution pattern without standard reference materials.

Table 2. LC signatures for which proposed structures were derived.

Signature ID	Proposed Chemical Formula	Proposed Structure from MS/MS data
1	C_5H_5N	
3	$C_7H_{17}N$	
5	$C_8H_{19}N$	
11	$C_{13}H_{19}NO$	
13	$C_{14}H_{19}NO$	
15, 16	$C_{14}H_{21}NO$	
25, 29	$C_{15}H_{21}NO$	 (or analog)
40	$C_{16}H_{23}NO_2$	
43	$C_{17}H_{25}NO_2$	
49	$C_{22}N_{28}N_2O$	 (fentanyl)
51	$C_{23}H_{30}N_2O$	 (3-methylfentanyl)
57	$C_{23}H_{30}N_2O_2$	 (Signature Q from Signature Science Ref[8])
65	C_8H_9NO	
66	$C_9H_{11}NO$	
82	$C_{15}H_{22}N_2O$	

Enantiomeric Profiling of 3-Methylfentanyl

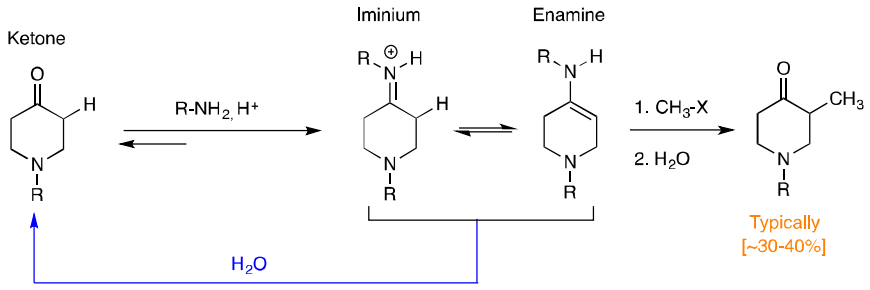
Also investigated was the degree to which the various synthetic routes produced the *R* vs. *S* enantiomers of 3-methylfentanyl, as this number might in principle be an additional "signature" by which the routes can be distinguished. Several attempts at using LC to separate the two enantiomers were unsuccessful, whether using a typical reverse phase column or a chiral column recommended by Phenomenex. GC, however, was able to cleanly separate the two species (with retention times of 33.28 and 33.65 min; previous work suggests that the early eluter corresponds to the more stable *trans*/3*S* configuration[d]). Three of the synthetic routes produced enough 3MF to yield peaks with sufficient signal-to-noise. The average ratios of the *cis* to *trans* 3MF were 3.88 ± 0.23 for $VVV, 3.99 ± 0.25 for $VVS and 4.26 ± .85 for LVVV. Statistical comparisons of these averages suggests there is no significant different in 3MF enantiomeric composition for these three routes, one of which generated the starting material using LDA. Additionally, these values are somewhat higher than those from the Espoto/Winek study (1.73 – 2.96 *cis*:*trans*) but are consistent is the observation that the *cis* 3MF is preferentially synthesized. As the remaining three routes studied presently did not produce sufficient 3MF, however, we cannot draw more insightful conclusions on the potential utility of this measure as an additional signature.

[d] The work of Esposito and Winek posits that the compound associated with the lower abundance ratio of both *m*/*z* = 160 and 203 to the base peak *m*/*z* = 259 corresponds to the more stable *trans* configuration of 3MF where the methyl group and amide nitrogen are in equatorial and axial positions, respectively. Esposto, F.; Winek, C. *J. Forensic Sci.* **1991**, *36*, 86-92.

Observations on Low Yields of 3-Methylfentanyl

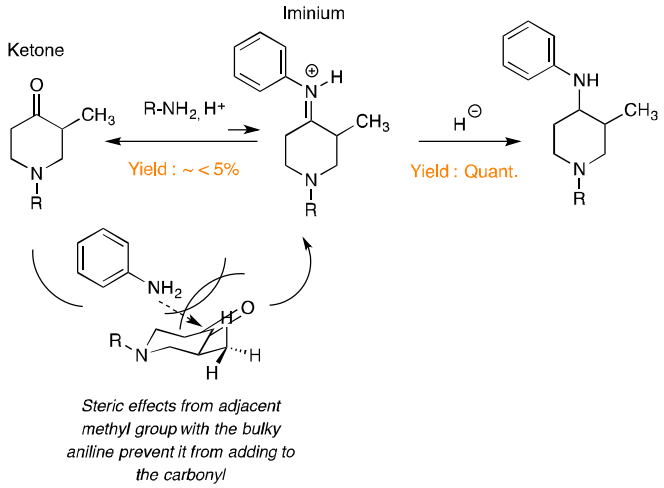
During the synthesis of 3-methylfentanyl, it was noted that the overall yield of the target compound via some routes was extremely low (<4-5%) by GC-MS analysis, particularly for routes $VSS, VHVV, and VLVV. As it is an estimate based on GC-MS analysis of the crude mixtures along the steps comprising a given route, the actual final yield of the product is not known but can be placed at <10% with great certainty.

Two factors observed during the synthesis can be highlighted to help explain these low yields. The first one has to do with the routes where the α-methylation needed to be performed. The methylation step is actually a two-step procedure that involves the formation of an enolate ion from the ketone or imine species generated from the starting material and subsequent alkylation to yield the α-methylated product (Scheme 3). The enolate (or enamine arising from the imine) is prone to hydrolysis reverting back to the starting ketone. Furthermore, it is water sensitive, which would also convert it back to the starting material. Another added issue with this initial manipulation is the use of a hindered, amine base to form the enolate (or enamine) that unfortunately is also water sensitive even when carefully controlling the dryness of the organic solvent (THF for example) with molecular sieves or other dehydrating agents. Yields for this transformation are usually low, ranging between 30-40% when not conducted under stringently dry conditions (i.e. argon atmosphere, flame-dried glassware, heat-activated molecular sieves, etc.)



Scheme 3. Methylation of starting ketone and competing pathways for the reaction.

