

scalable pythonic fitting

Jonas Eschle on behalf of zfit

jonas.eschle@cern.ch





A brief history



- A few years ago: analyses transition from C++ to Python
 - Scikit-HEP was created
 - Change of philosophy: non-monolithic packages
- Fitting packages still in in C++
 - Many scattered, specialized packages
 - Speed crucial aspect (and non-trivial in python)

Fitting in Python

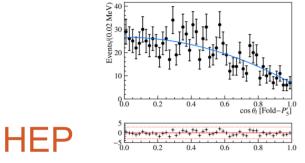


A lot of projects are around

- RooFit
- HEP Python
- Non-HEP

No real model fitting ecosystem/library for HEP that is well integrated into Python

HEP Model Fitting in Python



advanced features, simply extendable

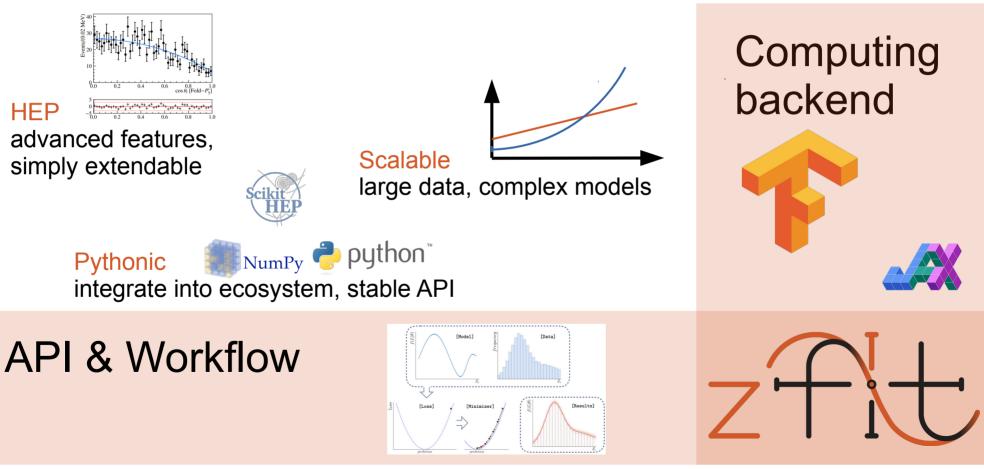


Scalable

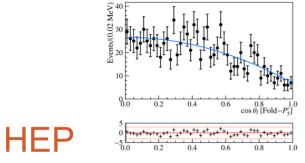
large data, complex models



HEP Model Fitting in Python



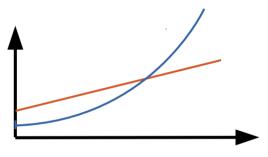
HEP Model Fitting in Python



advanced features, simply extendable



Scalable

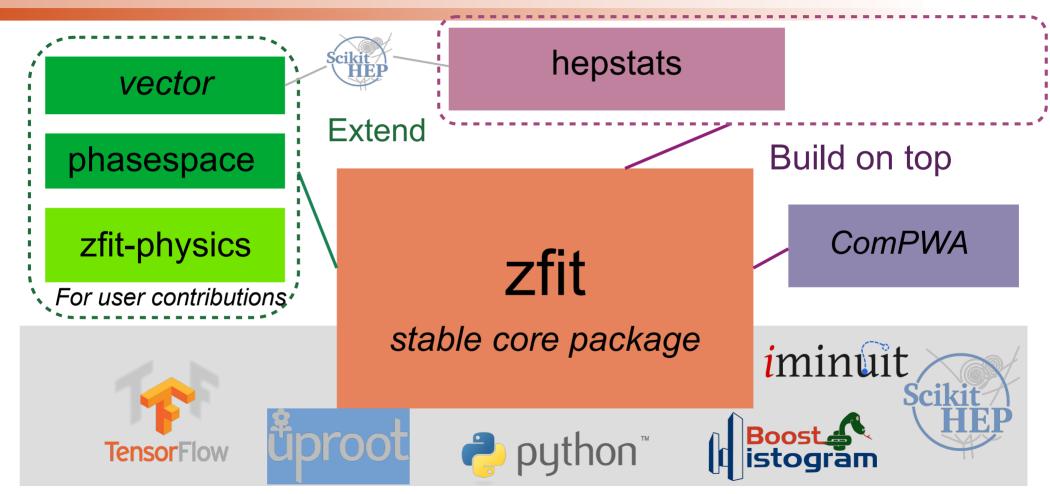


large data, complex models

Pythonic NumPy ♀ python[™] integrate into ecosystem, stable API

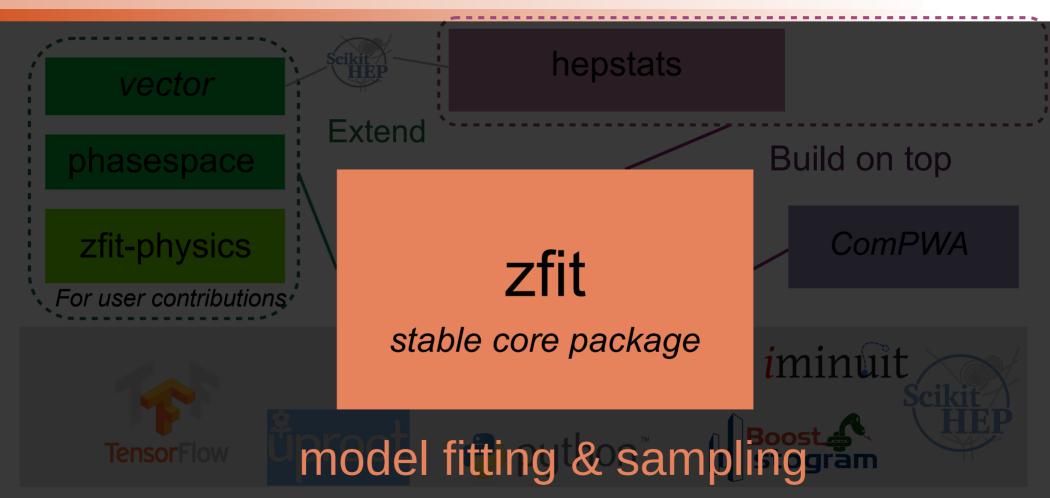






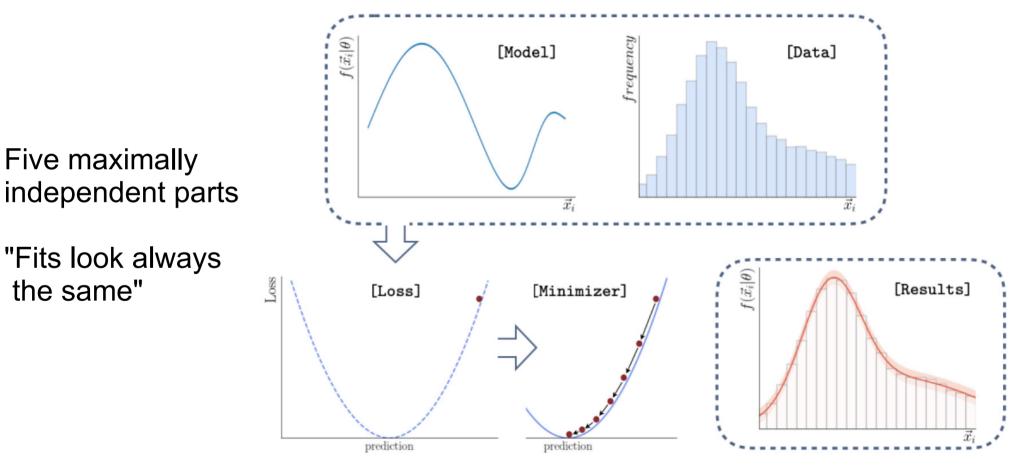






API & Workflow





2022

