WAVE®
LIFE SCIENCES

Blending KNIME Data ETL and Analytics Capabilities with an Enterprise LIMS, ELN and Cheminformatics Platform

Ken Longo

June 1, 2022

The Big Picture

Wave Life Sciences is a genetic medicines company focused on delivering life-changing treatments for people battling devastating diseases.

Early in the drug discovery process we use a combination of well-established enterprise software solutions for molecule registration, cheminformatics and data acquisition.

We use KNIME to interface with these platforms and perform a range of critical data ETL, modeling and analytics services.

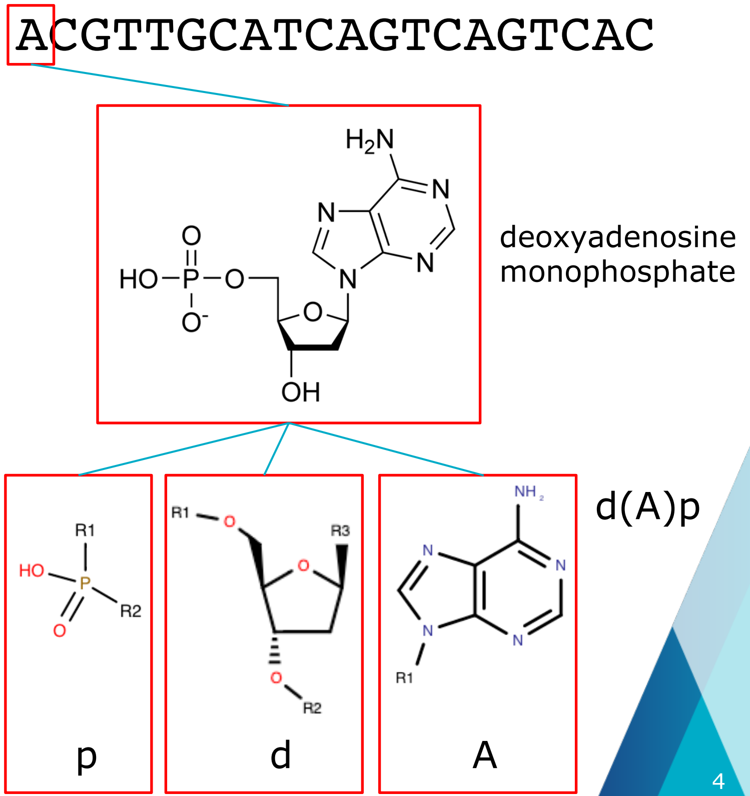
WAVE LIFE SCIENCES

A Story of Growth & Change at Wave…

- Grew from a small biotechnology start-up of <10 people into a 225+ person company
- Sponsors clinical trials worldwide in the areas of Huntington’s Disease, Frontotemporal Dementia, Amyotrophic Lateral Sclerosis and Muscular Dystrophy
- Has a pipeline of several early-stage drug discovery programs that require an informatics infrastructure
- As we evolved our systems of organizing chemical entities and the data connected to them, our use of KNIME grew and changed also

Introduction to Wave Life Sciences

- Wave Life Sciences is a **genetic medicines** company focused on delivering life-changing treatments for people battling devastating diseases.
- Requires solutions for several domains:
  - **Bioinformatics**
  - **Cheminformatics**
  - **Data processing/ETL** – assay data
  - **Analysis** (graphical, inference, prediction)
  - Clinical, manufacturing & commercial informatics
- **Challenge & opportunity:** interdisciplinary company with need for cross-domain expertise *and* platforms that integrate and connect



ACGTTGCATCAGTCAGTCAC

deoxyadenosine monophosphate

d(A)p

p

d

A

Blending KNIME with Common Drug Discovery Tools

Data ETL & Modeling

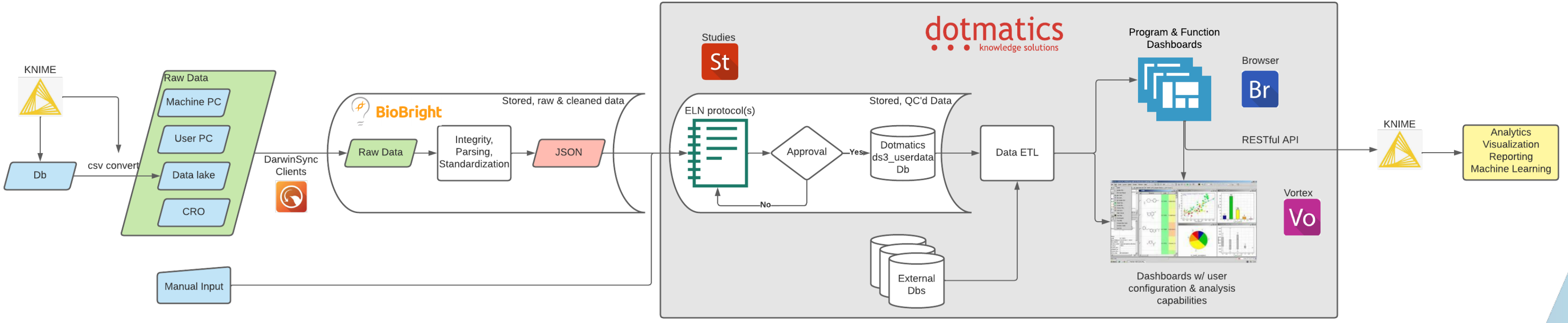
Data Lakes

LIMS

ELNs

Dashboards

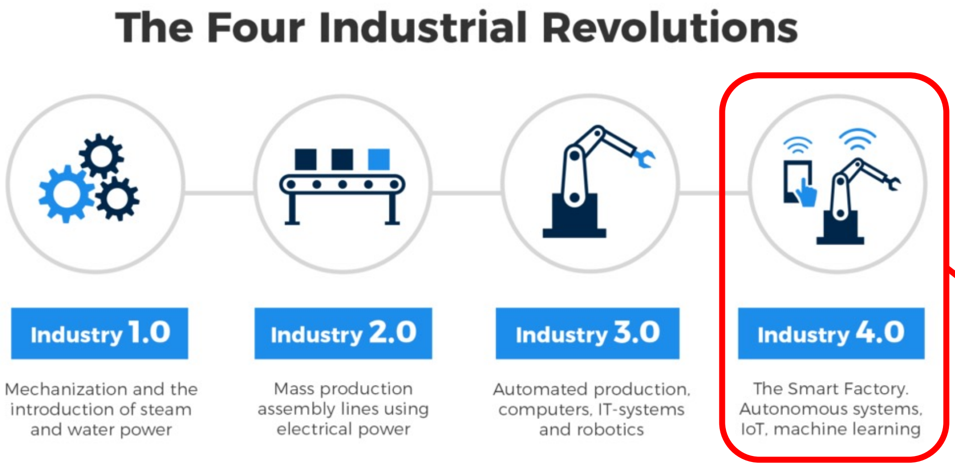
Analytics Visualization Reporting Prediction



Collectively, this developing system provides curated data and an electronic audit trail that supports IP and regulatory filings *and* quick and easy access to FAIRified data for deeper analytics.