The second factor, which we believe is more deleterious to overall yield, is the low-yielding reductive amination step. When carrying out the reductive amination on the generated 3-methylpiperidinones, it was observed by GC-MS that the conversion of the starting 3-methylketone to the final amine product was amazingly low (<5% conversion noted by GC-MS). The reductive amination is a two-step process – imine formation followed by hydride reduction. GC-MS studies revealed that the initial formation of the iminium intermediate was the low yielding step thus affecting the downstream reduction to the amine (Scheme 4). The reaction was studied to find a way to increase the overall yield of the amine, but even when using large excess of the amine to the ketone and performing the reduction at elevated temperature (80 °C), a larger yield was never accomplished. Although reductive aminations of 3-methylcyclohexanones have been reported in good yields (>80%), these involve the use of smaller, inherently more nucleophilic amines to form the iminium species.



Scheme 4. Steric factor between aniline and the methyl group governing the first step of the reductive amination step resulting in an overall low yielding transformation.

Data Transformation

In statistical analyses such as those described below, each chemical signature (i.e. the peak area of each chromatographic peak fed into the software) is considered to be a "predictor," and it is important to understand the distribution characteristics of the predictor values before any such analysis is performed. Histograms and boxplots can yield decent qualitative information as to the distribution of predictor values (i.e. peak areas) in a given sample. However, Q-Q plots are more powerful in quickly assessing the normality of a data set (a fundamental assumption of many statistical regression and classification algorithms). The left panel of Figure 1 gives the Q-Q plot for the raw peak areas for one of the LVVV data sets; clearly these data do not follow a normal distribution. Most mass spectral data, particularly those from the metabolomics community, rely on log-normalizing the data prior to other transformations. This treatment was given to the current data sets, the results of which are given in the right-

hand panel of Figure 1 for the representative dataset. The data conform to the reference line, indicating the base-ten logarithm yields reasonably normally distributed data.

Critical to both the stability of statistical calculations and the robustness of the resultant classification models are appropriate data preprocessing and transformation of these predictor values. For example, the most common and straightforward treatment involves centering and scaling the data, or subtracting the mean predictor value and standard deviation of predictor values within a given dataset, respectively. This ensures that each predictor has a zero mean and a common standard deviation of unity, so all predictors (signatures) have equal contributions to the model, regardless of absolute abundance.

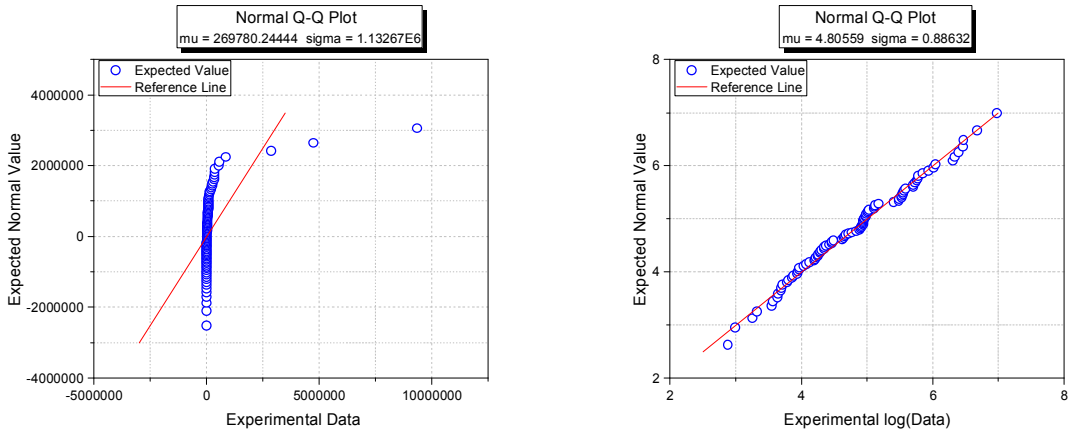


Figure 1. Normal Q-Q plots for a representative LVVV sample. Left panel: Raw LC+GC data showing non-normally distributed data. Right panel: Log_{10}-transformed raw LC+GC data showing log-normal behavior of the analytical data.

Accounting for zeros (i.e. "missing values") is another problematic issue that must be addressed. When a peak is recorded as "0" in the data set, it does not strictly mean that that signature is absent. Rather, it simply means that it is under the detection

limit of the instrument. Often, the detection limit itself is used in place of a zero value. A random number between zero and this limit is also possible. We chose a value of 0.0001 to replace all zeroes, which was subsequently converted to negative four in log-space. This practice is often seen in various "omics" communities. Centering and scaling of the $\log_{10}$-transformed data was then done for each crude sample using means and standard deviations of the corresponding data. Note that as with our previous work on fentanyl, it is important that LC data, GC data, and ICP-MS be treated separately, as they are acquired on different instruments with different responses and dynamic ranges.

For the three *unknown* data sets, the same statistics for mean centering and scaling were used. Specifically, the average mean and standard deviation of the training sets were used to transform test data, since it is assumed that the test data are fully blind (i.e. their underlying statistics are unknown).

Dimensionality and Feature Selection

A major issue with statistical analysis of chemical data is that while a very large amount of variables can be measured (i.e. peak areas) it is generally difficult to provide a similarly large number of sample replicates. This results in what is often referred to as the "curse of dimensionality." Lacking the sample size to support the results derived from the high-dimensional data set degrades the predictive power of classification and machine learning algorithm.[9] Thus, the issue of "over-fitting" the data is problematic, one that overemphasizes patterns that are not reproducible (i.e., outliers, noise, etc.) and makes it difficult to generalize to new samples. While we will work with the full datasets in the

current work (with acknowledging the negative impact of over-fitting) we also refer to our previous work to reduce the dimensionality of the dataset through statistical means.

Partial least squares (PLS) was used to assist in reducing the dimensionality of the predictor (i.e. variable) space. It has been well established that there exists a clear benefit in choosing PLS over other conventional methods (e.g. principal components analysis, PCA) for this reduction when followed by attempted classification. This fact results from PLS' seeking to optimize group separation guided by information between groups, rather than PCA's optimizing on overall variance. Either way, efforts to eliminate predictors (referred to as "feature selection") that contain little to no information about the underlying variance should improve the performance of any classification model.

While a handful of methods exist for feature selection, we chose to focus on *selectivity ratios* (SR). We found previously that this metric was useful in the objective identification of important CAS. A high SR value indicates the spectral variable contributes much toward discrimination amongst samples (i.e. explains a significant amount of total variance). A lower threshold SR value of 1.5 was chosen to select features that would be used to generate a reduced predictor set for classification purposes. This value also allowed for each route to have at least one LC signature identified as an important feature.

