Supplementary Information

High-throughput identification of FLT3 wild-type and mutant kinase substrate preferences and application to design of sensitive *in vitro* kinase assay substrates

Minervo Perez[1,2], John Blankenhorn[1], Kevin J. Murray[3], Andy W. Tao[2], Laurie L. Parker[1]

Institutional Affiliations:

[1]University of Minnesota, Department of Biochemistry, Molecular Biology and Biophysics, 420 Washington Avenue SE, Minneapolis, MN 55455

[2]Purdue University, Department of Medicinal Chemistry and Molecular Pharmacology, 201 S. University Street, West Lafayette, IN 47907

[3]University of Minnesota, Department of Veterinary Population Medicine, 319 15th Avenue South East, Minneapolis, MN 55455

Figures:

Tables

Methods

Extraction and reformatting of phosphopeptide sequences from peptide ID results—The KinaMINE data formatter (Kinamine.jar) uses the Distinct Peptide Report and the FASTA file that was used in the proteomics search engine as input, filters the peptides from the report with a threshold of 1% FDR, to consolidate the sequences of all peptides that were phosphorylated in the experiment. It then outputs a .csv table (the "Positive Substrates" file, which is named by the user at the time of running the script) of those tyrosine-phosphorylated sequences, with each amino acid separated into an individual column and the phosphotyrosine aligned. This table also contains the accession number of the protein each peptide was from, which is used to extract the sequences of those proteins from the inputted FASTA file and calculate the "Substrate Background Frequency" (frequency of the 20 canonical amino acids found in each of the proteins individually; SBF), also output as a .csv. This .csv file also reports the total number of tyrosine residues within those protein sequences and the number of those tyrosine residues that were observed as phosphorylated in the experiment for subsequent use in determining FLT3's "normalization score" in the Screener module of KINATEST-ID (described below).

Phosphopeptide list comparison filtering—To select the sequences that were phosphorylated in common between the WT and the two mutant forms of FLT3, we developed a filtering script in R ("Similarity and Difference Finder.R") to extract sequence lists and generate corresponding Substrate Background Frequency tables for the proteins corresponding to the selected peptides. This script provides either the intersection or symmetric difference between those sets as two new output tables containing only the information relevant to the sequences desired.

Approximating most likely "true negative" sequence list from substrate dataset—The accession numbers for proteins that remain in the Substrate Background Frequency list after the previous filter are submitted to the reviewed human Uniprot/SwissProt database (http://uniprot.org/uploadlist/) to generate a FASTA file containing the sequences of those

proteins. The FASTA file is converted separately to .csv format using a script obtained from (https://www.researchgate.net/post/Converting_a_fasta_file_to_a_tab-delimited_file10). This file and the filtered Positive Substrates list file (generated as described in the previous section) are used as input for the “NegativeMotifFinder.R” to extract additional tyrosine-containing sequences from those proteins that could in principle have been phosphorylated but were not detected (outputting a “Negative Motifs” .csv file that is named by the user upon running the script). “Negative Motifs” files and corresponding “Positive Substrates” files are later used by the Kinatestpart1.R script to calculate Matthews Correlation Coefficient (MCC) values that give a general threshold for which peptides will or will not be phosphorylated by the kinase of interest.

Tables

Substrate	Substrate Sequence	Molecular Weights (g/mol)	[M+(1)H]	[M+(2)H]	[M+(3)H]	[M+(4)H]	[M+(5)H]	[M+(6)H]	[M+(7)H]	[M+(8)H]
FL-ABLtide	EAIYAAPFAKKBGGGAPTYSPPPPPGGRKKRRQRRLL	4346.15	N/A	2174.4	1449.8	1087.8	870.3	725.4	622.0	544.5
FLT3tide	FTDRLQQYISTRGGBGG	2109.37	2109.9	1055.7	703.8	N/A	N/A	N/A	N/A	N/A
A	GGDEDNDNYCNPNEEGGBGG	2265.26	2264.2	1132.1	N/A	N/A	N/A	N/A	N/A	N/A
B	GGDED**S**D**D**Y**F**NPNEEGGBGG	2283.24	2284.9	1143.1	N/A	N/A	N/A	N/A	N/A	N/A
C	GGDED**S**D**I**Y**A**NPNEEGGBGG	2205.22	2206.6	1114.9	N/A	N/A	N/A	N/A	N/A	N/A
D	GGDED**S**DNY**F**NPNEEGGBGG	2282.26	2281.9	1140.6	N/A	N/A	N/A	N/A	N/A	N/A
E	GGDED**S**D**I**Y**F**NPNEEGGBGG	2281.31	2282.9	1152.7 (M+Na)	N/A	N/A	N/A	N/A	N/A	N/A
F	GGDED**S**DNY**F**N**F**NEEGGBGG	2332.32	2334.7	1167; 1179 (M+Na); 1186.6 (M+K)	N/A	N/A	N/A	N/A	N/A	N/A
G	GGDED**SND**Y**F**N**T**NEEGGBGG	2286.25	2287.9	1155.1 (M+Na)	N/A	N/A	N/A	N/A	N/A	N/A
H	GGDED**HNQ**Y**EQ**PNEEGGBGG	2341.33	2343.1	1172.1; 1182.6 (M+Na)	781.8	N/A	N/A	N/A	N/A	N/A

Table S1. A summary of the FLT3 Artificial Substrate (FAS) candidate sequences synthesized and assayed in vitro with recombinant FLT3 variants. Abltide (EAIYAAPFAK; the substrate has been incorporated with an SH3 recognition and cell penetrating sequence and termed FL-Abltide) is a previously known FLT3 peptide substrate and has been used a reference substrate to monitor kinase activity.[1] The substrate sequences derived from the KINATEST-ID pipeline are underlined and were synthesized within the terbium binding motif shell (amino acids not underlined; sequence generated using the Aligner module of KINATEST-ID) with a biotinylated lysine (B) as an enrichment tag. The “Molecular Weights” column summarizes the theoretical weight (molecular weight) of the synthesized peptide sequences. The additional right-hand columns summarize the

major observed mass (M) to charge (m/z) signals [M+(n)H] for each peptide's LC-MS analysis. Sodium (Na) or potassium (K) adducts were present in the second charge state [M+2H] for FAStides-E-H. Charge states not observed are denoted as N/A.

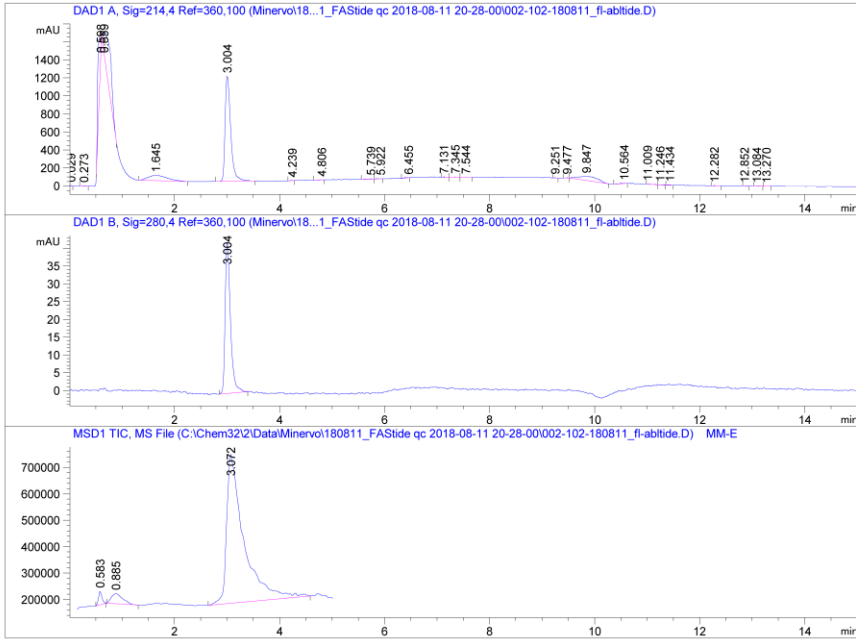
Scoring Model	Database size	MCC	Sensitivity	Specificity	Accuracy	Precision	EER	AROC	Threshold
WT-2H	888	0.39	0.73	0.81	80.28	0.31	0.20	0.86	17
WT-OVLP	559	0.38	0.79	0.76	76.44	0.29	0.24	0.85	17
WT-16H	1559	0.34	0.91	0.55	60.95	0.28	0.39	0.78	12
D835Y-16H	2010	0.35	0.92	0.56	60.82	0.30	0.39	0.78	13
ITD-16H	344	0.43	0.54	0.92	88.58	0.45	0.11	0.89	32
SHARED-16H	244	0.45	0.66	0.89	86.50	0.42	0.59	0.89	45