WAVE LIFE SCIENCES

Industry 4.0 & Pharma 4.0 Framework



The Four Industrial Revolutions

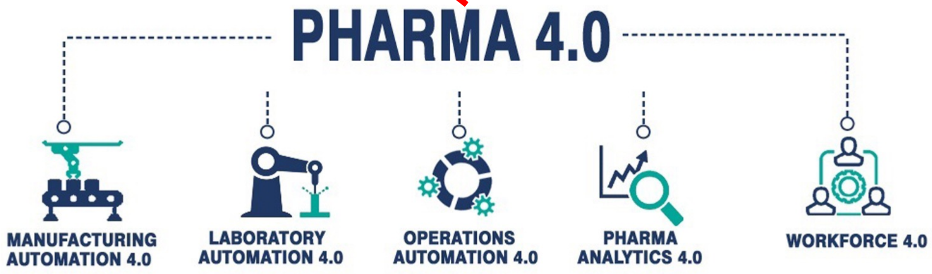
Industry 1.0
Mechanization and the introduction of steam and water power

Industry 2.0
Mass production assembly lines using electrical power

Industry 3.0
Automated production, computers, IT-systems and robotics

Industry 4.0
The Smart Factory. Autonomous systems, IoT, machine learning

Open for Innovation
KNIME



PHARMA 4.0

- MANUFACTURING AUTOMATION 4.0
- LABORATORY AUTOMATION 4.0
- OPERATIONS AUTOMATION 4.0
- PHARMA ANALYTICS 4.0
- WORKFORCE 4.0

Our KNIME Setup

KNIME Server

- AWS-hosted
- Test and Prod environments
- Cron job, microservice and WebPortal workflows

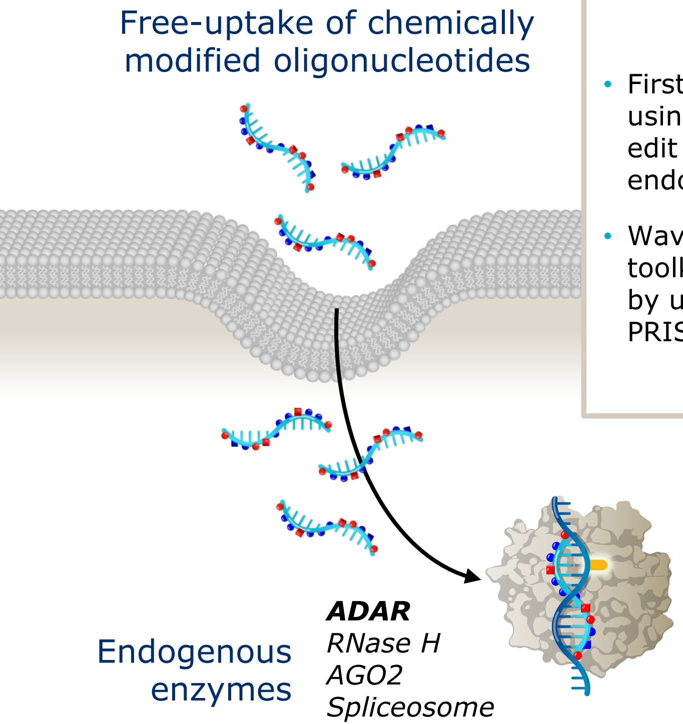
KNIME WebPortal

- Heavy internal use by scientists/drug discovery programs

Integrations

- R
- Python

Unlocking RNA editing with PRISM platform to develop AIMers: A-to-I editing oligonucleotides



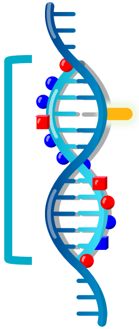
- First publication (1995) using oligonucleotide to edit RNA with endogenous ADAR[1]
- Wave goal: Expand toolkit to include editing by unlocking ADAR with PRISM oligonucleotides

✓ Learnings from biological concepts

✓ Applied to ASO structural concepts

✓ Applied PRISM chemistry

AIMer: Wave's **A-to-I** editing oligonucleotides

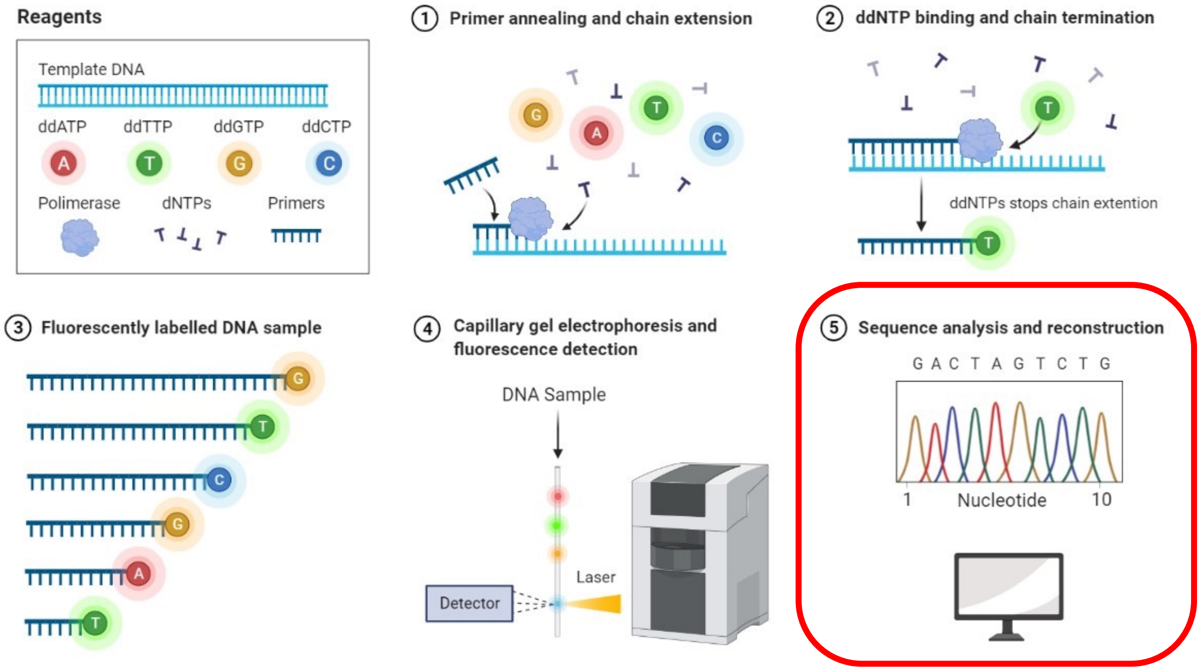


ADAR enzymes

- Catalyze conversion of A-to-I (G) in double-stranded RNA substrates
- A-to-I (G) edits are one of the most common post-transcriptional modifications
- ADAR1 is ubiquitously expressed across tissues, including liver and CNS