Samples and Predictors by PLS

Though PLS was employed primarily for feature selection, its results can also be used to assess the in- and between-class separation through scores values of the latent variables. Figure 2 shows a two-dimensional scores plot on LVs 1 and 2 for the full

LC+GC data set (141 LC inputs and 26 GC inputs). Note that in-class variability is low, since scores of the various routes cluster tightly (i.e. the three synthetic replicates of each route are similar to each other). Between-class separation, however, is observed to be quite high (i.e. the different synthetic routes are quite different from each other). Also, similar routes lie along various lines within the space. That is, scores from samples using the commercial starting material all lie along one line, while routes than use LDA-mediated methylation lie along another. The hydride-mediated route VHVV appears separate from the others though is closest to its cousin, VLVV. In summary, samples with similar chemistries can be visualized easily through the reduced dimensionality of PLS scores plots.

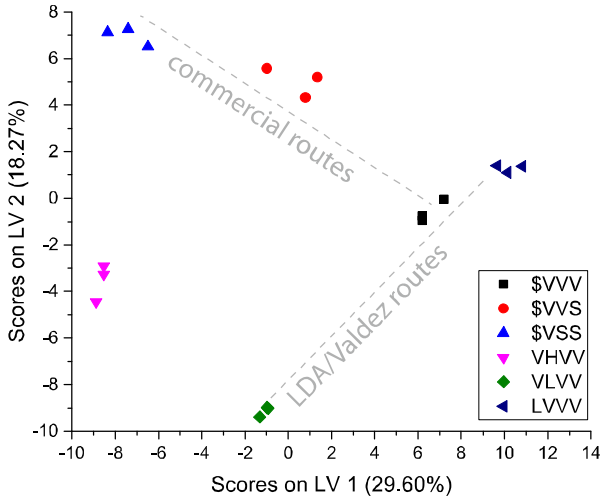
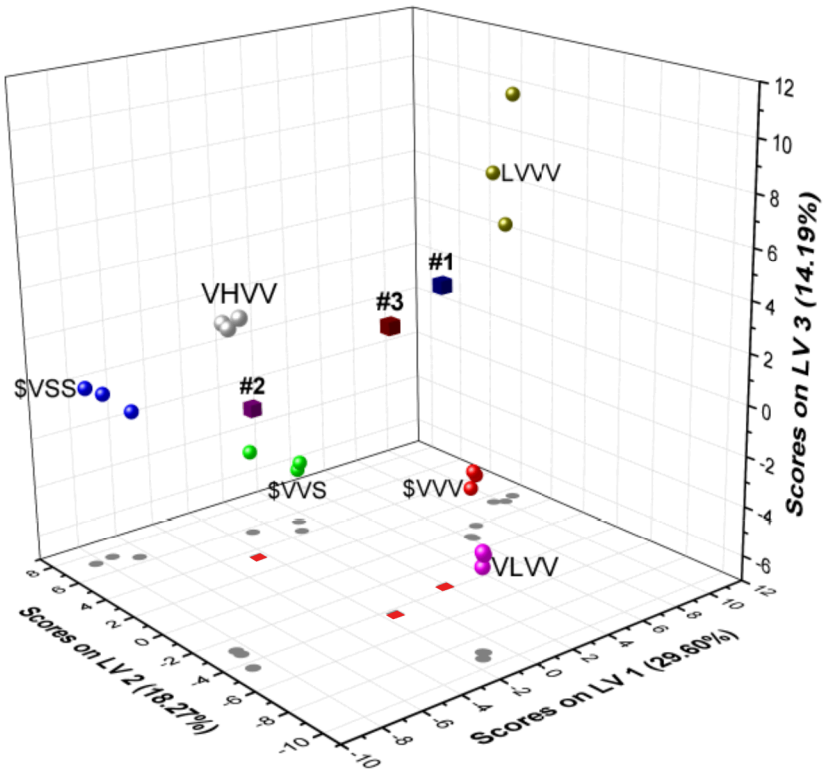


Figure 2. PLS-derived scores plot for the first and second latent variables (LVs) of the 18 training data sets. Gray dashed lines indicate groupings according to synthetic similarities.

PLS scores can also be used to assess in a qualitative sense the similarity of the unknown samples to those that formed the training set for our classification models.

Figure 3 shows a three-dimensional scores plot of a cross-validated (see below for details) PLS model. The 18 training set data are given as spheres. Added to this plot are the scores (as cubes) associated with the three datasets for the unknown pathways derived using the model. The x-y projection (gray shadows) is equivalent to the graph given in Figure 2. Note that the square shadows, corresponding to the unknown sample data (given in red for emphasis), do not necessarily closely cluster with the known sample scores. [e] This fact highlights both the necessity of statistical analyses to solving classification problems due to complex, ambiguous data *and* that different chemists in different laboratories, though following a standard procedure, clearly affect dramatically the ultimate signature profile of a crude synthesis material.



[e] It is useful to point out now that the "unknown" samples 1, 2, and 3 were later revealed to result from routes LVVV, $VSS, and VLVV, respectively.

Figure 3. Three-dimensional PLS-derived scores plot for first three LVs using the full LC+GC data set. X-Y projection in gray is given to help visualize separation between classes (plot same as in Figure 2, unknown sample shadows are in red). Cubes are blind data scores using the 18-sample training set; routes are labeled to be consistent with discussion found in main text.

Table 2 gives the results of feature selection using the PLS analysis. SR values are group-specific, so they can be tied directly to various synthetic routes and used as objectively determined CAS. These predictors were then isolated from the raw data sets and processed as described above to yield a reduced data set of 40 LC inputs and 7 GC inputs. We refer to this dataset as "Red2." Information on a second reduced dataset, "Red1," is found below.

Table 2. Route specific CAS identified using a PLSDA-derived selectivity ratio, SR_{min} = 1.5.* These predictors were used to form the Red2 dataset.

Method (Class)	LC-MS	GC-MS	ICP-MS
$VVV	7, 116, 130, 136	G	--
$VVS	16	I, J	--
$VSS	2, 4, 22, 42, 65, 68, 124	L	--
VHVV	9, 71, 98, 108, 114, 115, 123, 125, 126	F, U	--
VLVV	20, 34, 87, 99, 113, 118, 120	--	--
LVVV	19, 23, 25, 26, 27, 30, 35, 36, 37, 38, 77, 137	R	--

* Added to help aid in discrimination between $VVS and $VSS: 15, 31, 32, 43, 47, 51, 52, 54, 62, 67, 90, 96, 97, 132, 134. These predictors were added to those from Red2 to make the Red1 dataset.