PyHEP 2022 - zfit: extending to binned fits

Complete fit

Disclaimer: unbinned fits way more developed, binned very new in pre-release normal_np = np.random.normal(2., 3., size=10_000)

```
obs = zfit.Space("x", limits=(-2, 3))
```

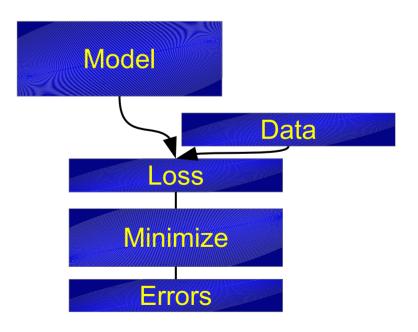
```
mu = zfit.Parameter("mu", 1.2, -4, 6)
sigma = zfit.Parameter("sigma", 1.3, 0.5, 10)
gauss = zfit.pdf.Gauss(mu=mu, sigma=sigma, obs=obs)
```

data = zfit.Data.from_numpy(obs=obs, array=normal_np)

nll = zfit.loss.UnbinnedNLL(model=gauss, data=data)

```
minimizer = zfit.minimize.Minuit()
result = minimizer.minimize(nll)
```

```
param_errors = result.hesse()
param_errors_asymmetric, new_result = result.errors()
```





11

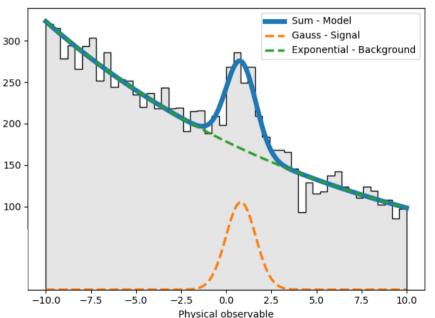
PyHEP 2022 - zfit: extending to binned fits

Example: Mass fit

- Sum, Product, (Convolution)
- Gauss, (double) Crystalball,...
- Exponential, Polynomials,...
- Histograms, SplineInterpolation,...

```
lambd = zfit.Parameter("lambda", -0.06, -1, -0.01)
frac = zfit.Parameter("fraction", 0.3, 0, 1)
```

```
gauss = zfit.pdf.Gauss(mu=mu, sigma=sigma, obs=obs)
exponential = zfit.pdf.Exponential(lambd, obs=obs)
model = zfit.pdf.SumPDF([gauss, exponential], fracs=frac)
```