[1] Woolf et al., PNAS Vol. 92, pp. 8298-8302, 1995

Base Editing Pipeline Analyzes Sanger Sequencing Data

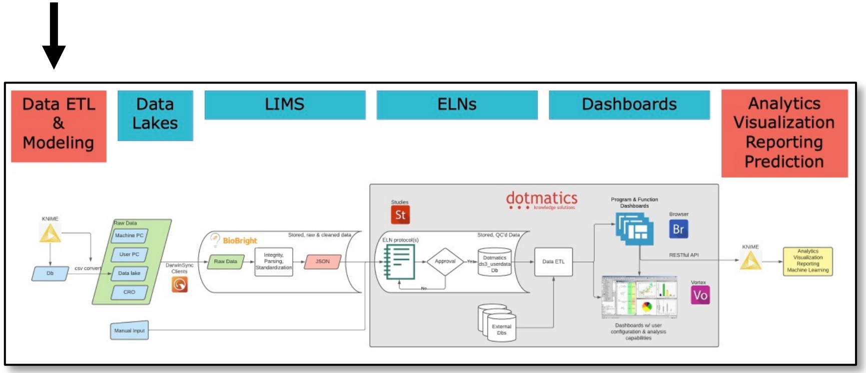

Reagents
Template DNA
ddATP
ddTTP
ddGTP
ddCTP
A
T
G
C
Polimerase
dNTPs
Primers
1 Primer annealing and chain extension
2 ddNTP binding and chain termination
ddNTPs stops chain extention
3 Fluorescently labelled DNA sample
4 Capillary gel electrophoresis and fluorescence detection
DNA Sample
Laser
Detector
5 Sequence analysis and reconstruction
G A C T A G T C T G
1
Nucleotide
10

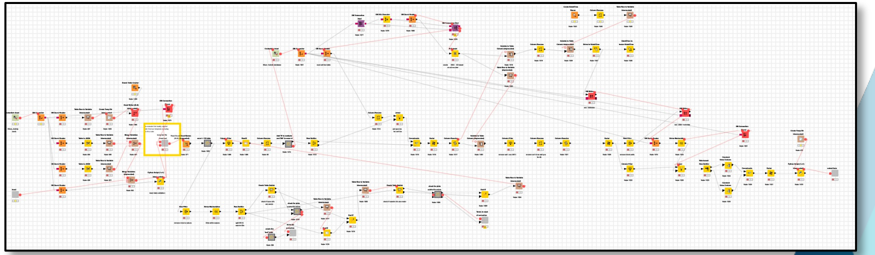
- Sanger sequencing technology has been around in various forms since 1977
- Modern applications produce a sequence chromatogram “ab1” file
- The colors of the chromatogram are translated into the “AGCT” letter code
- We infer the %base editing from the AUCs of the traces

Sanger Base-Editing Pipeline

The **Sanger pipeline** for **RNA A->I editing measurement** is comprised of three KNIME workflows:

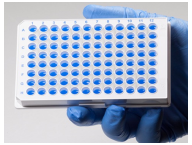
- **Metadata Input and Submission File Generation**
  - 96-well plate format input file to store metadata
  - Exportation of vendor Sample Submission Form
- **Automated Data QC and Processing**
  - CRON job
  - QC of Ab1 files
  - Base representation at all transcript locations along AIMer calculated via editR
- **Data Retrieval**
  - Export edit site percent editing results
  - Download chromatograms

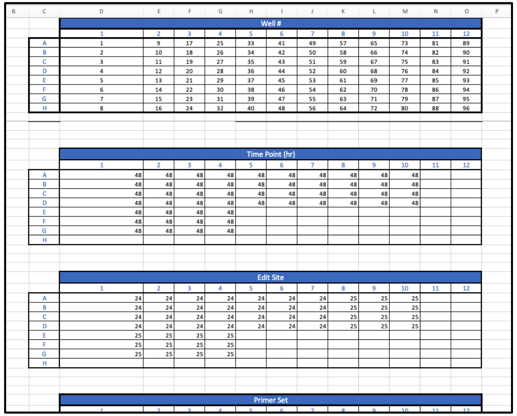




Metadata Registration and Submission File Generation (1)

User input





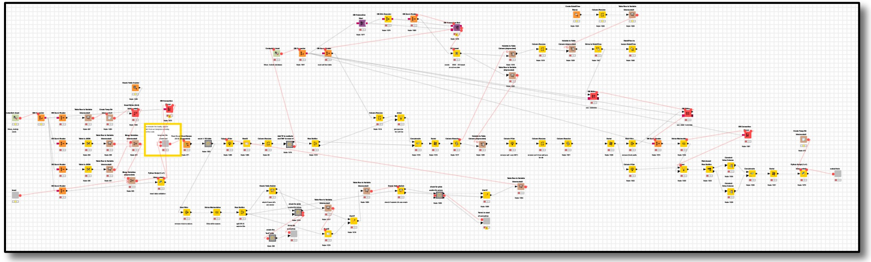
	Well #											
	1	2	3	4	5	6	7	8	9	10	11	12
A	1	9	17	25	33	41	49	57	65	73	81	89
B	2	10	18	26	34	42	50	58	66	74	82	90
C	3	11	19	27	35	43	51	59	67	75	83	91
D	4	12	20	28	36	44	52	60	68	76	84	92
E	5	13	21	29	37	45	53	61	69	77	85	93
F	6	14	22	30	38	46	54	62	70	78	86	94
G	7	15	23	31	39	47	55	63	71	79	87	95
H	8	16	24	32	40	48	56	64	72	80	88	96

	Time Point (hr)											
	1	2	3	4	5	6	7	8	9	10	11	12
A	48	48	48	48	48	48	48	48	48	48		
B	48	48	48	48	48	48	48	48	48	48		
C	48	48	48	48	48	48	48	48	48	48		
D	48	48	48	48	48	48	48	48	48	48		
E	48	48	48	48								
F	48	48	48	48								
G	48	48	48	48								
H												