Table S2. Performance metrics and comparison of PSM models. *Scoring model* gives the substrate lists used to develop the scoring model. *Database size* represents the number of substrates in the list. The Matthew's correlation coefficient (*MCC*) is a performance metric for binary classifiers with values within a -1 to +1 scale.[2,3] A value close to 0 indicates a model's prediction is random while a value close to 1 indicates a perfect prediction. A value close to -1 indicates the model is making inverse predictions. *Sensitivity* is the true positive rate (recall) of each scoring model. *Specificity* is the true negative rate (relative to the input dataset). *Accuracy* is the measure of a prediction model's ability to correctly predict an outcome's true classification (i.e. positives vs. negatives from the input dataset). *Precision* is the rate of a model's ability for predicting true results from all its predictions. Equal error rate (*EER*) is the rate where the acceptance and rejection errors are the same. Area under the receiver operator curve (*AROC*) describes the number of correct predictions (of true positives or true negatives) at each given score. *Threshold* is the chosen value for binary classification for each predictive model. The lack of balance in the dataset (i.e. more true positives than true negatives or the inverse) is a potential caveat of these metrics. MCC-based performance metrics are shown to be compatible with imbalanced datasets. However, optimized classifying metrics for imbalanced datasets have been developed[2,4] but require implementation of Bayesian statistics or the development of a support vector machines.[2,5] Based on the satisfactory performance of KINATEST-ID for the applications pursued so far, these

have not yet been examined. However, advanced performance metrics should be considered in future updates of the KINATEST-ID predictive models.

FL-ABLtide

MS Peak Purity Range Report
Data File C:\Chem32\...o\180811_FAStide qc 2018-08-11 20-28-00\002-102-180811_fl-abltide.D
Sample Name: 180811_fl-abltide



==========

Fraction Information

==========

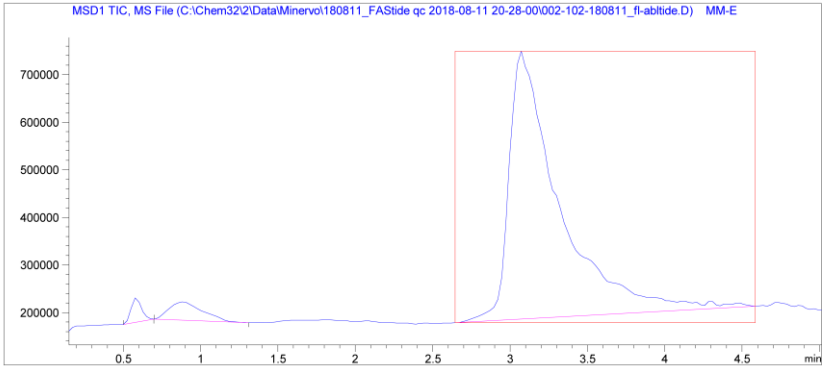
Fraction collection off

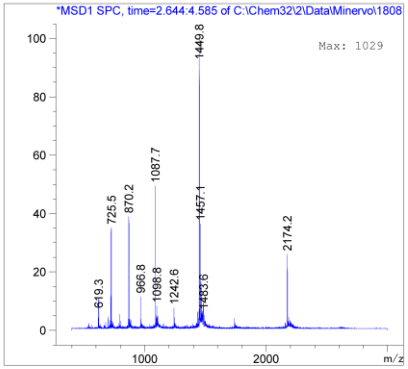
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...o\180811_FAStide qc 2018-08-11 20-28-00\002-102-180811_fl-abltide.D
Sample Name: 180811_fl-abltide





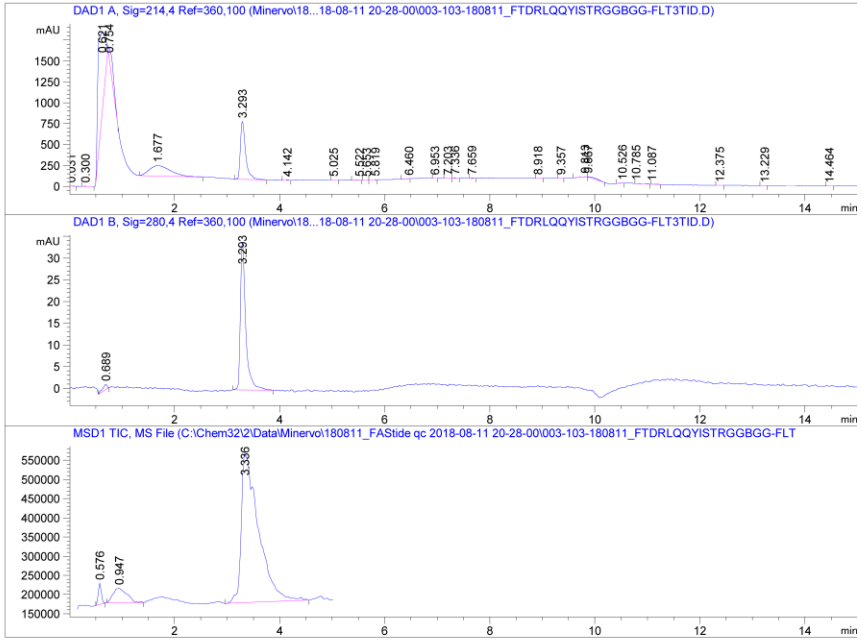
MSCalcPurity Error # 407!!!

Peak #3 at 3.072 min (2.644 to 4.585 min)
-> No purity results available. <-

*** End of Report ***

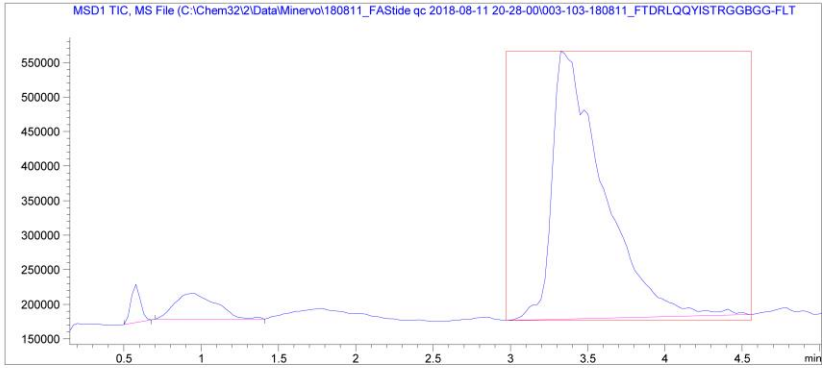
FLT3tide

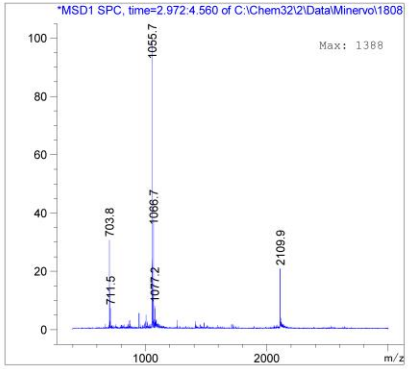
MS Peak Purity Range Report
Data File C:\Chem32\...e qc 2018-08-11 20-28-00\003-103-180811_FTDRLQQYISTRGGBGG-FLT3TID.D
Sample Name: 180811_FTDRLQQYISTRGGBGG-FLT3TIDE




```
================================================================================
                        Fraction Information
================================================================================
Fraction collection off
================================================================================
No Fractions found.
================================================================================
```


MS Peak Purity Range Report
Data File C:\Chem32\...e qc 2018-08-11 20-28-00\003-103-180811_FTDRLQQYISTRGGBGG-FLT3TID.D
Sample Name: 180811_FTDRLQQYISTRGGBGG-FLT3TIDE





MSCalcPurity Error # 1!!!