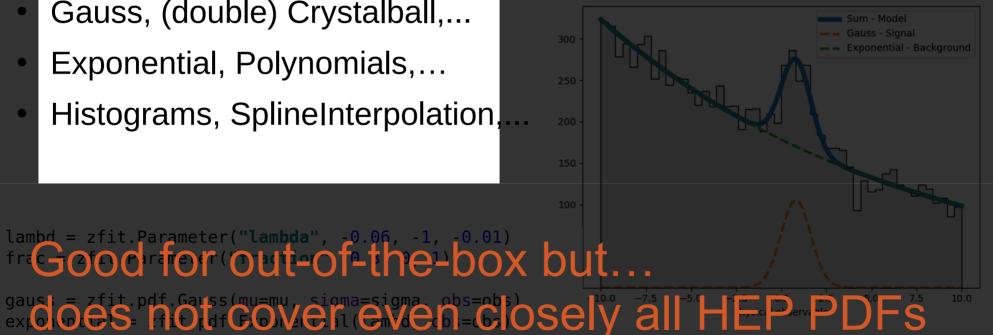




Example: Mass fit

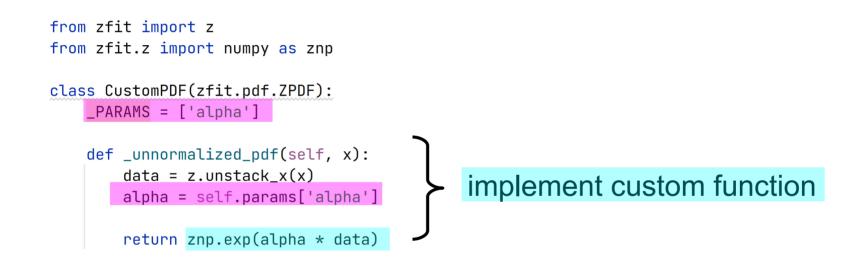


- Sum, Product, (Convolution)
- Gauss, (double) Crystalball,...
- Exponential, Polynomials,...
- Histograms, SplineInterpolation,



Custom PDF





Custom PDF



```
from zfit import z
  from zfit.z import numpy as znp
   class CustomPDF(zfit.pdf.ZPDF):
       _PARAMS = ['alpha']
       def _unnormalized_pdf(self, x):
           data = z.unstack_x(x)
           alpha = self.params['alpha']
           return znp.exp(alpha * data)
custom pdf = CustomPDF(obs=obs, alpha=0.2)
integral = custom_pdf.integrate(limits=(-1, 2))
sample = custom_pdf.sample(n=1000)
                                                            use functionality of model
          = custom pdf.pdf(sample)
prob
```

16 Sep 2022

sample

prob

PyHEP 2022 - zfit: extending to binned fits

use functionality of model

Example of zfit Base Classes

Or register integral

• integrate \rightarrow integrate

Can also override:

- pdf \rightarrow pdf
- sample \rightarrow _sample



from zfit.z import numpy as znp

class CustomPDF(zfit.pdf.ZPDF): PARAMS = ['alpha']

def _unnormalized_pdf(self, x):

alpha = self.params['alpha']

return znp.exp(alpha * data)

integral = custom_pdf.integrate(limits=(-1, 2))

= custom pdf.sample(n=1000)

custom pdf = CustomPDF(obs=obs, alpha=0.2)

= custom pdf.pdf(sample)

data = $z.unstack_x(x)$

from zfit import z

Arbitrary analytic shapes

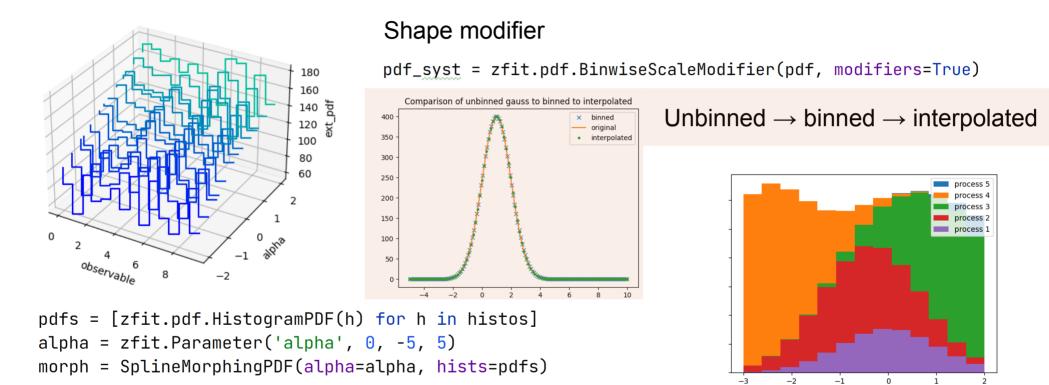
```
class P5pPDF(zfit.pdf.ZPDF):
    PARAMS = ['FL', 'AT2', 'P5p']
    N OBS = 3
    def unnormalized pdf(self, x):
        FL = self.params['FL']
        AT2 = self.params['AT2']
        P5p = self.params['P5p']
        costheta l, costheta k, phi = ztf.unstack x(x)
        sintheta k = tf.sqrt(1.0 - costheta k * costheta k)
        sintheta l = tf.sqrt(1.0 - costheta l * costheta l)
        sintheta 2k = (1.0 - \text{costheta } k + \text{costheta } k)
        sintheta 2l = (1.0 - costheta l * costheta l)
        sin2theta k = (2.0 * sintheta k * costheta k)
        cos2theta l = (2.0 * costheta l * costheta l - 1.0)
        pdf = ((3.0 / 4.0) * (1.0 - FL) * sintheta 2k +
               FL * costheta k * costheta k +
               (1.0 / 4.0) * (1.0 - FL) * sintheta 2k * cos2theta l +
               -1.0 * FL * costheta k * costheta k * cos2theta l +
               (1.0 / 2.0) * (1.0 - FL) * AT2 * sintheta 2k *
               sintheta 2l * tf.cos(2.0 * phi) + tf.sgrt(FL * (1 - FL))
               * P5p * sin2theta k * sintheta l * tf.cos(phi))
```