Cross-validation and Prediction using Support Machine Vectors

In our previous work with fentanyl, we employed PLS Discriminant Analysis (PLSDA) for both feature selection and classification. PLSDA, however, generally classifies using binary results. That is, the probability that an unknown sample is

classified to a particular route will either be a "1" or a "0." In Eigenvector's Solo software's GUI, then, classification results from PLSDA offer little flexibility in assessing the degree to which unknowns belong to the other five routes in addition to that of the "most probable route." Additionally, invoking PLSDA for classification relies on specifics assumptions about the distributions of the predictors. Assuming that underlying distributions are either not well-characterized or just unknown, nonlinear discriminant analyses were investigated as being perhaps more suitable for classification.[10]

Support vector machines (SVM) are a class of nonlinear statistical models that have become some of the most flexible modeling tools available.[11] The development of a basic classification model using SVMs involves only two variables, the cost- and kernel-parameters. The more important of the two, the cost parameter (*c*), quantitates the penalty associated with erroneous classification (lower values equates with high tolerance for misclassification). Cross-validation of the model assigns the values of these two parameters, but discussion of how to properly cross-validate is also important. Due to the small sample size, random sampling of test data was chosen. That is, the software randomly chooses a subset of the training data to reserve as test data. The remaining training data is used to develop a model; the test data that was set aside is used for cross-validation. This is done for a certain number of repetitions to attain stable estimates of model performance. In general, it is desirable to have about 75-80% of the data present in the training set and the remainder reserved for validation. For the current work, four sets were used and 100 cross-validation repetitions were used to attain good performance stability. The results of cross-validation yielded models able to assign correctly all 18 samples. Note that this method of cross-validation was applied *regardless* of the

classification technique used, and keeping this method consistent allows for proper comparison of the models' results.

Three classification methods were studied for their ability to assign synthesis routes to the three unknown samples. They all vary in the details of their algorithms, but they all seek in one way or another to discover an underlying variable space that represents significant differences among the signature profiles of various synthetic routes. These techniques are known as support vector machines discriminant analysis (SVMDA), neural networks (NN), and logistic regression (LR). SVMDA and NN algorithms are popular tools when underlying relationships between predictors and response(s) are potentially nonlinear and unknown. Logistic regression is also used frequently and, while being a type of linear regression model, can predict responses for *categorical* classifications. That is, this type of model is good for predicting *discrete* outcomes like those encountered presently (i.e. either the route *is* VLVV, for example, or it *is not*).

In all, three datasets were analyzed for their ability to classify the three unknown crude 3MF samples:

1) Full GC+LC data set using all 167 predictors (141 from LC, 26 from GC).
2) A reduced data (Red2) set using only predictors with SR > 1.5 as determined from PLSDA analysis of the Full GC+LC data set (40 from LC, 7 from GC).
3) A second reduced data set (Red1) which includes all predictors from Red2 *plus* additional predictors with SR > 1.5 as determined from the PLSDA analysis of data from $VSS and $VVS *alone* (55 from LC, 7 from GC).

Again, it is important to stress that the reduced data sets were generated in an attempt to address the problem of over-fitting the data. It is likely, however, that still too many variables remain given the small size of the training set (n = 18 sample data sets).

Results from unknown route classification using Solo's SVM Discriminant Analysis (SVMDA) package are given in the left-most column in Figure 4. Results of running the model with different subsets of the original data are color-coded: red = full data set, green = "Red1" set, and blue = "Red2" set. Recall that Red2 (blue) used only those predictors for which SR > 1.5 using the full data set. Due to the general inability of any algorithm to discriminate confidently between $VVS and $VSS for unknown Test Sample Two, the set Red1 was developed to include those additional predictors.

Note that for all three unknown samples, classification probabilities are low; not one route stands apart significantly from the others. This is likely due to the fact that the model is too tightly tuned regardless of what sample set size you consider (full vs. reduced). The current SVMDA model used, generated largely as a black box, cannot be therefore easily generalized to new samples. The combination of the overly tuned model (i.e. high cost parameter reducing generalization) with the model being overfit (i.e. too high a predictor space for the number of samples) results in great uncertainty in the SVMDA predictions. Despite these facts, note that maximum probabilities predict Routes One, Two and Three to be LVVV, $VVS, and VLVV, respectively.

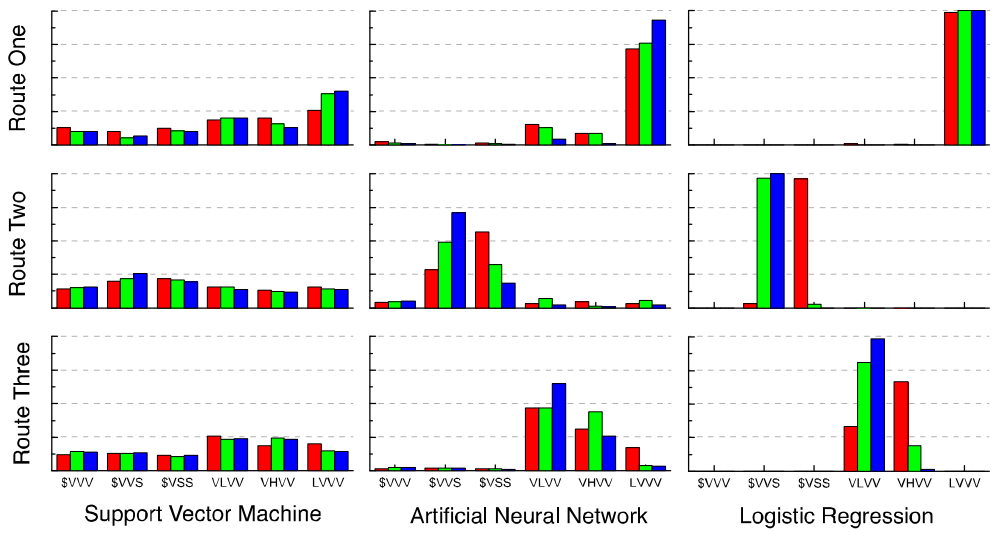


Figure 4. Classification results summary. Left: SVM; Middle: NN; Right: LR. Red: full data set; green: Red1 data set; blue: Red2 data set. Routes are given in descending order. Numerical probabilities have been removed for clarity, but gray lines represent probability, p = 0.25, 0.5, 0.75, and 1.0.