Peak #3 at 3.336 min (2.972 to 4.557 min)
-> No purity results available. <-

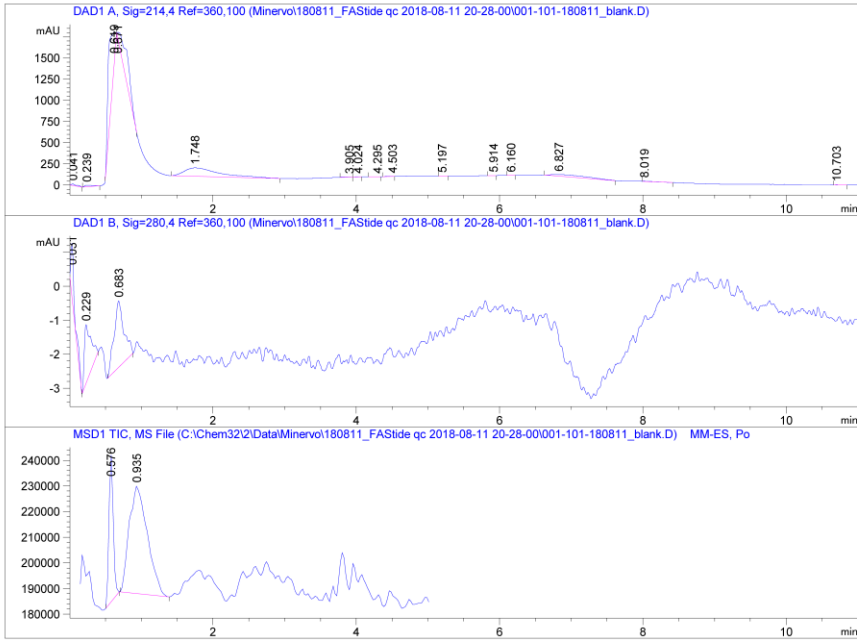
*** End of Report ***

HPLC-MS Analytical Blank

MS Peak Purity Range Report
Data File C:\Chem32\...inervo\180811_FAStide qc 2018-08-11 20-28-00\001-101-180811_blank.D
Sample Name: 180811_blank

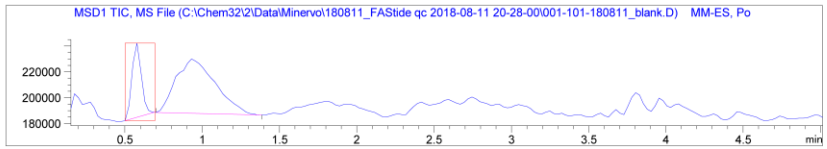
DMSO IN PBS diluted with hplc water with 0.1% fa

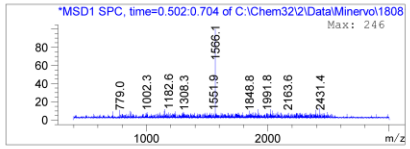
```
Acq. Operator    : SYSTEM                        Seq. Line :   1
Acq. Instrument  : ChemStation Online LC-MS       Location :   101
Injection Date   : 8/11/2018 8:29:36 PM               Inj :   1
                                                 Inj Volume : 50.000 µl
Acq. Method      : C:\Chem32\2\Data\Minervo\180811_FAStide qc 2018-08-11 20-28-00\180223_
                   ANALYTICA BLANK-HPLC-MS.M
Last changed     : 8/11/2018 8:28:00 PM by SYSTEM
Analysis Method  : C:\Chem32\2\Data\Minervo\180811_FAStide qc 2018-08-11 20-28-00\180223_
                   ANALYTICA BLANK-HPLC-MS.M (Sequence Method)
Last changed     : 8/12/2018 12:30:52 AM by SYSTEM
Sample Info      : DMSO IN PBS diluted with hplc water with 0.1% fa
```




```
================================================================================
                          Fraction Information
================================================================================
Fraction collection off
================================================================================
No Fractions found.
================================================================================
```

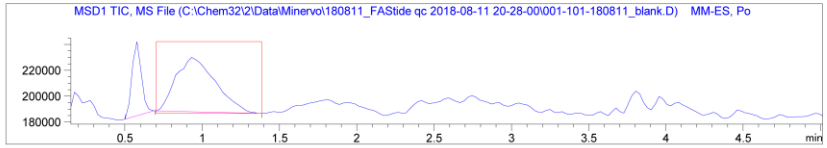

MS Peak Purity Range Report
Data File C:\Chem32\...inervo\180811_FAStide qc 2018-08-11 20-28-00\001-101-180811_blank.D
Sample Name: 180811_blank

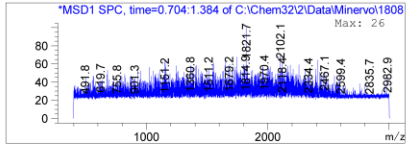




MSCalcPurity Error # 407!!!

Peak #1 at 0.576 min (0.502 to 0.694 min)
-> No purity results available. <-





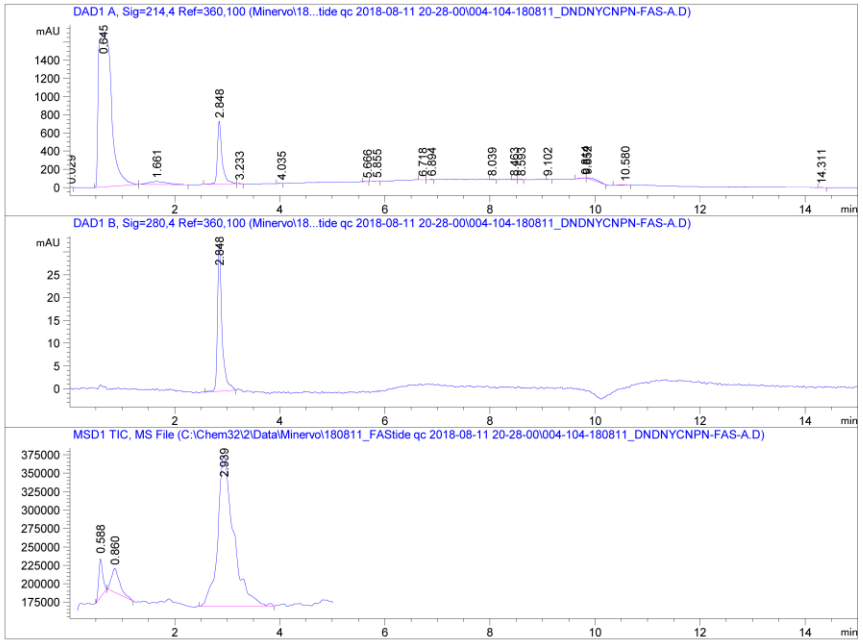
MSCalcPurity Error # 407!!!