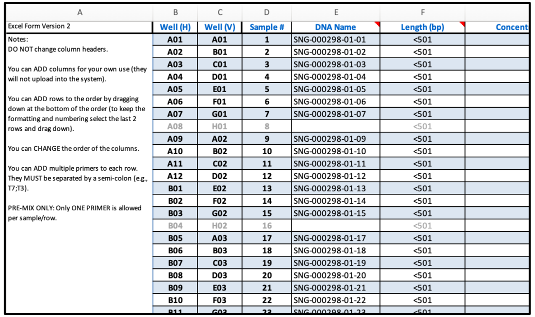
	Edit Site											
	1	2	3	4	5	6	7	8	9	10	11	12
A	24	24	24	24	24	24	24	25	25	25		
B	24	24	24	24	24	24	24	25	25	25		
C	24	24	24	24	24	24	24	25	25	25		
D	24	24	24	24	24	24	24	25	25	25		
E	25	25	25	25								
F	25	25	25	25								
G	25	25	25	25								
H												

	Primer Set											
	1	2	3	4	5	6	7	8	9	10	11	12

Standardized experimental metadata with built-in validation; stored in Db



Vendor submission



Excel Form Version 2

Notes:
DO NOT change column headers.

You can ADD columns for your own use (they will not upload into the system).

You can ADD rows to the order by dragging down at the bottom of the order (to keep the formatting and numbering select the last 2 rows and drag down).

You can CHANGE the order of the columns.

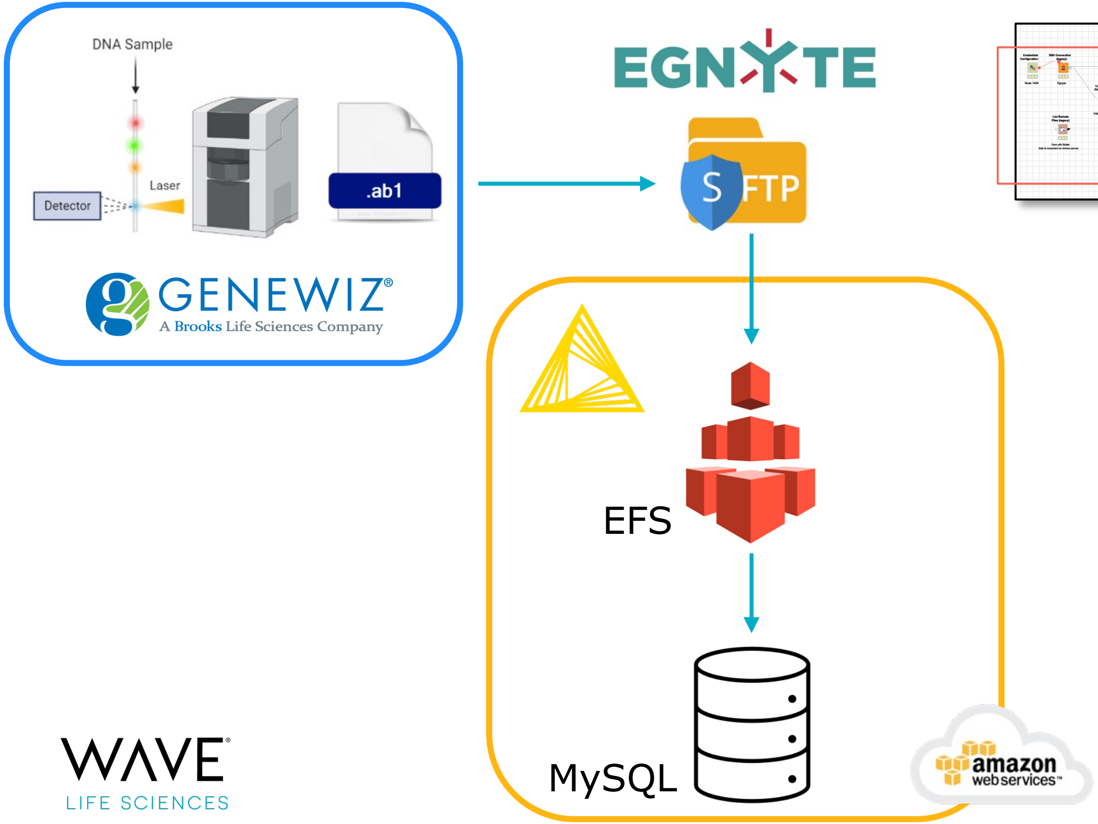
You can ADD multiple primers to each row. They MUST be separated by a semi-colon (e.g., T7;T3).

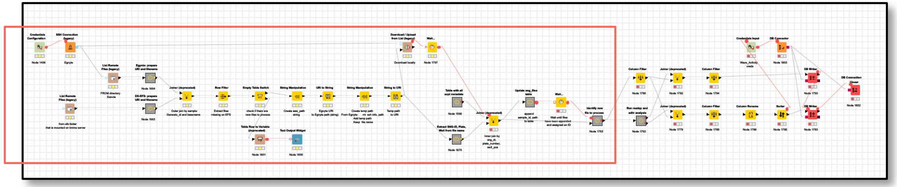
PRE-MIX ONLY: Only ONE PRIMER is allowed per sample/row.

Well (H)	Well (V)	Sample #	DNA Name	Length (bp)	Concent
A01	A01	1	SNG-000298-01-01	<501	
A02	B01	2	SNG-000298-01-02	<501	
A03	C01	3	SNG-000298-01-03	<501	
A04	D01	4	SNG-000298-01-04	<501	
A05	E01	5	SNG-000298-01-05	<501	
A06	F01	6	SNG-000298-01-06	<501	
A07	G01	7	SNG-000298-01-07	<501	
A08	H01	8		<501	
A09	A02	9	SNG-000298-01-09	<501	
A10	B02	10	SNG-000298-01-10	<501	
A11	C02	11	SNG-000298-01-11	<501	
A12	D02	12	SNG-000298-01-12	<501	
B01	E02	13	SNG-000298-01-13	<501	
B02	F02	14	SNG-000298-01-14	<501	
B03	G02	15	SNG-000298-01-15	<501	
B04	H02	16		<501	
B05	A03	17	SNG-000298-01-17	<501	
B06	B03	18	SNG-000298-01-18	<501	
B07	C03	19	SNG-000298-01-19	<501	
B08	D03	20	SNG-000298-01-20	<501	
B09	E03	21	SNG-000298-01-21	<501	
B10	F03	22	SNG-000298-01-22	<501	
B11	G03	23	SNG-000298-01-23	<501	