For example, create amplitude with ComPWA and fit with zfit

16 Sep 2022

More histograms





pdfs = [zfit.pdf.HistogramPDF(h) for h in histos]
sumpdf = zfit.pdf.BinnedSumPDF(pdfs)

Complete fit: Data



normal_np = np.random.normal(2., 3., size=10_000)

```
obs = zfit.Space("x", limits=(-2, 3))
```

mu = zfit.Parameter("mu", 1.2, -4, 6)
sigma = zfit.Parameter("sigma", 1.3, 0.5, 10)
gauss = zfit.pdf.Gauss(mu=mu, sigma=sigma, obs=obs)

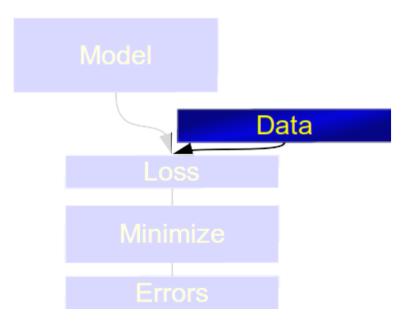
data = zfit.Data.from_numpy(obs=obs, array=normal_np)

nll = zfit.loss.UnbinnedNLL(model=gauss, data=data)

```
minimizer = zfit.minimize.Minuit()
result = minimizer.minimize(nll)
```

```
param_errors = result.hesse()
```

param_errors_asymmetric, new_result = result.errors()

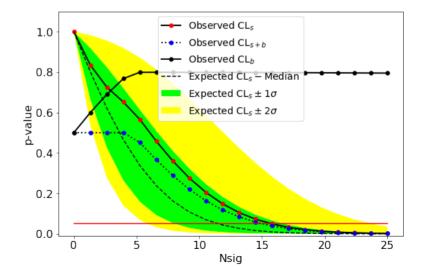


Back to HEP ecosystem: hepstats

- Inference library for hypothesis tests
- Takes model, data, loss from zfit
- sWeights, CI, limits, ... asymptotic or toys
- New: can also handle multi-dimensional PDFs

```
calculator = AsymptoticCalculator(loss, minimizer)
poinull = POIarray(Nsig, np.linspace(0.0, 25, 20))
poialt = POI(Nsig, 0)
ul = UpperLimit(calculator, poinull, poialt)
ul.upperlimit(alpha=0.05, CLs=True)
```

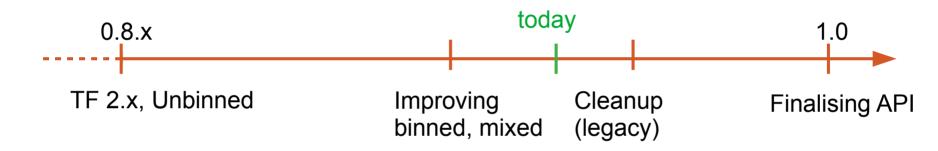








Public testing stage (pip install zfit)



A lot of experience and proven API, but also design flaws (global parameters, ...)

Continue to incorporate feedback and adaptability to other libraries

Binned fits: still rough edges (!)





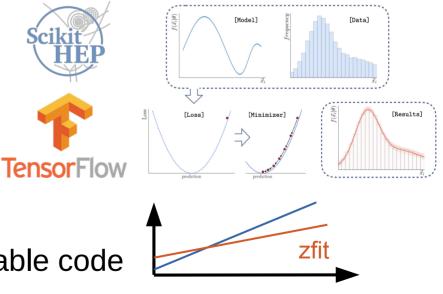
build stable model fitting ecosystem for HEP

- Integrate into HEP ecosystem functionality limited; stable API
- Technical requirements

performance; maintainability

• HEP requirements

advanced features; simply extendable code







• LHCb collision: https://physicsworld.com/wp-content/uploads/2018/08/LHCb-collision.png

Backup Slides

PyHEP 2022 - zfit: extending to binned fits

https://zfit.github.io/zfit/





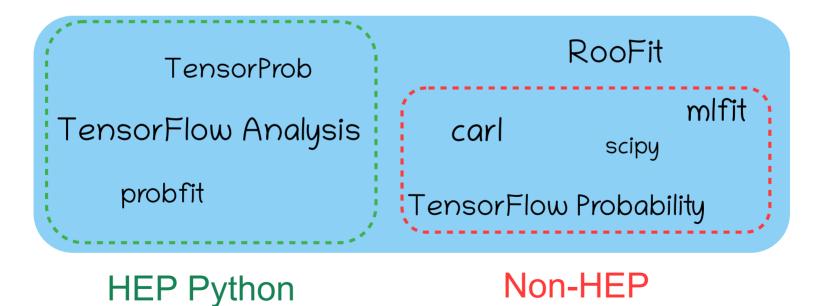
zfit@physik.uzh.ch Join the discussion!