Peak #2 at 0.935 min (0.704 to 1.384 min)
-> No purity results available. <-

*** End of Report ***

FAStide-A

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\004-104-180811_DNDNYCNPN-FAS-A.D
Sample Name: 180811_DNDNYCNPN-FAS-A



==========

Fraction Information

==========

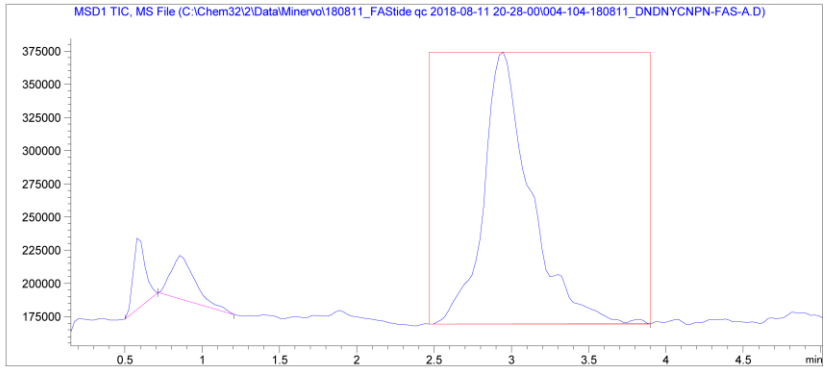
Fraction collection off

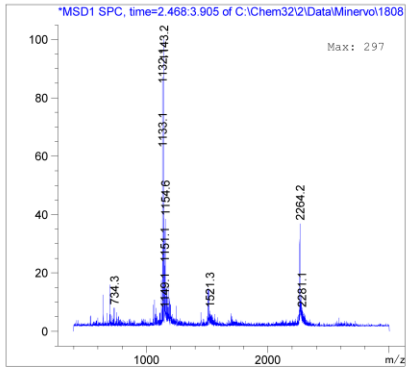
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\004-104-180811_DNDNYCNPN-FAS-A.D
Sample Name: 180811_DNDNYCNPN-FAS-A





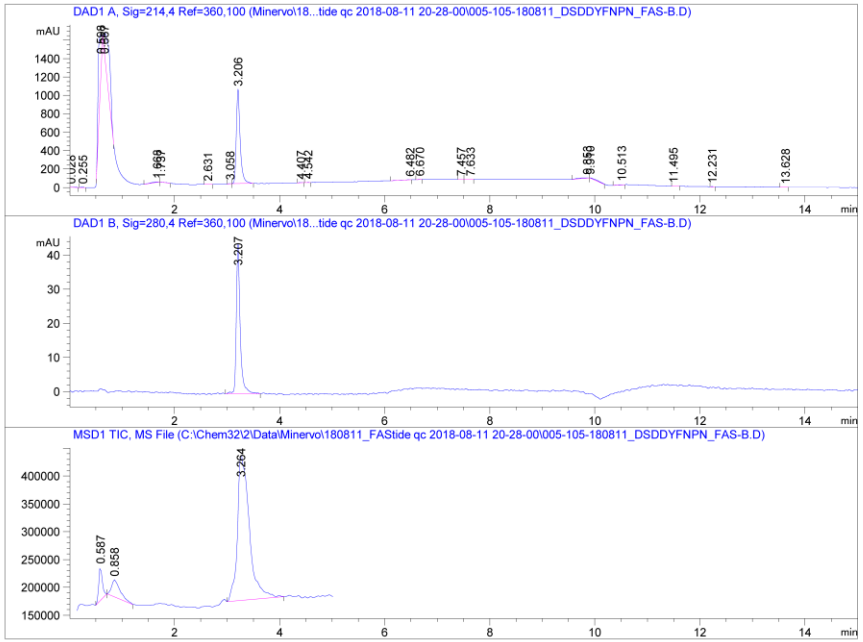
MSCalcPurity Error # 407!!!

Peak #3 at 2.939 min (2.468 to 3.897 min)
-> No purity results available. <-

*** End of Report ***

FAStide-B

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\005-105-180811_DSDDYFNPN_FAS-B.D
Sample Name: 180811_DSDDYFNPN_FAS-B



==========

Fraction Information

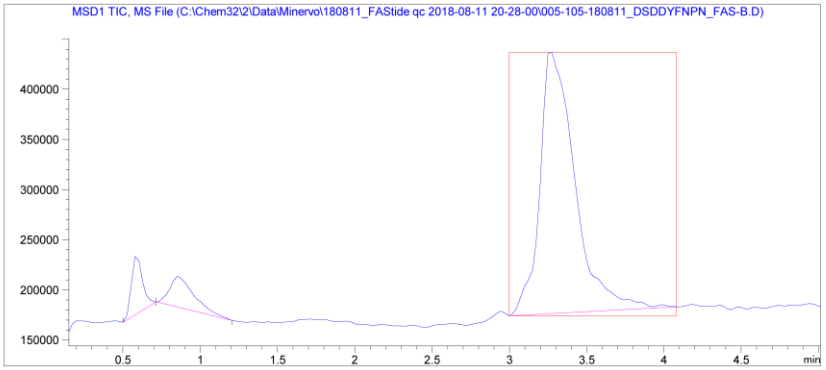
==========

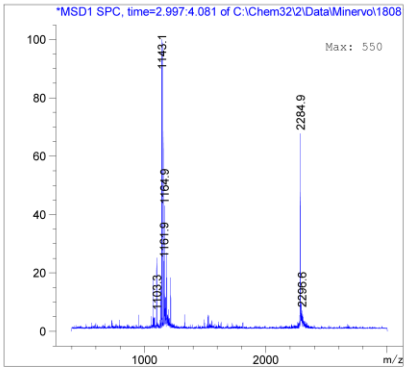
Fraction collection off

==========

No Fractions found.

==========





MSCalcPurity Error # 407!!!

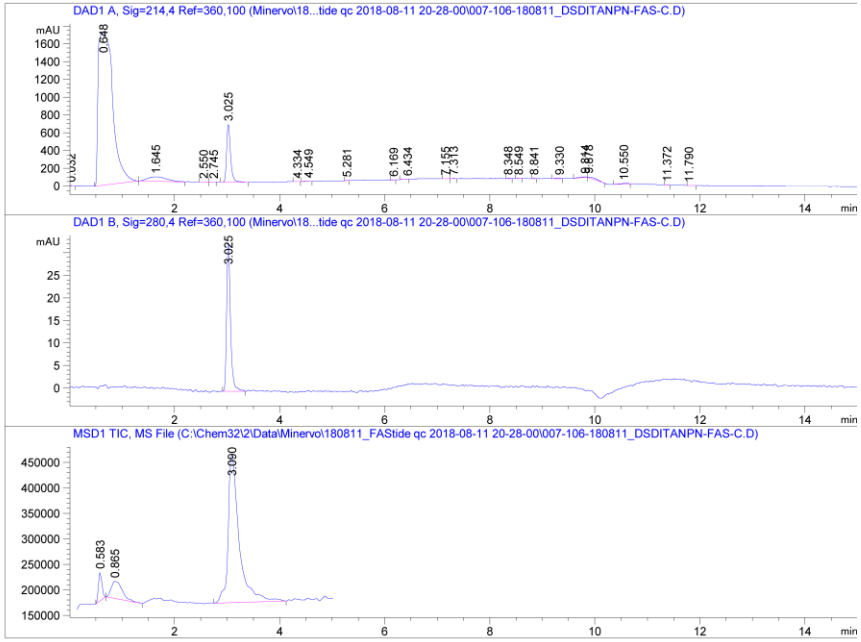
Peak #3 at 3.264 min (2.997 to 4.081 min)
-> No purity results available. <-

*** End of Report ***

ChemStation Online LC-MS 8/16/2018 12:57:41 PM SYSTEM Page 2 of 2

FAStide-C

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\007-106-180811_DSDITANPN-FAS-C.D
Sample Name: 180811_DSDITANPN-FAS-C



==========

Fraction Information

==========

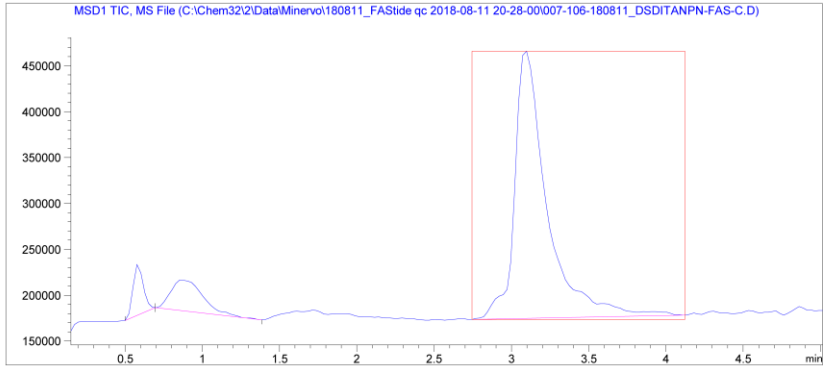
Fraction collection off

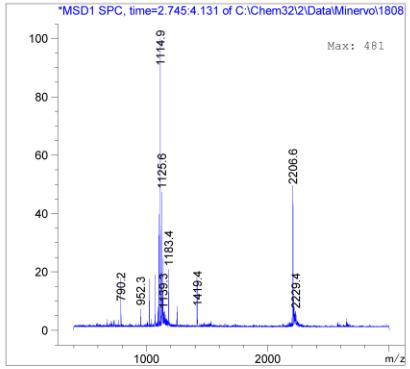
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\007-106-180811_DSDITANPN-FAS-C.D
Sample Name: 180811_DSDITANPN-FAS-C