Automated generation of downloadable vendor-preferred formatted sequencing submission file

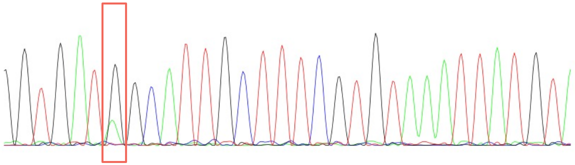
Automated ab1 file transfer to Server-mounted EFS (2)



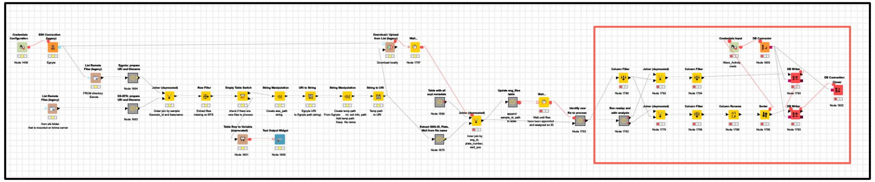


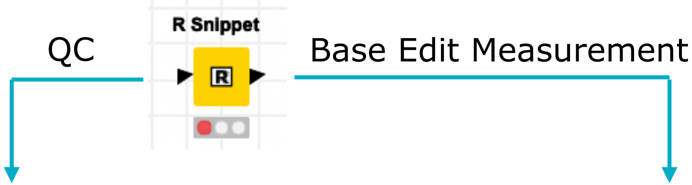
- ab1 files deposited by vendor to Egnyte repository
- CRON job runs on KNIME Server to detect presence of new files in repo
  - Transfer to mounted EFS
  - Store file metadata in database

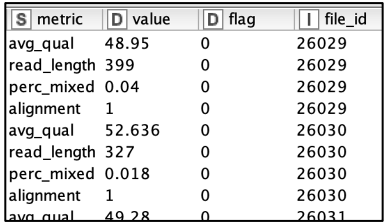
Automated Sanger Data QC and Processing (3)



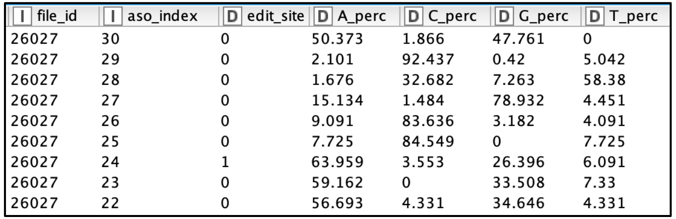
EFS







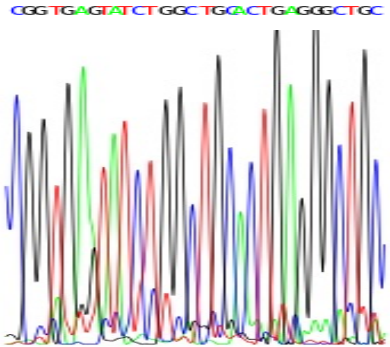
S metric	D value	D flag	I file_id
avg_qual	48.95	0	26029
read_length	399	0	26029
perc_mixed	0.04	0	26029
alignment	1	0	26029
avg_qual	52.636	0	26030
read_length	327	0	26030
perc_mixed	0.018	0	26030
alignment	1	0	26030
avg_qual	49.28	0	26031

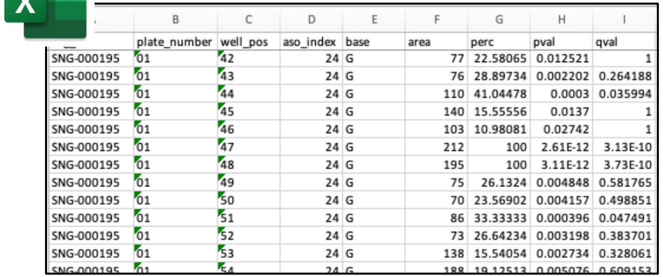


I file_id	I aso_index	D edit_site	D A_perc	D C_perc	D G_perc	D T_perc
26027	30	0	50.373	1.866	47.761	0
26027	29	0	2.101	92.437	0.42	5.042
26027	28	0	1.676	32.682	7.263	58.38
26027	27	0	15.134	1.484	78.932	4.451
26027	26	0	9.091	83.636	3.182	4.091
26027	25	0	7.725	84.549	0	7.725
26027	24	1	63.959	3.553	26.396	6.091
26027	23	0	59.162	0	33.508	7.33
26027	22	0	56.693	4.331	34.646	4.331

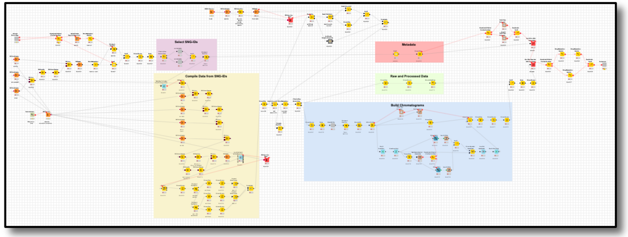
- ab1 files QC'ed based on
  - Quality metrics
  - Compound alignment to sequence
- %base editing calculated across entire compound

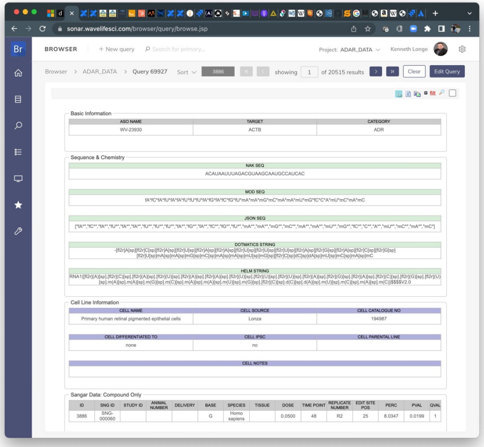
Data Retrieval (4)