It is generally considered that NNs can be used to derive predictions on class assignments that are at least as competitive as or better than other machine learning counterparts.[10,12] Results from the NN analysis of the three data sets is shown in the middle panel of Figure 4. It is apparent that the trends are similar to SVMDA with regards to the most-probable route, though there appears to be more "confidence" in class assignments. This observation, however, may simply be due to a less tightly modeled training set, or that the hidden variables calculated by the NN algorithm can at least *rule out* certain synthesis routes more efficiently that the current SVMDA models.

Logistic regression (LR) is typically used for predictions where the outcome is discrete (i.e. "is" or "isn't" or, stated otherwise, the dependent variable is *catagorical*).[10] This type of model is particularly effective when the goal is solely classification. Note that classification probabilities are given for the three routes and three data set sizes in the

right-hand panel of Figure 4. The binary nature of the classification is most likely responsible for overall higher probabilities as compared to the other two routes. This fact may make the classifications look artificially confident. Nevertheless, the results seem to agree with those from the other two methods.

Models in Conjunction

Because none of the models presently is perfect (indeed, there is never *one right* answer when dealing with classification statistical analyses) it is often useful to look at results in concert. Considering all of the data and models in Figure 4, there exists agreement among all three algorithms in assignment of the unknown crude samples despite being all black box approaches. Results are summarized as follows:

1) Test Sample One is clearly assigned to LVVV; NN and LR have almost negligible probabilities of being otherwise save for minor assignment to VLVV or VHVV.
2) Test Sample Two seems to be either $VVS or $VSS with LR favoring $VVS for all reduced data sets and $VSS for the unabridged data.
3) Test Sample Three is *either* VHVV or VLVV according to SVM and NN; LR favors VLVV over others.
4) In general there is more ambiguity in assignment for sample datasets with more variables, but even more parsimonious data sets present issues with classification. Without less black box approaches, identifying the reasons and quantifying the resultant effects is challenging.

5) Consulting the chemists that performed the blind syntheses, two of the three routes were assigned correctly.

Table 3. Summary table of predicted and actual routes for unknown 3MF crude samples.

Unknown Sample ID	Predicted	Actual
Test Sample One	LVVV	**LVVV**
Test Sample Two	$VVS	**$VSS**
Test Sample Three	VLVV	**VLVV**

Chemical Context for Classifications

Test Sample One

Classification of Test Sample One appears to be the least ambiguous; clearly all algorithms favor assignment of LVVV to this blind sample. This route is quite unlike the others in that a) methylation of the piperidine ring was performed as part of the synthesis and b) this step was performed first, before reaction with 2-phenylethyl bromide. Though it may be considered closely related to VLVV in terms of reagents, clearly the order of synthetic steps causes significant underlying differences as revealed by the large separation by LV scores seen in Figures 1 and 2.

Test Sample Two

Overall, there appears to be the most confusion when classifying Test Sample Two. Indeed, this was the only one of the three test samples to be incorrectly identified on average. That is, the analyses seem to favor $VVS when the actual route used was $VSS. Recall that the differences between these two routes sits with the reductive amination step: $VVS relies on the Valdez method, which uses sodium triacetoxyborohydride at RT whereas $VSS relied on Siegfried's use of sodium borohydride, methanol, and reduced temperatures. At the end of either synthesis, little

conclusive evidence of their use (in the form of unique small molecules) will be found, particularly by LC. GC might be able to recover evidence of volatile methanol, but its presence will most likely be lost in the solvent delay. Solid Phase Micro-Extraction (SPME) methods coupled to GC-MS might then be a useful addition to the suite of analytical tools used.

Test Sample Three

Though this blind route was on average classified correctly as VLVV, there was some non-negligible probability that the route was VHVV, particularly for the SVM and NN analyses. Like the ambiguity with Test Sample Two, we believe there is a rational, chemical explanation for this fact. The difference between VLVV and VHVV is basically the reagents used to perform the methylation of 3-methylfentanyl's central piperidine ring: either lithium diisopropylamine (LDA) or sodium hydride. There are several issues with expecting unique signatures associated with this method:

1) Neither sodium nor hydride ions will be detected; ICP-MS might be thought of as helpful but in general sodium is found to be an unreliable element to monitor due to its ubiquity in synthesis.
2) Lithium may be useful but for the current data sets, all samples seemed to have some nonnegligible amount with no apparent trends or correlations to specific routes.
3) Direct evidence of diisopropylamine (from LDA's use) will be difficult as acylation in the final step converts this to *N*,*N*-diisopropylpropionamide, which is already a signature associated with using DIPEA as the base during acylation for

the Valdez step. Both VLVV and VHVV use this base, negating this amide compound a unique signature for the LDA route, VLVV.

Ultimately, however, the data analyses on average predict correctly that VLVV was the method used to synthesize Test Sample Three.

ICP-MS

Both semi-quantitative and quantitative ICP-MS results were used to assess any added value to the classification models. Semi-quantitative data was subjected to feature selection algorithms in an attempt to first find elements that were of statistical importance to describing overall variance. From the training data those elements were determined to be Co, Cu, Zn, Sn, and Sb. Samples were then rerun for quantitation purposes. Those data were then considered by statistical analysis as part of a "quantitative" LC+GC+ICP data set.

Quantitative data from the five elements selected from statistical analysis of semiquant data was incorporated into the unabridged LC+GC data set. Classification results using these data are given in Figure 5 in blue[f]. In general, even the quantitative data only adds to the ultimate misclassification of the unknown samples. Overall, the use of elemental data seems to only confound the statistical classification of new data relative to the training set. This is most likely due to a combination of factors, the testing of which was outside the scope of this work. Complete sample digestion would seem to homogenize the samples, making them more representative of the entire synthesis process

[f] Though not very instructive, analysis of the full GC+LC data set *including semiquant* ICP-MS data is also given for completion in green. The conclusion that ICP data only degrades model accuracy still applies.

(reagent trace metals, for example). However, these elemental profiles are most likely extremely dependent on synthetic chemist, particular reagent lots, glassware used, etc. and may only appear reproducible under highly controlled conditions like those used to generate the training sample set. Further research into these ideas is recommended.