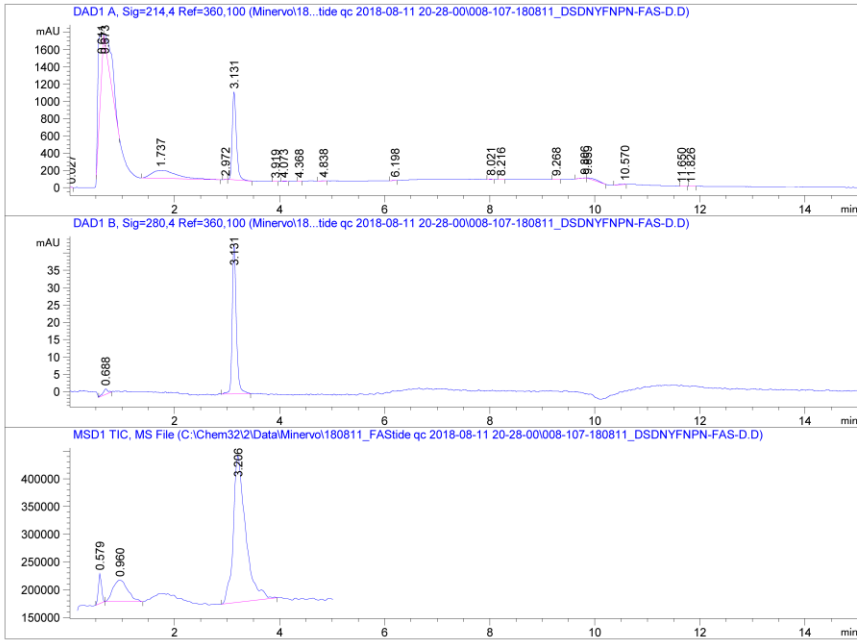
MSCalcPurity Error # 407!!!

Peak #3 at 3.090 min (2.745 to 4.123 min)
-> No purity results available. <-

*** End of Report ***

FAStide-D

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\008-107-180811_DSNYFNPN-FAS-D.D
Sample Name: 180811_DSNYFNPN-FAS-D



==========

Fraction Information

==========

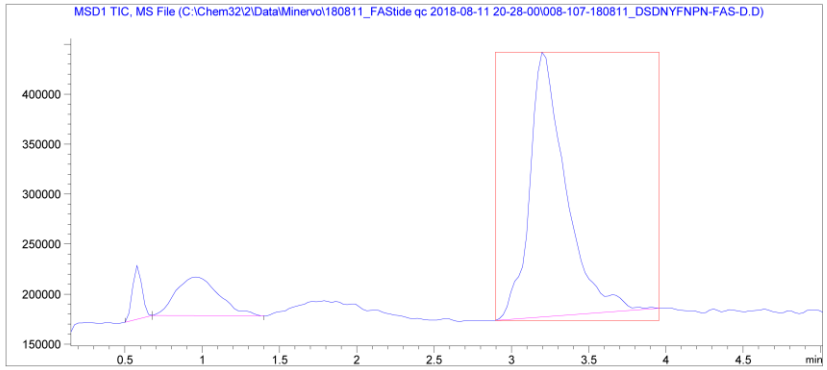
Fraction collection off

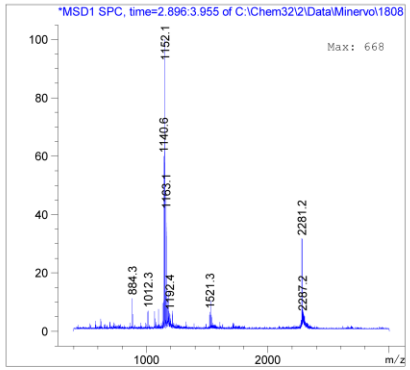
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\008-107-180811_DSNYFNPN-FAS-D.D
Sample Name: 180811_DSNYFNPN-FAS-D





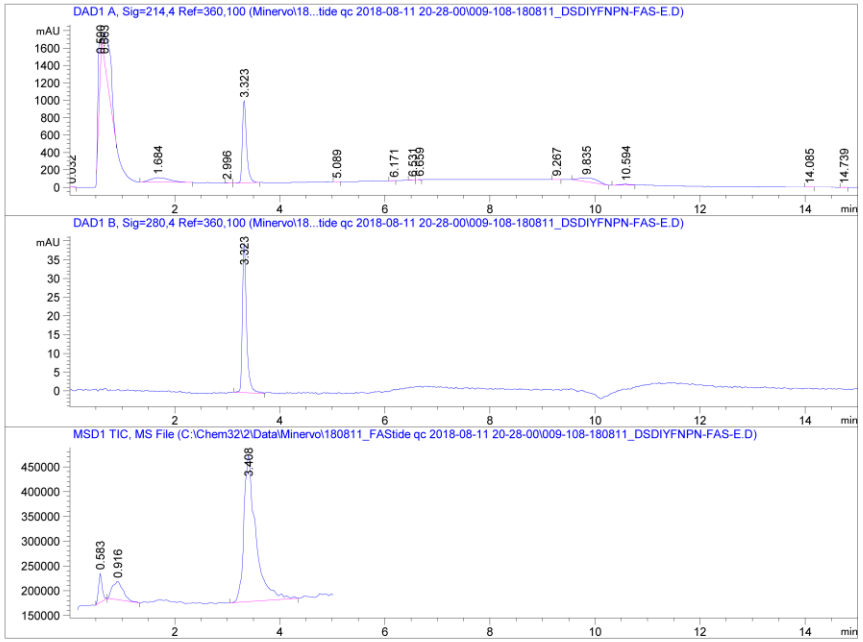
MSCalcPurity Error # 407!!!

Peak #3 at 3.206 min (2.896 to 3.956 min)
-> No purity results available. <-

*** End of Report ***

FAStide-E

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\009-108-180811_DSDIYFNPN-FAS-E.D
Sample Name: 180811_DSDIYFNPN-FAS-E



==========

Fraction Information

==========

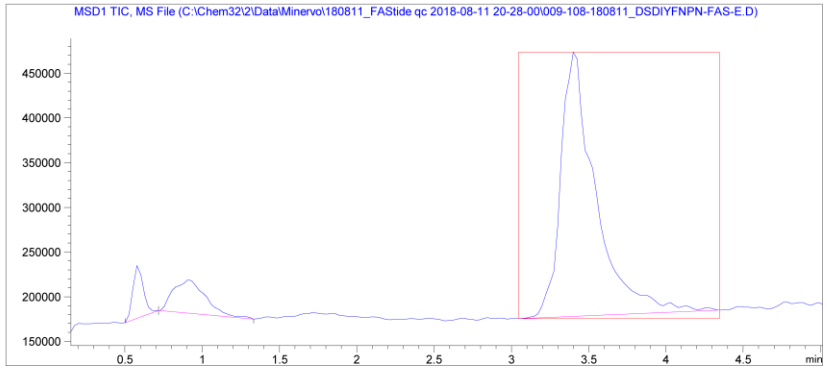
Fraction collection off

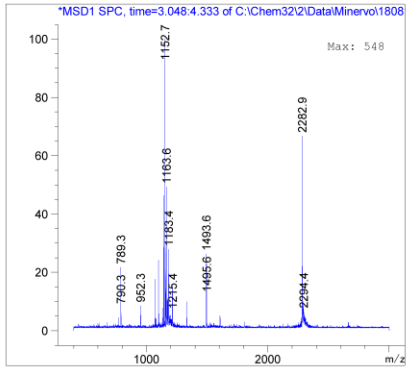
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\009-108-180811_DSDIYFNPN-FAS-E.D
Sample Name: 180811_DSDIYFNPN-FAS-E