	B	C	D	E	F	G	H	I
	plate_number	well_pos	aso_index	base	area	perc	pval	qval
SNG-000195	01	42	24	G	77	22.58065	0.012521	1
SNG-000195	01	43	24	G	76	28.89734	0.002202	0.264188
SNG-000195	01	44	24	G	110	41.04478	0.0003	0.035994
SNG-000195	01	45	24	G	140	15.55556	0.0137	1
SNG-000195	01	46	24	G	103	10.98081	0.02742	1
SNG-000195	01	47	24	G	212	100	2.61E-12	3.13E-10
SNG-000195	01	48	24	G	195	100	3.11E-12	3.73E-10
SNG-000195	01	49	24	G	75	26.1324	0.004848	0.581765
SNG-000195	01	50	24	G	70	23.56902	0.004157	0.498851
SNG-000195	01	51	24	G	86	33.33333	0.000396	0.047491
SNG-000195	01	52	24	G	73	26.64234	0.003198	0.383701
SNG-000195	01	53	24	G	138	15.54054	0.002734	0.328061
SNG-000195	01	54	24	G	188	19.12513	0.005076	0.609153




sonar.wavelifesci.com/browser/query/browse.jsp
Br
BROWSER
+ New query
Search for primary...
Project: ADAR_DATA
Kenneth Longo
Browser > ADAR_DATA > Query 69927
Sort
3886
showing 1 of 20515 results
Clear
Edit Query
filt

Basic Information

ASO NAME	TARGET	CATEGORY
WV-23930	ACTB	ADR

Sequence & Chemistry

NAK SEQ

ACAUAAUUUAGACGUAAGCAAUGCCAUCAC

MOD SEQ

fA*fC*fA*fU*fA*fA*fU*fU*fU*fA*fG*fA*fC*fG*fU*mA*mA*mG*mC*mA*mA*mU*mG*fC*C*A*mU*mC*mA*mC

JSON SEQ

["fA*","fC*","fA*","fU*","fA*","fA*","fU*","fU*","fU*","fA*","fG*","fA*","fC*","fG*","fU*","mA*","mA*","mG*","mC*","mA*","mA*","mU*","mG*","fC*","C*","A*","mU*","mC*","mA*","mC"]

DOTMATICS STRING

-[fl2r]A[sp][fl2r]C[sp][fl2r]A[sp][fl2r]U[sp][fl2r]A[sp][fl2r]A[sp][fl2r]U[sp][fl2r]U[sp][fl2r]U[sp][fl2r]A[sp][fl2r]G[sp][fl2r]A[sp][fl2r]C[sp][fl2r]G[sp][fl2r]U[sp]mA[sp]mA[sp]mG[sp]mC[sp]mA[sp]mA[sp]mU[sp]mG[sp][fl2r]C[sp]dC[sp]dA[sp]mU[sp]mC[sp]mA[sp]mC

HELM STRING

RNA1{[fl2r](A)[sp].[fl2r](C)[sp].[fl2r](A)[sp].[fl2r](U)[sp].[fl2r](A)[sp].[fl2r](A)[sp].[fl2r](U)[sp].[fl2r](U)[sp].[fl2r](U)[sp].[fl2r](A)[sp].[fl2r](G)[sp].[fl2r](A)[sp].[fl2r](C)[sp].[fl2r](G)[sp].[fl2r](U)[sp].m(A)[sp].m(A)[sp].m(G)[sp].m(C)[sp].m(A)[sp].m(A)[sp].m(U)[sp].m(G)[sp].[fl2r](C)[sp].d(C)[sp].d(A)[sp].m(U)[sp].m(C)[sp].m(A)[sp].m(C)}$$$$V2.0

Cell Line Information

CELL NAME	CELL SOURCE	CELL CATALOGUE NO
Primary human retinal pigmented epithelial cells	Lonza	194987

CELL DIFFERENTIATED TO	CELL IPSC	CELL PARENTAL LINE
none	no	

CELL NOTES

Sangar Data: Compound Only

ID	SNG ID	STUDY ID	ANIMAL NUMBER	DELIVERY	BASE	SPECIES	TISSUE	DOSE	TIME POINT	REPLICATE NUMBER	EDIT SITE POS	PERC	PVAL	QVAL
3886	SNG-000060				G	Homo sapiens		0.0500	48	R2	25	8.0347	0.0199	1

- **WebPortal** downloadable Excel files containing experimental metadata, base editing statistics and chromatogram plots
- **Dotmatics** Browser-based dashboard for program teams

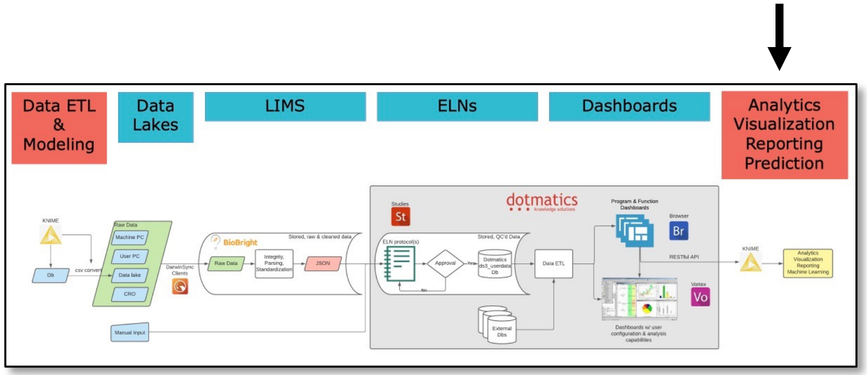
WAVE
LIFE SCIENCES

Measuring (& Predicting) Tolerability in Mice

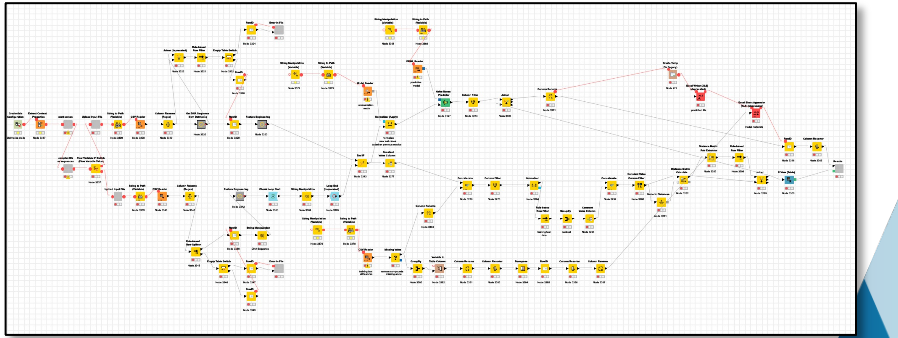
We **monitor behavior** of mice in response to chemical compounds repeatedly over several weeks:

- Activity
- Motor changes
- Breathing
- Physical characteristics
- Body weight

We use these data to build **predictive models** that help to inform our medicinal chemistry design and make molecules safer.