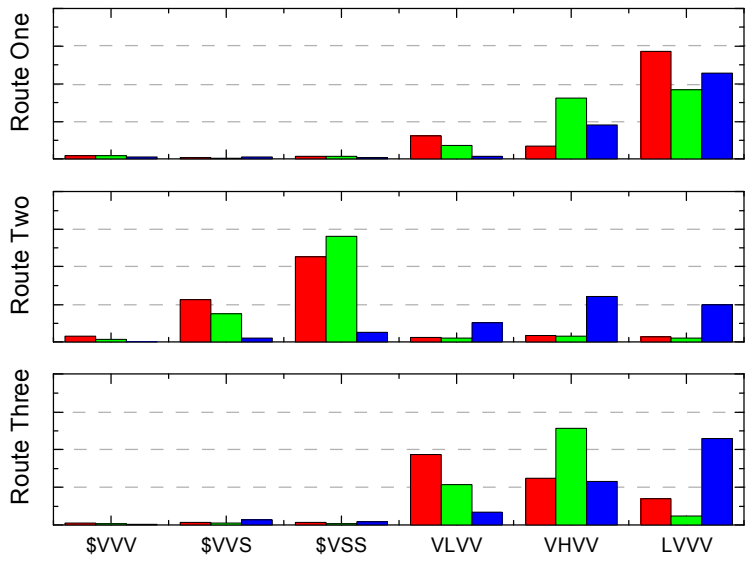


Figure 5. Comparison of NN classification analyses for LC+GC data (red), LC+GC with full semiquant ICP data (green), and LC+GC with quantitative ICP data from Co, Cu, Zn, Sn, and Sb (blue). Grey dashed lines represent p = 0.25, 0.50, and 0.75 for classification.

Conclusions

A variety of machine learning techniques were used to identify methods of synthesis of unknown crude 3-methylfentanyl samples based on LC-, GC, and ICP-MS analytical data. PLS scores showed that replicates of synthesis training data were similar and that samples of different routes were easily separable. More importantly, PLS was used to generate two reduced data sets in an attempt to address overfitting associated with the "curse of dimensionality." Neural networks, support vector machines, and logistic regression all predicted similar routes for the three unknown samples with varying levels

of apparent confidence. Model results were taken in concert due to the black box nature of the classification algorithms. Model predictions were correct for the most part, with a clear chemical rationale present for samples that were not classified with high certainty. ICP was shown to only confuse classification, and it was posited that this issue was due to high variability in elemental composition with respect to chemist, equipment, and synthetic conditions. Further investigation into the ultimately utility of ICP as a CAS tool is recommended.

Acknowledgements

The authors would like to thank Drs. Robert Sanner and H. Paul Martinez for their assistance in synthesizing the Test Samples. The authors would also like to thank the Department of Homeland Security, Science and Technology Directorate, Chemical Forensic Program for their funding of this work. Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344.

Appendix A. Comprehensive signature list of LC-QTOF data. "Source" refers the to route for which a particular compound was originally detected.