- Backend & TF
- Amplitude
- K*II toys
- K*mumu Wilson coeffs
- Other fitting packages
- Zfit (associated) packages
- Zfit project
- Zfit elements examples

Fitting in Python



A lot of projects are around





Backend & TensorFlow

Backend: a comparison



- TensorFlow: supports the most features to this day
- PyTorch: missing advanced math (complex support, ...)
- Numpy/SciPy: Too slow, no gradient, no GPU
- JAX: very promising, but no globals (cache,...), only static known shapes (adaptive algorithms, accept-reject...), only JAX/Numpy arrays compatible
- SymPy: limited to mathematical expressions (no control-flow,...) but can convert to any other backend (used by TensorWaves)

Tracing

Includes GPU support, optimizations, caching,...

Recent rise of big data industry created libraries that support this

Autograd "analytic" gradient of function

Backend: tracing and autograd

execute Python once, remember (algebraic) computation









build *the* stable model fitting ecosystem for HEP

• Integrate into HEP ecosystem

functionality limited; stable API

• Technical requirements

performance; maintainability

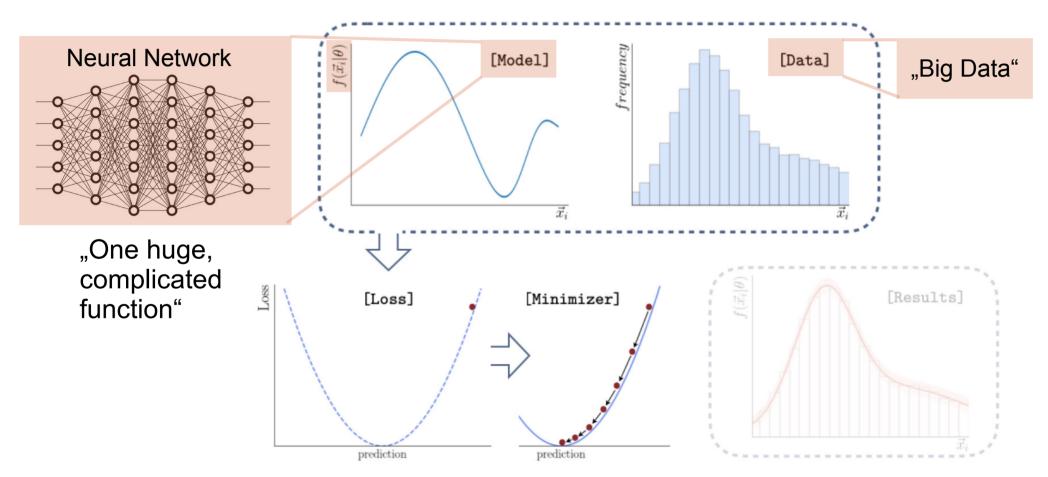
• HEP requirements

advanced features; simply extendable code

Deep Learning lessons for model fitting

Deep Learning





Main backend: TensorFlow

• By Google, highly popular (150k \bigstar , 4th on \bigcirc)



ZH

<u>\</u>

Main backend: TensorFlow

- By Google, highly popular (130k \bigstar , 4th on \bigcirc)
- Used in multiple physics libraries and analyses



Main backend: TensorFlow

- By Google, highly popular (150k \bigstar , 4th on \bigcirc)
- Consists of "two parts":
 - High level API for building neural networks (NOT used!)
 - Low level API with Numpy-style syntax

tf.sqrt, tf.random.uniform,...

- Two modes:
 - "numpy"-like (full Python flexibility)
 - "compiled" (very performant)





GPU/Multi CPU support

Delegating the workload

C++ |

the workload		ZTIC
library (RooFit,)	Numpy based	zfit
	SciPy	TF Probability
	∲Numba 💭 NumPy	TensorFlow Intel NVIDIA.

Low level handling

Parallelization/GPU

HEP specific content/API

Models

Gradients

Computational

optimizations

Delegating the workload



	C++ library (RooFit,)	Numpy based	zfit
HEP specific content/API			
Models		SciPy	TF Probability
	"Stopping	on tha	
	"Stepping shoulders	of a giant"	
	onodiaoro	SNumba Numby	TensorFlow
Low level handling		P	thon

Delegating the workload



	C++ library (RooFit,)	Numpy based	zfit
HEP specific content/API			API & Workflow
Models		SciPy	TF Probability
Gradients	llood 9 mo	intainad (I)	
Computational optimizations	Used & ma by industry		
Parallelization/GPU	by modeling	∲Numba MumPy	TensorFlow
Low level handling		e p	thon

Can we express model fitting as static graphs?

Yes!





Definition of computation, shape etc. (add static knowledge)
 Compilation of the graph

3) Execution of computation (re-use optimized graph)

Inside TF, hidden to end-user

HPC: the more is know *before* the execution, the better

TensorFlow takes care of *how* to use this knowledge

Graph elements



... do not have to be constant!

Parameters

Can change their value

Random numbers

Generate newly on every graph execution: MC integration,...