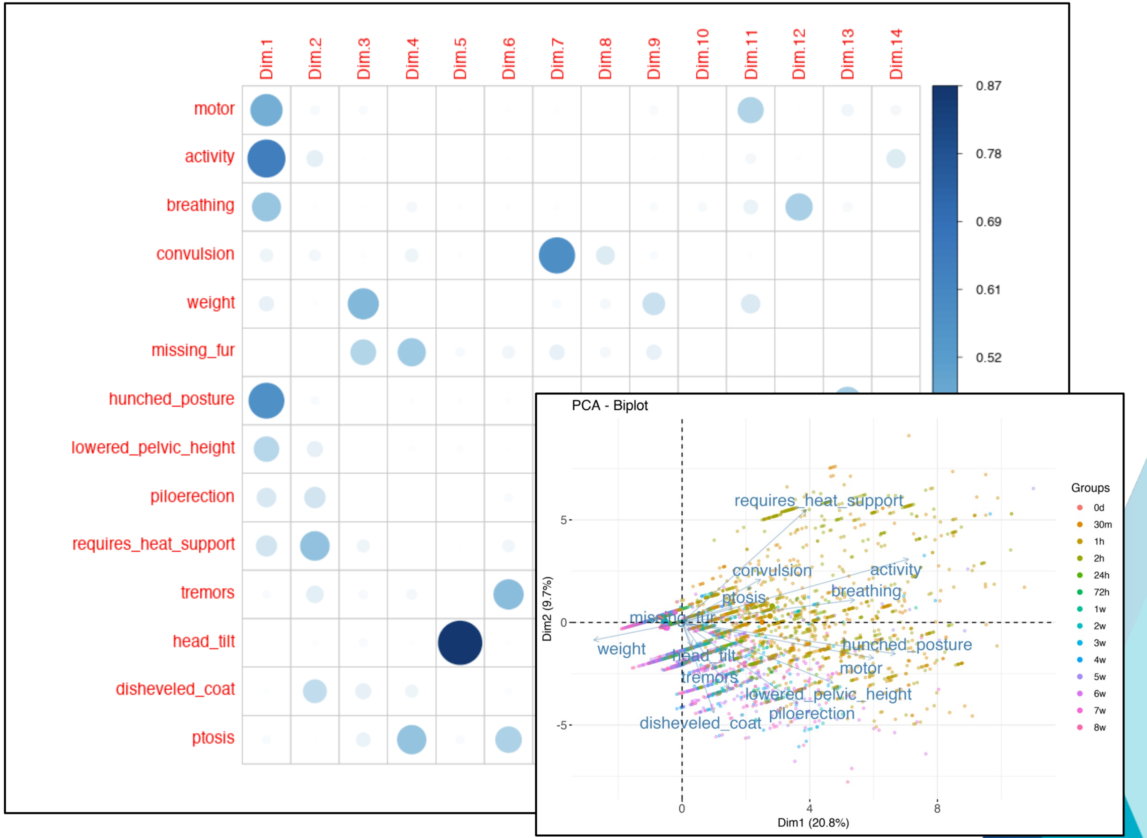


Data ETL & Modeling
Data Lakes
LIMS
ELNs
Dashboards
Analytics Visualization Reporting Prediction
KNIME
Db
csv convert
Raw Data
Machine PC
User PC
Data lake
CRO
DarwinSync Clients
Manual Input
BioBright
Stored, raw & cleaned data
Raw Data
Integrity, Parsing, Standardization
JSON
dotmatics knowledge solutions
Studies
St
ELN protocol(s)
Approval
Yes
No
Stored, QC'd Data
Dotmatics ds3_userdata Db
External Dbs
Data ETL
Program & Function Dashboards
Browser
Br
RESTful API
Vortex
Vo
Dashboards w/ user configuration & analysis capabilities
KNIME
Analytics Visualization Reporting Machine Learning



Behavioral Features Using Principal Components

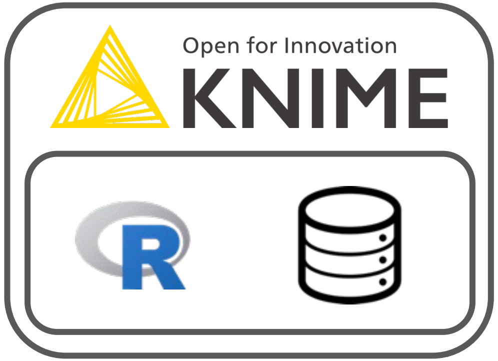
- We used **PCA** to analyze data from mice **treated with molecules**
- PCA reduces a large number of complex variables and responses (behavior, body weight, etc) to a **smaller number of patterns**
- We ask **how different treatments influence these patterns**, and their relationships to chemical features of molecules



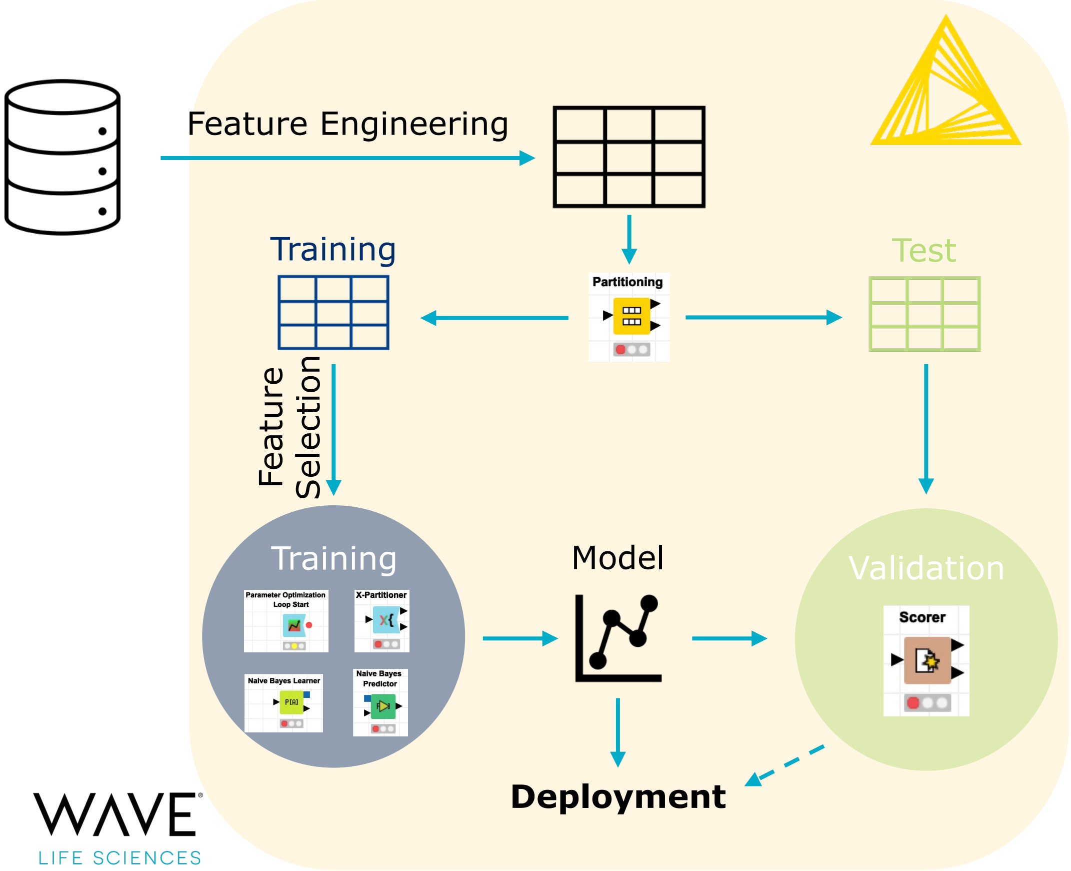
KNIME Server Webportal GUI for *in vivo* Wild-type Tolerability Prediction