ID	RT (min)	Exact Mass	Formula	Source	ID	RT (min)	Exact Mass	Formula	Source
1	1.7	80.0478	C5H6N	$VSS	72	13	290.2688	C16H36NO3	$VVV
2	2	152.0702	C8H10NO2	$VSS	73	13.1	451.2955		VLVV
3	2.1	116.1432	C7H18N	VHVV	74	13.1	453.2887	C31H37N2O	$VVV
4	2.2	108.0804	C7H10N	$VSS	75	13.2	230.2482	C14H32NO	VHVV
5	2.3	130.1584	C8H20N	$VVV	76	13.2	274.2742	C16H36NO2	$VSS
6	2.3	315.2432	C20H31N2O	$VVV	77	13.2	348.1863	?	VHVV
7	2.3	333.2539	C20H33N2O2	$VVV	78	13.2	455.3031	C29H40N2NaO	$VVV
8	2.3	373.3042	C20H41N2O4	VLVV	79	13.3	330.222	C24H28N	VLVV
9	5.4	113.0598	C6H9O2	$VVV	80	13.4	226.179	C13H24NO2	$VSS
10	5.5	222.1497	C13H20NO2	VLVV	81	13.4	228.1594	C12H22NO3	$VVV
11	6.1	206.1539	C13H20NO	LVVV	82	13.4	247.1805	C15H23N2O	LVVV
12	6.4	252.1586	C14H22NO3	$VVV	83	13.4	415.2961	C25H39N2O3	VLVV
13	8.7	218.1533	C14H20NO	$VVV	84	13.5	216.1385	C14H18NO	$VVV
14	8.7	236.1651	C14H22NO2	$VVV	85	13.5	266.1725	C15H24NO3	VHVV
15	9.4	220.1702	C14H22NO	LVVV	86	13.6	318.3007	C18H40NO3	LVVV
16	10.1	220.1702	C14H22NO	LVVV	87	13.6	440.273	C29H34N3O	VLVV
17	10.1	264.1965	C16H26NO2	VLVV	88	13.6	493.323	C34H41N2O	LVVV
18	10.4	198.1295	C14H16N	$VVV	89	13.7	303.2074	C18H27N2O2	$VVV
19	10.5	250.1803	C15H24NO2	LVVV	90	13.7	304.1586	?	$VSS
20	10.5	264.1965	C16H26NO2	VLVV	91	13.8	158.1538	C9H20NO	$VVV
21	10.5	290.1786	C17H24NO3	$VVV	92	13.9	213.1463	?	VHVV
22	10.6	287.1383	C16H19N2O3	$VSS	93	13.9	288.2528	C16H34NO3	$VSS
23	10.6	407.27	C27H37N2O2	LVVV	94	13.9	507.3396	C35H43N2O	LVVV
24	10.6	434.3192	C28H40N3O	$VVV	95	14	136.0747	C8H10NO	VLVV
25	10.7	232.1703	C15H22NO	LVVV	96	14	178.1237	C11H16NO	LVVV
26	10.7	244.1696	C16H22NO	LVVV	97	14.1	113.0591	C6H9O2	VHVV
27	10.7	389.2595	C26H33N2O	LVVV	98	14.1	191.0685	C9H12NaO3	VHVV
28	10.8	448.3349	C29H42N3O	$VVV	99	14.1	232.1695	C15H22NO	VLVV
29	10.9	232.1703	C15H22NO	LVVV	100	14.1	272.1622	C15H23NNaO2	VLVV
30	10.9	403.2747	C27H35N2O	LVVV	101	14.2	303.1712	C17H23N2O3	$VVV
31	11.1	275.2123	C17H27N2O	$VVV	102	14.2	309.1959	C20H25N2O	VLVV
32	11.2	308.1853	C17H26NO4	$VVV	103	14.3	274.1351	C18H16N3	$VSS
33	11.2	341.222	C21H29N2O2	LVVV	104	14.3	293.1102	?	$VSS
34	11.3	276.1608	C16H22NO3	VLVV	105	14.3	398.3666	C24H48NO3	$VSS
35	11.3	403.2747	C27H35N2O	LVVV	106	14.5	236.1628	C12H23NNaO2	VHVV
36	11.4	244.1696	C16H22NO	LVVV	107	14.8	150.0912	C9H12NO	$VVV
37	11.4	246.1847	C16H24NO	LVVV	108	14.8	169.0858	C9H13O3	VHVV
38	11.4	417.2903	C28H37N2O	LVVV	109	14.8	206.1176	C12H16NO2	$VVV
39	11.5	214.1201	C14H16NO	$VVV	110	14.8	228.0993	C12H15NNaO2	$VVV
40	11.5	262.1795	C16H24NO2	$VVV	111	15	274.1794	C17H24NO2	$VVV
41	11.6	355.2382	C22H31N2O2	LVVV	112	15	296.162	C17H23NNaO2	$VVV
42	11.7	136.0751	C8H10NO	LVVV	113	15	332.1883	C20H22N5	VLVV
43	11.7	276.1961	C17H26NO2	$VVV	114	15.4	169.0858	C9H13O3	VHVV
44	11.8	290.1786	C17H24NO3	VLVV	115	15.4	247.0956	?	VHVV
45	11.8	323.2124	C21H27N2O	VLVV	116	15.4	312.197	C20H26NO2	$VVV
46	11.9	218.212	C12H28NO2	VHVV	117	15.5	306.2433	C19H32NO2	$VVV
47	12	234.2069	C12H28NO3	VHVV	118	15.5	413.2229	C27H29N2O2	VLVV
48	12	281.2015	C19H25N2	LVVV	119	15.6	325.1916	C20H25N2O2	LVVV
49	12	337.2272	C22H29N2O	$VVV	120	15.7	379.2039	C26H28NaO	VLVV
50	12.1	292.1907	C17H26NO3	$VSS	121	15.7	443.2721	?	$VVV
51	12.2	351.2436	C23H31N2O	$VVV	122	15.8	393.2529	C25H33N2O2	$VVV
52	12.3	332.2222	C20H30NO3	$VSS	123	15.9	254.155	C17H20NO	$VVV
53	12.3	345.1974	C23H25N2O	$VVV	124	15.9	381.2176	C23H29N2O3	$VVV
54	12.4	185.1134	C8H18NaO3	VHVV	125	16.1	282.1859	C19H24NO	$VVV
55	12.5	295.2171	C20H27N2	LVVV	126	16.1	304.1685	C21H22NO	$VSS
56	12.5	349.2279	C23H29N2O	VHVV	127	16.1	348.1577	?	$VSS
57	12.5	367.236	C23H31N2O2	$VSS	128	16.2	354.2883	C18H36N5O2	VLVV
58	12.5	407.2695	C26H35N2O2	$VSS	129	16.3	200.202	C12H26NO	$VSS
59	12.5	453.2753		VLVV	130	16.3	393.2165	C24H29N2O3	$VVV
60	12.6	246.2429	C14H32NO2	VHVV	131	16.4	289.1443	C17H21O4	VLVV
61	12.6	262.237	C14H32NO3	$VSS	132	16.5	241.1479	?	$VSS
62	12.6	365.2232	C23H29N2O2	$VVV	133	16.5	246.1476	C13H32NNaO3	$VSS
63	12.6	492.3235	C30H42N3O3	$VVV	134	16.5	395.2341	C24H31N2O3	$VVV
64	12.7	94.0638	C6H8N	LVVV	135	16.6	235.1449	C13H19N2O2	LVVV
65	12.7	136.0751	C8H10NO	$VSS	136	16.7	405.2554	C26H33N2O2	$VVV
66	12.7	150.0912	C9H12NO	$VVV	137	17	339.2074	C21H227N2O2	LVVV
67	12.7	202.2177	C12H28NO	VHVV	138	17.3	339.2074	C21H227N2O2	LVVV
68	12.7	232.0937	C13H14NO3	$VSS	139	18	281.1739	?	$VSS
69	12.8	144.1371	C8H18NO	$VVV	140	18.1	353.2224	C22H29N2O2	LVVV-3
70	12.9	172.0731	C9H11NNaO	VHVV	141	20.4	287.2227	C17H27N4	LVVV
71	13	209.0793	???	VHVV					

Appendix B. Comprehensive GC signature list with chemical names given for library match factors (reverse, forward, or both) greater than 850. If the name is in italics, the match favor was between 800 and 850.

ID	RT (min)	Exact Mass	Chemical Name
A	3.8	79.0414	pyridine
B	5.17	72.0805	diisopropylethylamine
C	6.92	104.0614	styrene
D	7.62	57.0332	propanoic acid, anhydride
E	8.91	93.0559	aniline
F	10.32	91.0536	Benzeneacetaldehyde
G	10.7	78.0468	benzene
H	11.68	91.0537	2-chloroethyl benzene
I	11.91	72.0802	Acetamide, N,N-dipropyl-
J	12.28	86.096	N,N-bis(1-methylethyl)-propanamide
K	13.31	91.0539	2-bromoethyl benzene
L	16.08	104.0615	Propanoic acid, 2-phenylethyl ester
M	16.5	93.0568	Acetamide, N-phenyl
N	17.2	106.0648	Benzenamine, N-ethyl-
O	17.77	93.0569	Propanamide, N-phenyl
P	19.39	93.0568	Dipropionylaniline
Q	22.43	126.0907	4-Piperidone, 3-methyl-1-phenethyl
R	22.49	140.1065	Piperidin-4-one, 2,3-dimethyl-1-phenethyl-
S	23	140.1063	Piperidin-4-one, 2,3-dimethyl-1-phenethyl-
T	25.08	170.1166	1-phenethyl-4-propionyloxypiperidine
U	25.43	184.1326	*1-Aminocyclopentanecarboxylic acid, N-isobutoxycarbonyl-, isohexyl ester*
V	26.36	104.0613	*Butylphosphonic acid, di(2-phenylethyl) ester*
W	26.59	104.0616	*pimelic acid, di(phenethyl) ester*
X	33.27	245.1641	Fentanyl
Y	33.28	259.1783	3-Methylfentanyl
Z	33.65	259.1799	3-Methylfentanyl

Appendix C. PLSDA-derived CAS broken down by synthesis route. GC CAS have names only. Blanks indicate a reasonable formula could not be derived from the LC-QTOF exact mass data.