Control flow (if, while)

Steer the execution: Accept-reject sampling (while), etc.

Static, not constant



Similarity	Complicated Models	Large Data	Composed loss	Minimization	Results and uncertainties
HEP	Non-trivial functions	Whole Dataset	simultaneous, constraints	Global min, 2 nd derivative algorithm	Hesse, profiling
Deep Learning	Combine many, trivial functions	Many, small Batches	Anything! (GANs, RL,)	Local (!) min, 1 th derivative, many steps	None
Conclusion					



		Large Data Bui	Composed loss		
hep W	Non-trivial Mat is a	Whole Deep I	4	Global min, 2 nd clerivative Glg J D f A r y	Hesse, P ^{filing}
	Combine many, trivial functions	Many, small Batches	Anything! (GANs, RL,)	Local (!) min, 1 th derivative, many steps	None



	Complicated Models	Large Data	Composed loss		
	Non-trivial functions	Whole Dataset	simultaneous, constraints	Global min, 2 nd derivative algorithm	Hesse, profiling
	Combine many, trivial functions	Many, small Batches	Anything! (GANs, RL,)	Local (!) min, 1 th derivative, many steps	None
				Optimizers "analytic" derivatives!	No support, but simple
16 Sen 2022		PvHEP 2022 - zfit: e	xtending to binned fits		94

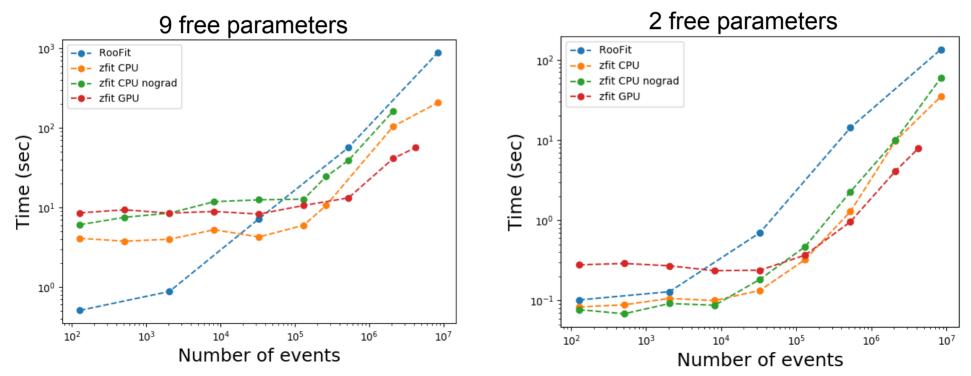


	Complicated Models	Large Data	Composed loss		
HEP	Non-trivial	whole httperfo	simultaneous,	Global min, ^{2nd} dorivativ al ÇOIMPL	Hesse, Iting
	Combine many, trivial functions	Many, small Batches	Anything! (GANs, RL,)	Local (!) min, 1 th derivative, many steps	None
				Optimizers "analytic" derivatives!	
16 Sep 2022		PyHEP 2022 - zfit: e:	xtending to binned fits		95

Scalability: Performance



Fitting time (lower is better): RooFit vs. zfit





Amplitude

16 Sep 2022

PyHEP 2022 - zfit: extending to binned fits

100

('K*(892)0', ('K+', 'pi-'), bw amplitude)] $COEFFS = {...}$ D2Kpipi0 = Decay('D0', ['K+', 'pi-', 'pi0'])

('K(2)*(1430)0', ('K+', 'pi-'), bw amplitude),

('K*(892)+', ('K+', 'pi0'), bw amplitude),

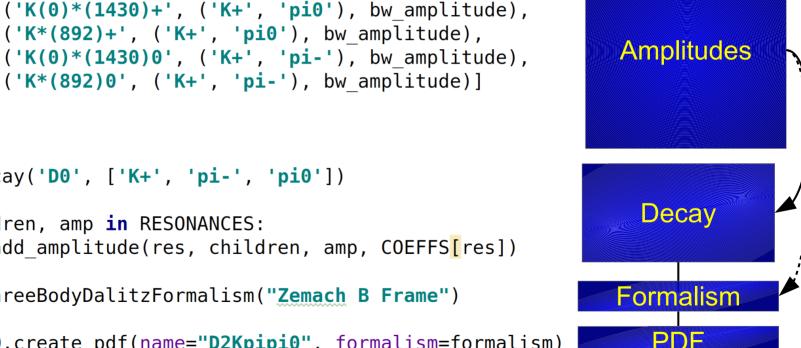
for res, children, amp **in** RESONANCES: D2Kpipi0.add amplitude(res, children, amp, C0EFFS[res])

RESONANCES = [('rho(770)', ('pi-', 'pi0'), bw amplitude),

formalism = ThreeBodyDalitzFormalism("Zemach B Frame")

pdf = D2Kpipi0.create pdf(name="D2Kpipi0", formalism=formalism)

Example amplitude







Angular toys

$B^0 \rightarrow K^{*0}l^+l^-$ angular: toy study

Sensitivity study

- draw toys (sample) from PDF
- Fit to sample

```
for i in range(ntoys):
```

```
# set initial sampling values
for param in params:
    param.set_value(...)
```

sampler.resample()

```
# set random initial values
for param in params:
    param.set_value(...)
```

result = minimizer.minimize(nll)

if result.converged:

. . .