The **WT Tolerability Predictive pipeline** uses KNIME for:

- **Predictive Model Building and Validation**
  - Feature Engineering and Selection
  - Model training and validation
- **User-operated WebPortal tool for predictions**
  - Fast calculation of predictions for new molecules
  - Distance measurement between new cases and existing model dataset



Predictive Model Building and Validation



Feature Engineering

Training

Partitioning

Test

Feature Selection

Training

Parameter Optimization Loop Start

X-Partitioner

Naive Bayes Learner

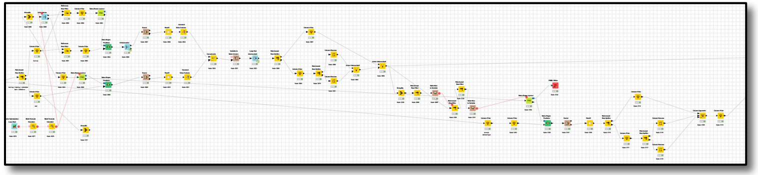
Naive Bayes Predictor

Model

Validation

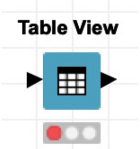
Scorer

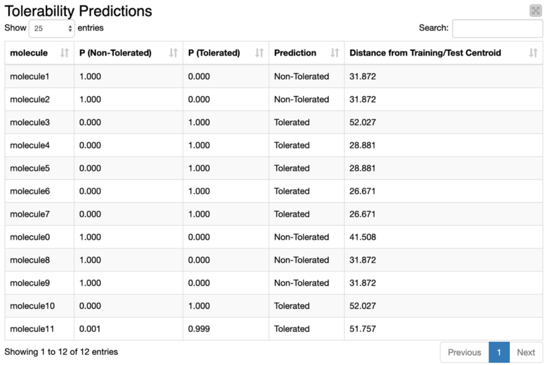
Deployment



- KNIME utilized for data cleaning, feature engineering/selection, and model building and evaluation
- Naïve Bayes Model built and validated for ***in vivo* tolerability prediction**
- 83% accuracy

Web Tool for Tolerability Prediction of New Compounds



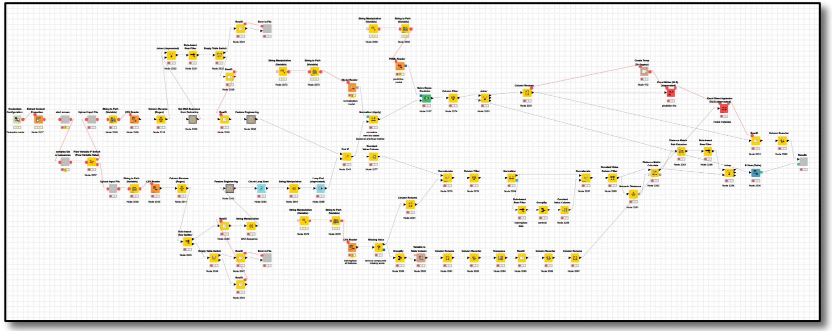


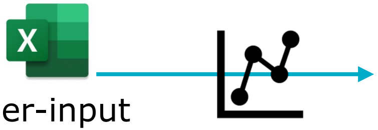
Tolerability Predictions

Show 25 entries Search:

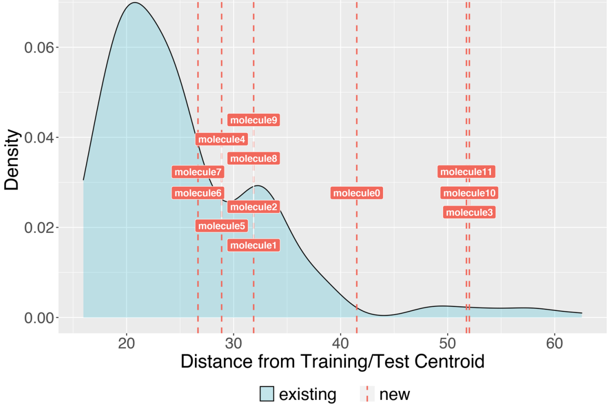
molecule	P (Non-Tolerated)	P (Tolerated)	Prediction	Distance from Training/Test Centroid
molecule1	1.000	0.000	Non-Tolerated	31.872
molecule2	1.000	0.000	Non-Tolerated	31.872
molecule3	0.000	1.000	Tolerated	52.027
molecule4	0.000	1.000	Tolerated	28.881
molecule5	0.000	1.000	Tolerated	28.881
molecule6	0.000	1.000	Tolerated	26.671
molecule7	0.000	1.000	Tolerated	26.671
molecule0	1.000	0.000	Non-Tolerated	41.508
molecule8	1.000	0.000	Non-Tolerated	31.872
molecule9	1.000	0.000	Non-Tolerated	31.872
molecule10	0.000	1.000	Tolerated	52.027
molecule11	0.001	0.999	Tolerated	51.757

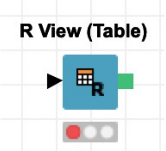
Showing 1 to 12 of 12 entries





User-input Molecule List





- User submits list of novel molecules
- Model returns tolerability score predictions
- Evaluate distance of predictions from historical tolerated molecule projections

WAVE LIFE SCIENCES

Summary & Areas for Further Development

- Internal use of component building & sharing
- Expand our microservices framework and RESTful access to data
- Continue to refine our continuous integration/deployment framework
- Expand the reach of KNIME to non-data scientists through internal training programs
- Continue to develop along an Industry/Pharma 4.0 trajectory

Thank you!

WAVE®
LIFE SCIENCES