$VVV

ID	R.T. (min)	Exact Mass	Formula/Name
7	2.3	333.2539	C20H33N2O2
116	15.4	312.197	C20H26NO2
130	16.3	393.2165	C24H29N2O3
136	16.7	405.2554	C26H33N2O2
G	10.7	78.0468	benzene

$VVS

ID	R.T. (min)	Exact Mass	Formula/Name
16	10.1	220.1702	C14H22NO
I	11.91	72.0802	Acetamide, N,N-dipropyl-
J	12.28	86.096	N,N-bis(1-methylethyl)-propanamide

$VSS

ID	R.T. (min)	Exact Mass	Formula/Name
2	2	152.0702	C8H10NO2
4	2.2	108.0804	C7H10N
22	10.6	287.1383	C16H19N2O3
42	11.7	136.0751	C8H10NO
65	12.7	136.0751	C8H10NO
68	12.7	232.0937	C13H14NO3
124	15.9	381.2176	C23H29N2O3
L	16.08	104.0615	Propanoic acid, 2-phenylethyl ester

VHVV

ID	R.T. (min)	Exact Mass	Formula/Name
9	5.4	113.0598	C6H9O2
71	13	209.0793	
98	14.1	191.0685	C9H12NaO3
108	14.8	169.0858	C9H13O3
114	15.4	169.0858	C9H13O3
115	15.4	247.0956	
123	15.9	254.155	C17H20NO
125	16.1	282.1859	C19H24NO
126	16.1	304.1685	C21H22NO
F	10.32	91.0536	Benzeneacetaldehyde
U	25.43	184.1326	

VLVV

ID	R.T. (min)	Exact Mass	Formula/Name
20	10.5	264.1965	C16H26NO2
34	11.3	276.1608	C16H22NO3
87	13.6	440.273	C29H34N3O
99	14.1	232.1695	C15H22NO
113	15	332.1883	C20H22N5
118	15.5	413.2229	C27H29N2O2
120	15.7	379.2039	C26H28NaO

LVVV

ID	R.T. (min)	Exact Mass	Formula/Name
19	10.5	250.1803	C15H24NO2
23	10.6	407.27	C27H37N2O2
25	10.7	232.1703	C15H22NO
26	10.7	244.1696	C16H22NO
27	10.7	389.2595	C26H33N2O
30	10.9	403.2747	C27H35N2O
35	11.3	403.2747	C27H35N2O
36	11.4	244.1696	C16H22NO
37	11.4	246.1847	C16H24NO
38	11.4	417.2903	C28H37N2O
77	13.2	348.1863	
137	17	339.2074	C21H227N2O2
R	22.49	140.1065	2,3-dimethyl-1-phenethylpiperidin-4-one

REFERENCES

[1] Smith, F. P.; Siegel, J. A. *Handbook for Forensic Drug Analysis*, 1st ed.; Academic Press: San Diego, CA, 2004.

[2] a) Williams, AM; Mulcahy, HA; Leif, RN; Valdez, CA; Vu, AK. Extraction and Detection of US VX and its Chemical Attribution Signatures in Food Matrices Report to DHS Chemical Forensics Program, October 2012. (FOUO); b) Williams, AM; Vu, AK. Extraction and Detection of Russian VX and its Chemical Attribution Signatures in Food Matrices. Report to DHS Chemical Forensics Program, August 2012. (FOUO); c) Hok, S.; Hart, R.B. "Pathway Signatures and Analysis of Small-Scale Production Methods of US VX." DHS Chemical Forensics and Attribution Program. Jan. 2010. (FOUO); d) Hok, S.; Valdez, A. C.; Alcaraz, A. "US-Swede Collaboration: Pathway Signatures and Analysis of Small-Scale Production Methods of Russian VX." DHS Chemical Forensics and Attribution Program. (FOUO); e) Mayer, B. P.; DeHope, A. J.; Mew, D. A.; Spackman, P. E.; Williams, A. M. *Anal. Chem.* **2016**, *88*, 4303-4310.

[3] Henderson, G. L. *J. Forensic Sci.* **1988**, *22*, 569-575.

[4] CDC. Influx of Fentanyl-laced Counterfeit Pills and Toxic Fentanyl-related Compounds Further Increases Risk of Fentanyl-related Overdose and Fatalities. https://emergency.cdc.gov/han/han00395.asp (accessed Sept 7, 2015).

[5] Janssen, P. A. J.; Gardocki, J. F. *Method for Producing Analgesia.* U.S. Patent 3,141,823, July 21, 1964.

[6] Valdez, C. A.; Leif, R. N.; Mayer, B. P. *PLoS ONE* **2014**, *9*, e108250.

[7] Seigfried. Synthesis of Fentanyl. https://www.erowid.org/archive/rhodium/chemistry/fentanyl.html (accessed Aug 31, 2015).

[8] Straight, S. D.; Gardner, M. W.; Krueger, C. J.; Muthu, P. J.; Harry, E. B.; Manley, T. E.; Acevedo, C. A. Chemical Attribution Signatures of Fentanyl Derivatives Identified by the Application of Data Analytics. DHS Chemical Forensics Program, September 2015.

[9] Hughes, G.F. *IEEE T. Inform. Theory* **1968**, *14,* 55–63.

[10] Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, 2013.

[11] Mahadevan, S.; Shah, S. L.; Marrie, T. J.; Slupsky, C. M. *Anal. Chem.* **2008**, *80*, 7562-7570.

[12] Goodacre, R.; Vaidyanathan, S.; Dunn, W. B.; Harrigan, G. G.; Kell, D. B. *Trends Biotechnol.* **2004**, *22*, 245-252.