$B^0 \rightarrow K^{*0}l^+l^-$ angular: toys

Sensitivity study

- draw toys (sample) from PDF
- Fit to sample

```
for i in range(ntoys):
```

set initial sampling values
for param in params:
 param.set_value(...)

sampler.resample()

```
# set random initial values
for param in params:
    param.set_value(...)
```

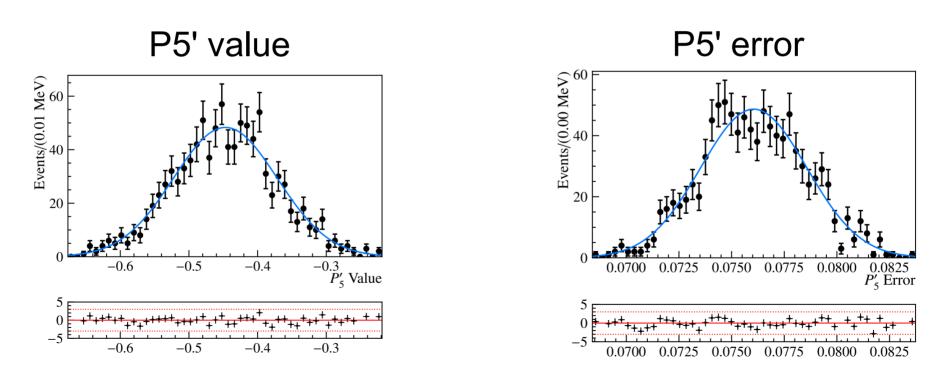
result = minimizer.minimize(nll)

if result.converged:

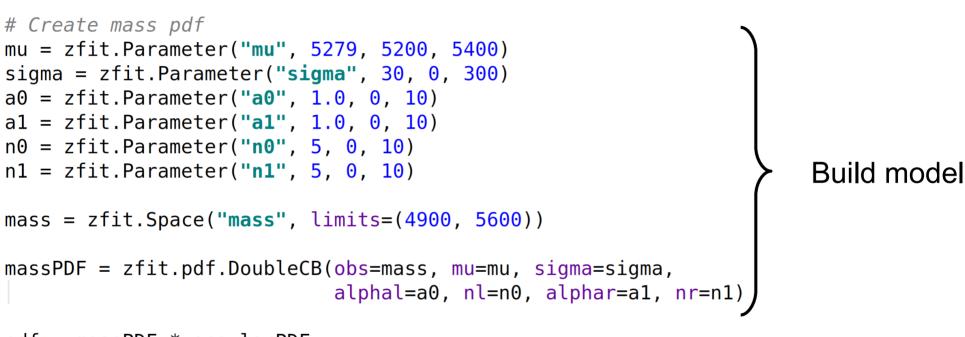
. . .

 $B^0 \rightarrow K^{*0}l^+l^-$ angular: toy study

Result of toy study



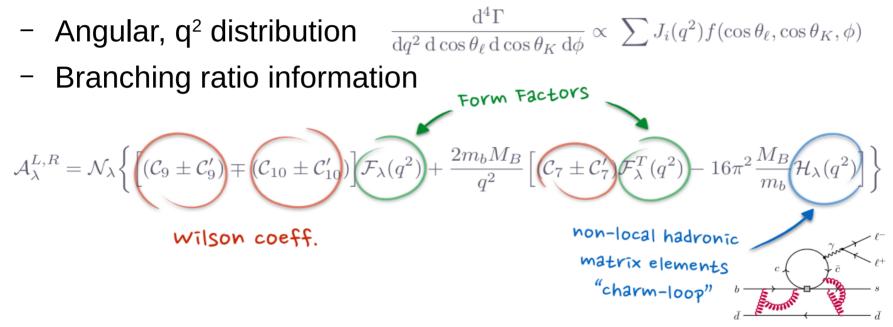
Extending with a mass shape



pdf = massPDF * angularPDF

$B^0 \rightarrow K^{*0} \mu^+ \mu^-$ full amplitude







Fitting libraries and comparison

Python model fitting in HEP



- Scalable: large data, complex models
- Pythonic: use Python ecosystem/language
- Specific HEP functionality:
 - Normalization: specific range, numerical integration,...
 - Composition of models
 - Multiple dimensions
 - Custom models
 - Non-trivial loss (constraints, simultaneous,...)





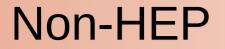
- Limited customization and extendibility
- Sub-optimal scalability for ever larger datasets and modern computing infrastructure
- Isolated, aging ecosystem, no cutting-edge software
- Not Python native
 - Memory allocation errors
 - Arbitrary C++ limitations
 - No real integration into the Python ecosystem



Probfit, TensorProb,...

- Lack generality and extendibility
- "experimental", but great proof of concept
 - API and Python in general
 - Computational backends (e.g. Cython, TensorFlow)
 - Building an ecosystem (iminuit,...)

General impression in comparison with other HEP packages





Scipy, Imfit, TensorFlow Probability,...

- Lack of specific HEP features
 - Normalization: specific range, numerical integration,...
 - Composition of models
 - Multiple dimensions
 - Custom models
- Irrelevant functionality supported in API
 - Survival function, ...