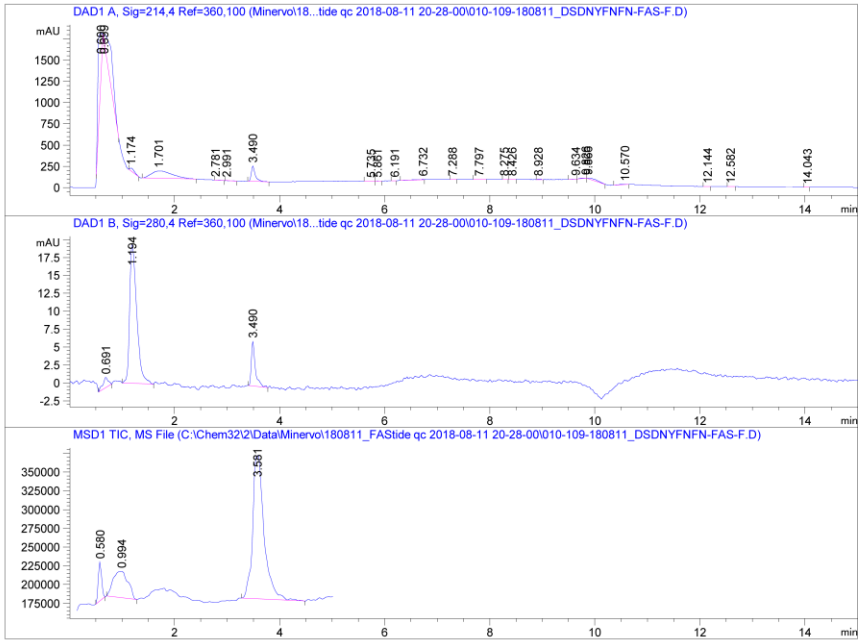
MSCalcPurity Error # 407!!!

Peak #3 at 3.408 min (3.048 to 4.345 min)
-> No purity results available. <-

*** End of Report ***

FAStide-F

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\010-109-180811_DSDNYFNFN-FAS-F.D
Sample Name: 180811_DSDNYFNFN-FAS-F



==========

Fraction Information

==========

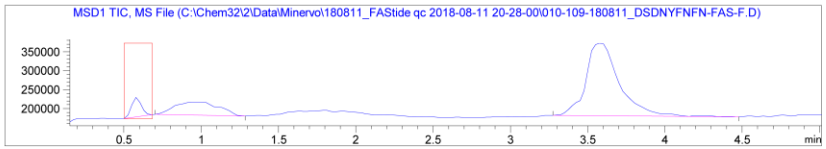
Fraction collection off

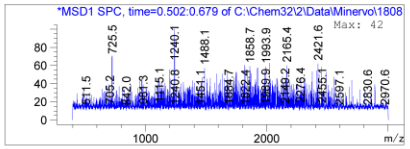
==========

No Fractions found.

==========

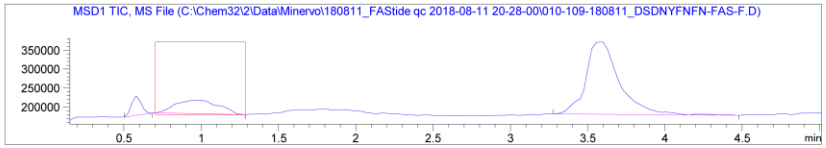
MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\010-109-180811_DSDNYFNFN-FAS-F.D
Sample Name: 180811_DSDNYFNFN-FAS-F

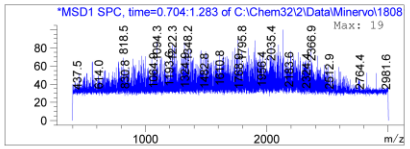




MSCalcPurity Error # 407!!!

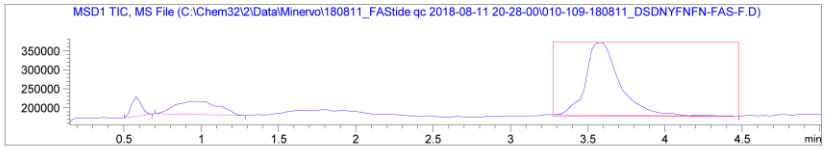
Peak #1 at 0.580 min (0.502 to 0.682 min)
-> No purity results available. <-

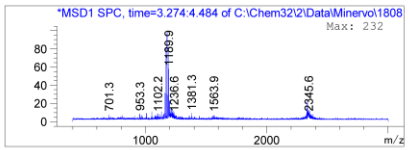




MSCalcPurity Error # 407!!!

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\010-109-180811_DSDNYFNFN-FAS-F.D
Sample Name: 180811_DSDNYFNFN-FAS-F

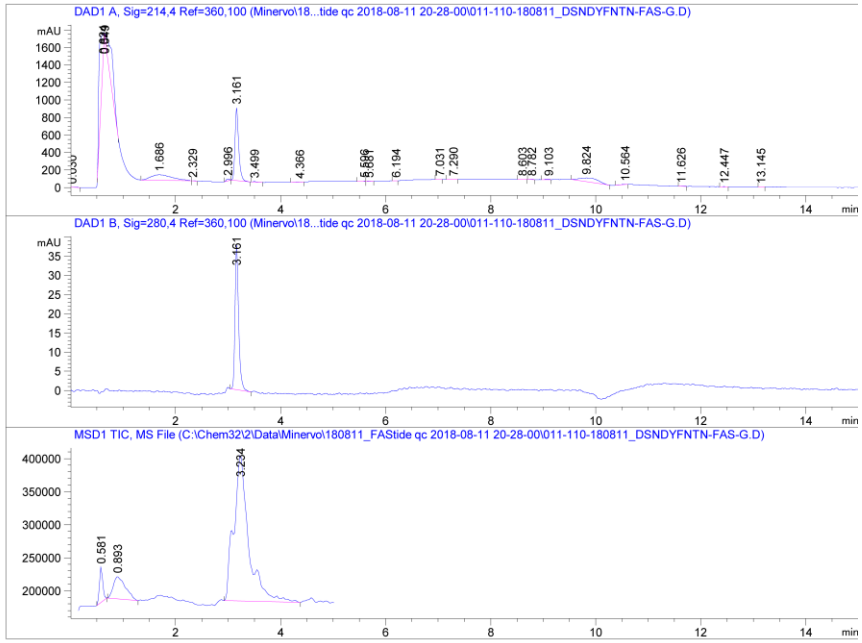




MSCalcPurity Error # 407!!!

FAStide-G

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\011-110-180811_DSNDYFNTN-FAS-G.D
Sample Name: 180811_DSNDYFNTN-FAS-G



==========

Fraction Information

==========

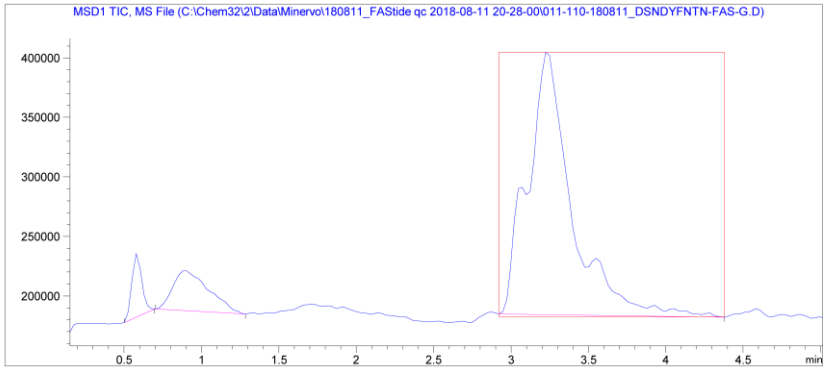
Fraction collection off

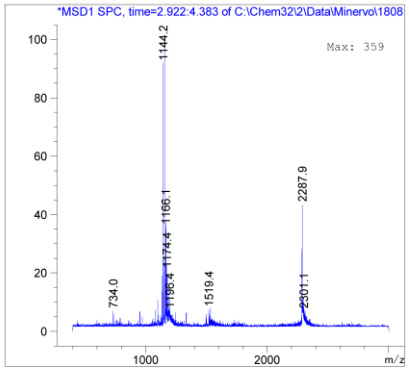
==========

No Fractions found.

==========

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\011-110-180811_DSNDYFNTN-FAS-G.D
Sample Name: 180811_DSNDYFNTN-FAS-G





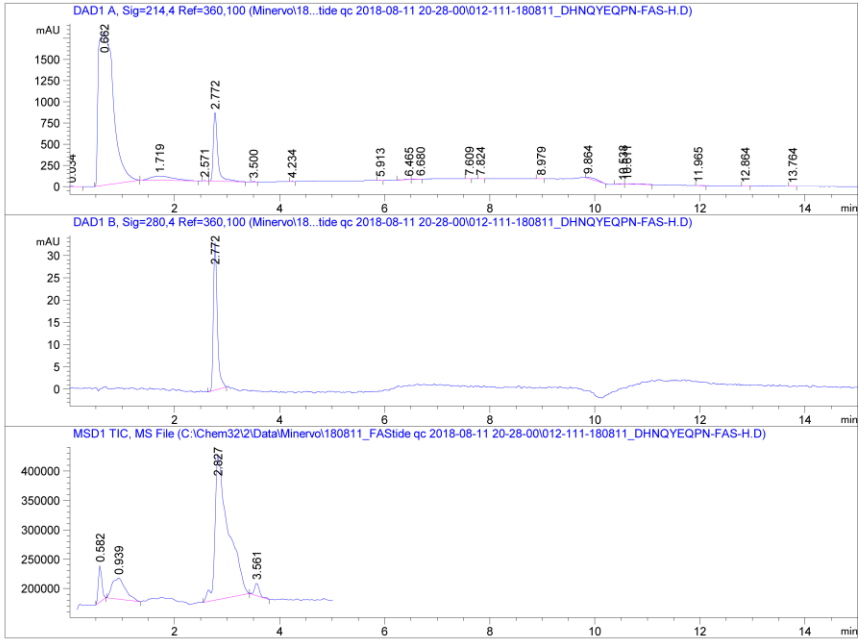
MSCalcPurity Error # 407!!!

Peak #3 at 3.234 min (2.922 to 4.377 min)
-> No purity results available. <-

*** End of Report ***

FAStide-H

MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\012-111-180811_DHNQYEQPN-FAS-H.D
Sample Name: 180811_DHNQYEQPN-FAS-H



==========

Fraction Information

==========

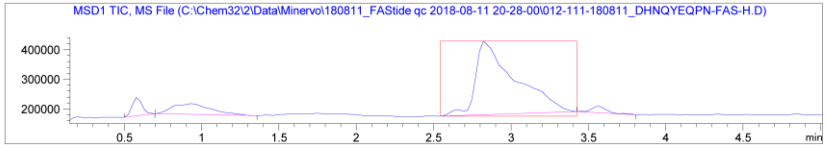
Fraction collection off

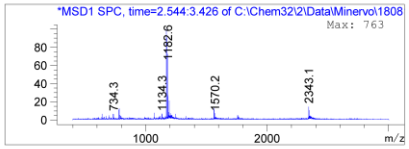
==========

No Fractions found.

==========

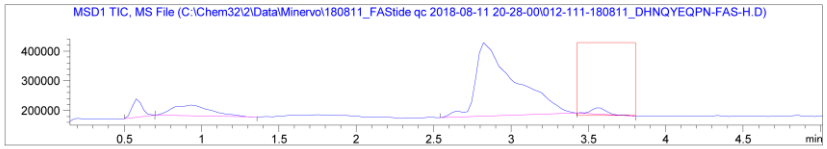
MS Peak Purity Range Report
Data File C:\Chem32\...811_FAStide qc 2018-08-11 20-28-00\012-111-180811_DHNQYEQPN-FAS-H.D
Sample Name: 180811_DHNQYEQPN-FAS-H

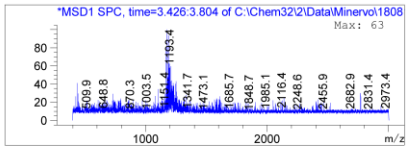




MSCalcPurity Error # 407!!!

Peak #3 at 2.827 min (2.544 to 3.426 min)
-> No purity results available. <-





MSCalcPurity Error # 407!!!

Peak #4 at 3.561 min (3.429 to 3.804 min)
-> No purity results available. <-

*** End of Report ***

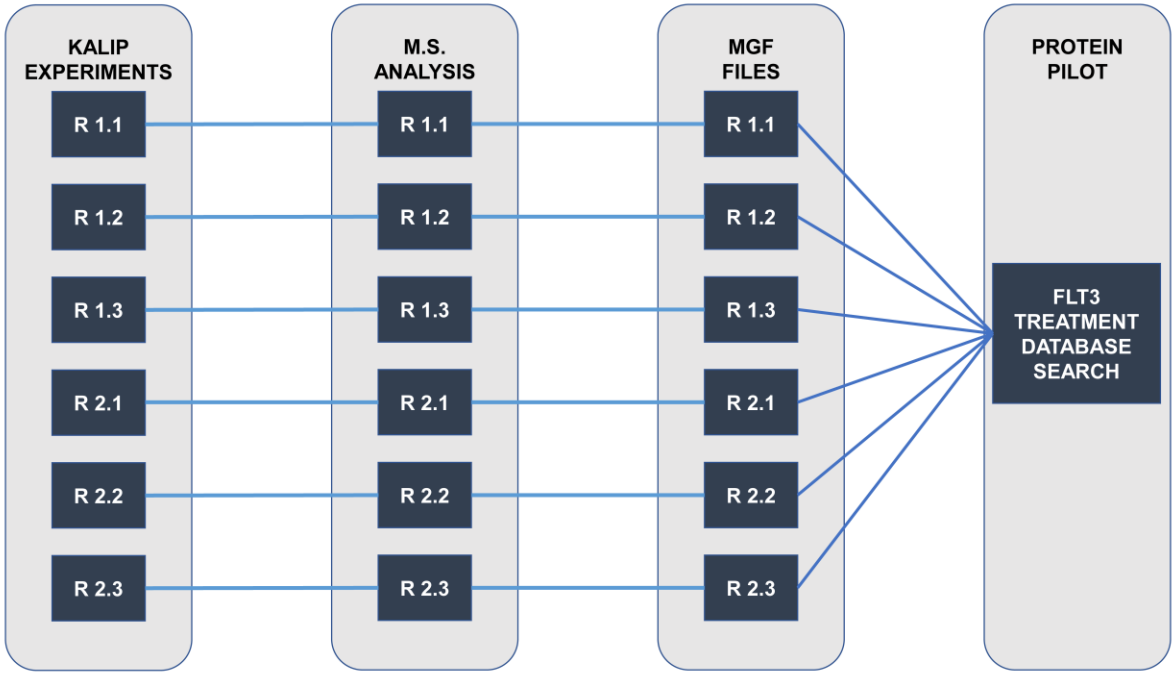


Figure S1 Schematic representation of raw mass spectrometer file combination for ProteinPilot database searches. Each KALIP kinase treatment (WT, D835Y and/or ITD) was performed with three biological replicates (R1). The KALIP process was then repeated later to generate a second independent KALIP technical experiment (R2). Replicates were individually analyzed on the mass spectrometer and then converted to MGF files, ProteinPilot 5.0 database search consisted of six mass spectrometer files for each kinase treatment (no kinase or kinase treatment).