TFA: approach & differences



- accept-reject as tf.while_loop, Dataset input,...
- ...and hide the tedious, unambiguous parts
 - automatic normalization, Tensor cache, ...
- Well defined structures, e.g.
 - String name order (like columns) in PDFs, data, limits,...
 - pdf("x") * pdf("y") => pdf("x", "y")
 1-dim
 1-dim
 2-dim
 - Local/recursive dependency resolution of Parameters



Zfit related packages

phasespace



- Package for phasespace generation of particles
- Covers functionality of TGenPhaseSpace (and more)
- Pure Python (& TensorFlow), integrates seemless with zfit

```
pion = GenParticle('pi+', PION_MASS)
kaon = GenParticle('K+', KAON_MASS)
kstar = GenParticle('K*', KSTARZ_MASS).set_children(pion, kaon)
gamma = GenParticle('gamma', 0)
bz = GenParticle('B0', B0_MASS).set_children(kstar, gamma)
```

```
weights, particles = bz.generate(n_events=1000)
```



Zfit: project description

PyHEP 2022 - zfit: extending to binned fits

zfit: stable core

zfit project

- Unbinned fits, binned WIP
- n-dim models with integral, pdf, sample
- zfit-physics: HEP specific content
 - BreitWigner, DoubleCB,...
 - Faster development, more content —
 - Ideal for contributions
 - Auto testing of new pdfs/func
 - Contribution guidelines •





build stable model fitting ecosystem for HEP

• Integrate into HEP ecosystem

functionality limited; stable API

• Technical requirements

performance; maintainability

• Analysis requirements

advanced features; simply extendable code



scalable pythonic fitting





build the stable model fitting ecosystem for HEP ...the time has come





build the stable model fitting ecosystem for HEP

- Integrate into HEP ecosystem
 functionality limited; stable API
- Technical requirements

performance; maintainability

• Analysis requirements

advanced features; simply extendable code

Ecosystem: API & Workflow



Establish a stable API

- High level libraries (statistics, plotting,...)
 - "code against an interface, not an implementation"
- Replace each component
 - Allow other libraries to implement custom parts

Many discussions with community to avoid splitting/duplication

Pythonic

- Pure Python («pip install zfit»)
- Integrated into python ecosystem
 - Load ROOT files (uproot, no ROOT dependence!)
 - Use Minuit for minimization (iminuit)
 - Data preprocessing with Pandas DataFrame
 - Plotting with matplotlib
 - High level statistics (lauztat, more WIP)
- Extendable classes
 - e.g. custom PDF





Scalable

- TensorFlow hidden backend, uses graphs
 - numpy-like syntax
 - parallelization on CPU/GPU, analytic gradient,...
- Writing functions simple for users *and* developers
 - No Cython, MPI, CUDA,... for state-of-the-art performance
 - No low-level maintenance required!
- Used in multiple physics libraries and analyses





Scalable: TensorFlow

- Deep Learning framework by Google
- Modern, declarative graph approach
- Built for highly parallelized, fast communicating CPU, GPU, TPU,... clusters
- Built to use «Big Data»







Zfit library examples

Minimize Python function



def func(x):
 x = np.array(x) # make sure it's an array
 return np.sum((x - 0.1) ** 2 + x[1] ** 4)

func.errordef = 0.5

params = [1, -3, 2, 1.4, 11]

result = minimizer.minimize(func, params)

PyHEP 2022 - zfit: extending to binned fits

Model, loss building

sum of two pdfs

sum_pdf = zfit.pdf.SumPDF([gauss, exponential], fracs=frac)

shared parameters

mu_shared = zfit.Parameter("mu_shared", 1., -4, 6)

gauss1 = zfit.pdf.Gauss(mu=mu_shared, sigma=sigma1, obs=obs)
gauss2 = zfit.pdf.Gauss(mu=mu_shared, sigma=sigma2, obs=obs)

simultaneous loss

nll1 = zfit.loss.UnbinnedNLL(model=gauss1, data=data1)
nll2 = zfit.loss.UnbinnedNLL(model=gauss2, data=data2)
nll_simultaneous2 = nll1 + nll2

From classical

to more TensorFlow





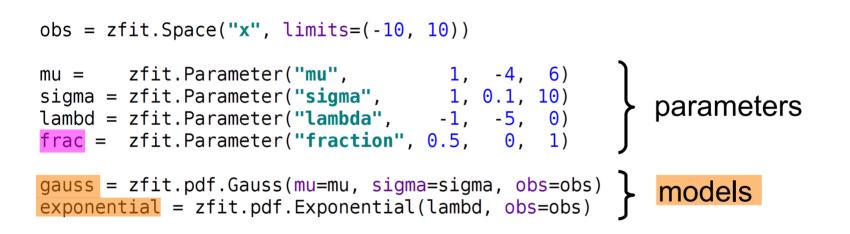
```
Simple combinations
func n = zfit.func.ZFunc(...) # pseudo code
func = func 1 + func 2 * func 3
Composite Parameter
pdf = zfit.pdf.Gauss(mu=tensor1, sigma=4)
Custom Loss
loss = zfit.loss.SimpleLoss(lambda: tensor loss)
```

up to pure TensorFlow

=> use all of zfit functionality like minimizers

Model building







Simultaneous fit