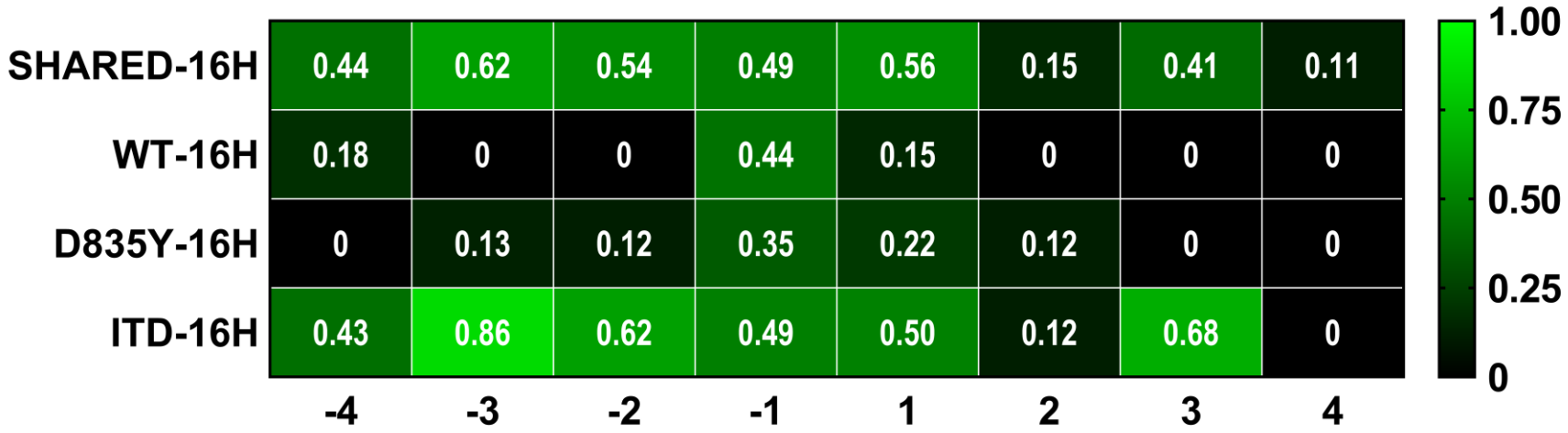


Figure S2 is a heat map representation of the Site Selectivity Matrix (SSM) values found in "Output file 2" and was generated as previously reported.[31] SSM values closer to 1 suggest that the kinase of interest would be more sensitive to changes in the particular residue at this position.

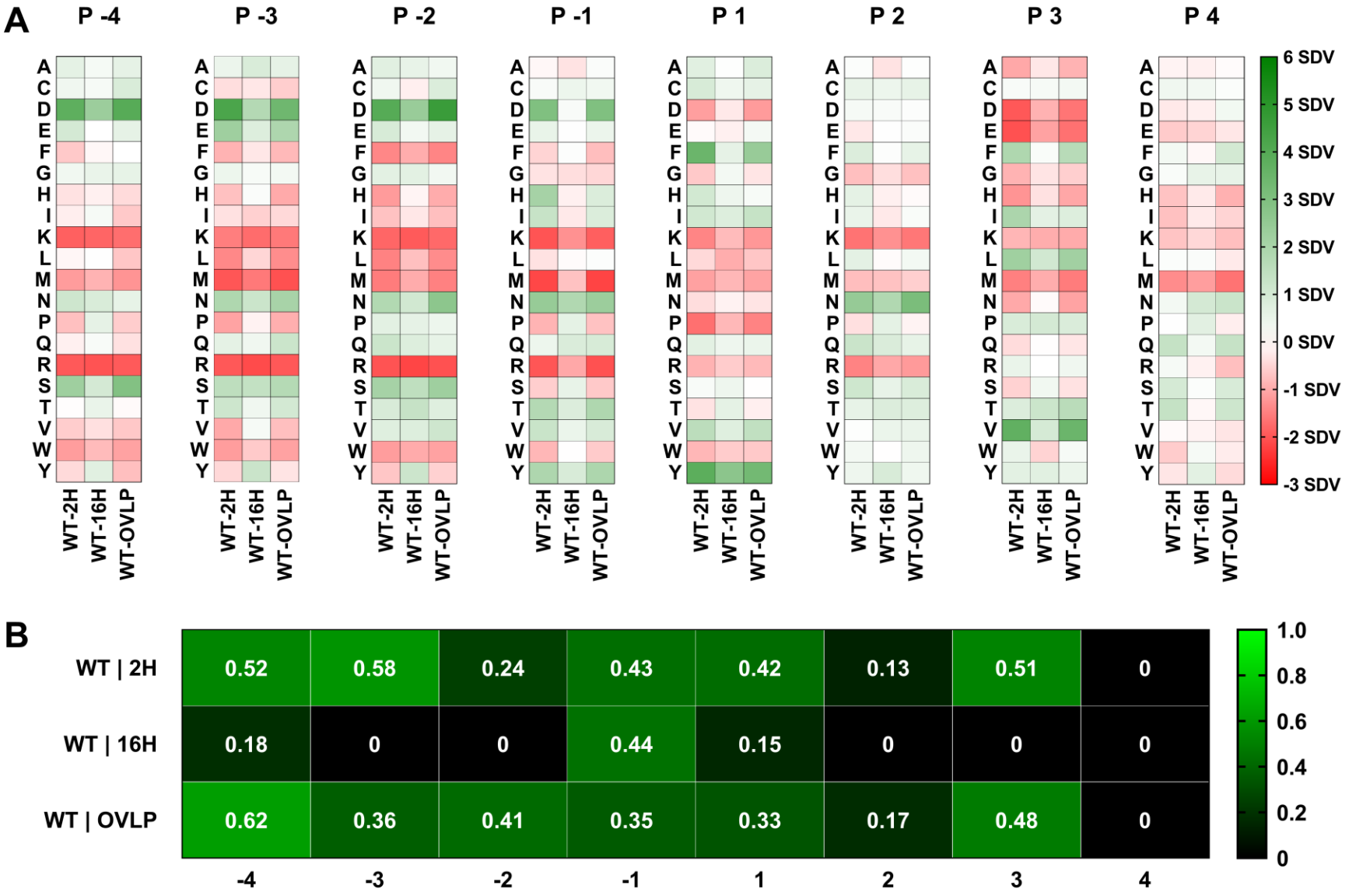

A
P -4
P -3
P -2
P -1
P 1
P 2
P 3
P 4
A C D E F G H I K L M N P Q R S T V W Y
WT-2H
WT-16H
WT-OVLP
6 SDV
5 SDV
4 SDV
3 SDV
2 SDV
1 SDV
0 SDV
-1 SDV
-2 SDV
-3 SDV
B
WT | 2H
WT | 16H
WT | OVLP
0.52 0.58 0.24 0.43 0.42 0.13 0.51 0
0.18 0 0 0.44 0.15 0 0 0
0.62 0.36 0.41 0.35 0.33 0.17 0.48 0
-4 -3 -2 -1 1 2 3 4
1.0
0.8
0.6
0.4
0.2
0

Figure S3. Heat map representation of FLT3-WT time course KALIP experiment, Site Selectivity Matrix and artificial substrate library sequence scoring comparison. (A) Observed representation of each amino acid at each position (-4 to +4 relative to phosphotyrosine) in the individual phosphoproteomics datasets for the kinase treatments at two hours (WT-2H) or sixteen hours (WT-16H), or for the sequences shared in the two datasets (WT-OVLP). Green = over-represented, white = neutral, red = under-represented. To summarize, differences were modest between the two treatment times. (B) We compared the three substrate lists' SSM values to identify positions with a value greater than 1, which is the previously reported threshold used to consider a position as "significant."[1] None of the KALIP dataset SSMs contained a position with a value greater than one, suggesting that all positions exhibited some flexibility for which particular amino acid was present.

References

1. Lipchik, A. M. *et al.* KINATEST-ID: A pipeline to develop phosphorylation-dependent terbium sensitizing kinase assays. *J. Am. Chem. Soc.* **137,** 2484–2494 (2015).
2. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **12,** e0177678 (2017).
3. Jiao, Y. & Du, P. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quant. Biol.* **4,** 320–330 (2016).
4. Narasimhan, H., Vaish, R. & Agarwal, S. On the Statistical Consistency of Plug-in Classifiers for Non-decomposable Performance Measures. *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* 1493–1501 (2014). at <http://dl.acm.org/citation.cfm?id=2968826.2968993>
5. Breu, F., Guggenbichler, S. & Wollmann, J. *Machine Learning: ECML 2004*. *15th European Conference on Machine Learning Pisa, Italy, September 2004 Proceedings* **3201,** (Springer Berlin Heidelberg, 2004).