Easy visualization of experimental sensitivity to parton distributions

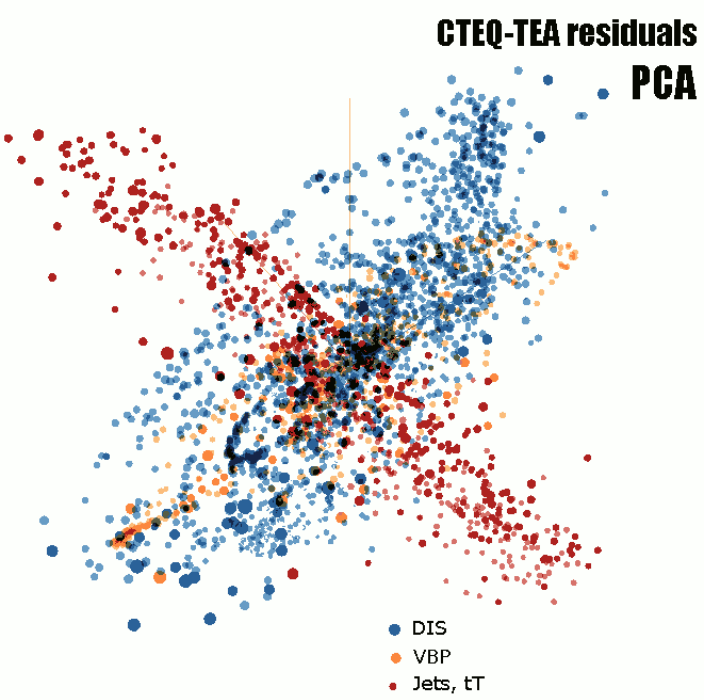
Pavel Nadolsky

Southern Methodist University

In collaboration with Bo Ting Wang, T. J. Hobbs, S. Doyle, J. Gao, T.-J. Hou, F. Olness

arXiv:1803.02777

PDFSense program, article, and figures:
http://tinyurl.com/PDFSense



How sensitive is an experiment to a PDF? Can we know it **before** doing the global fit?

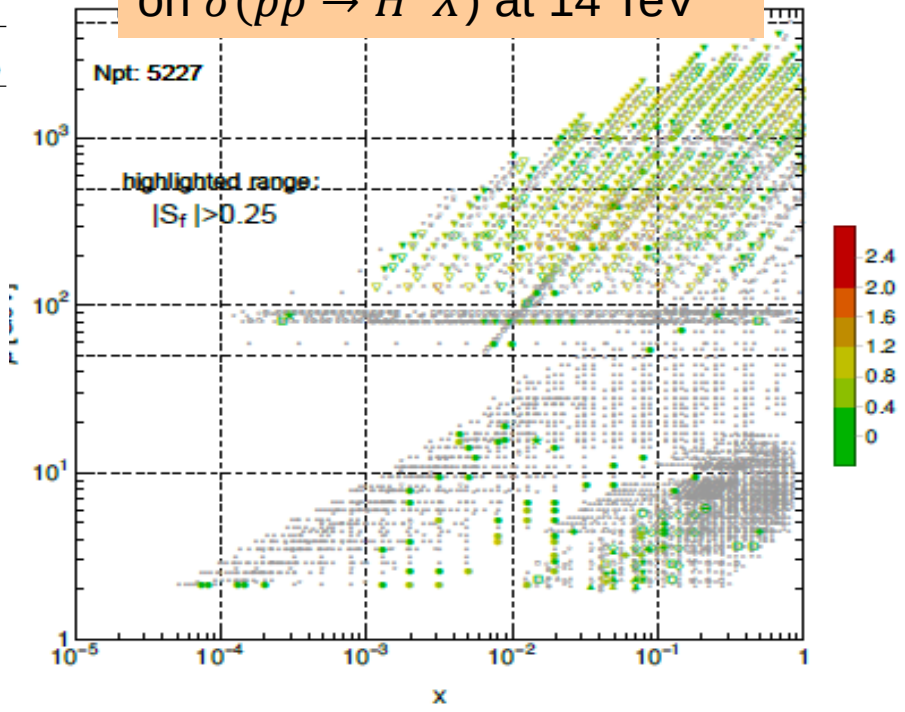
PDFSense estimates…

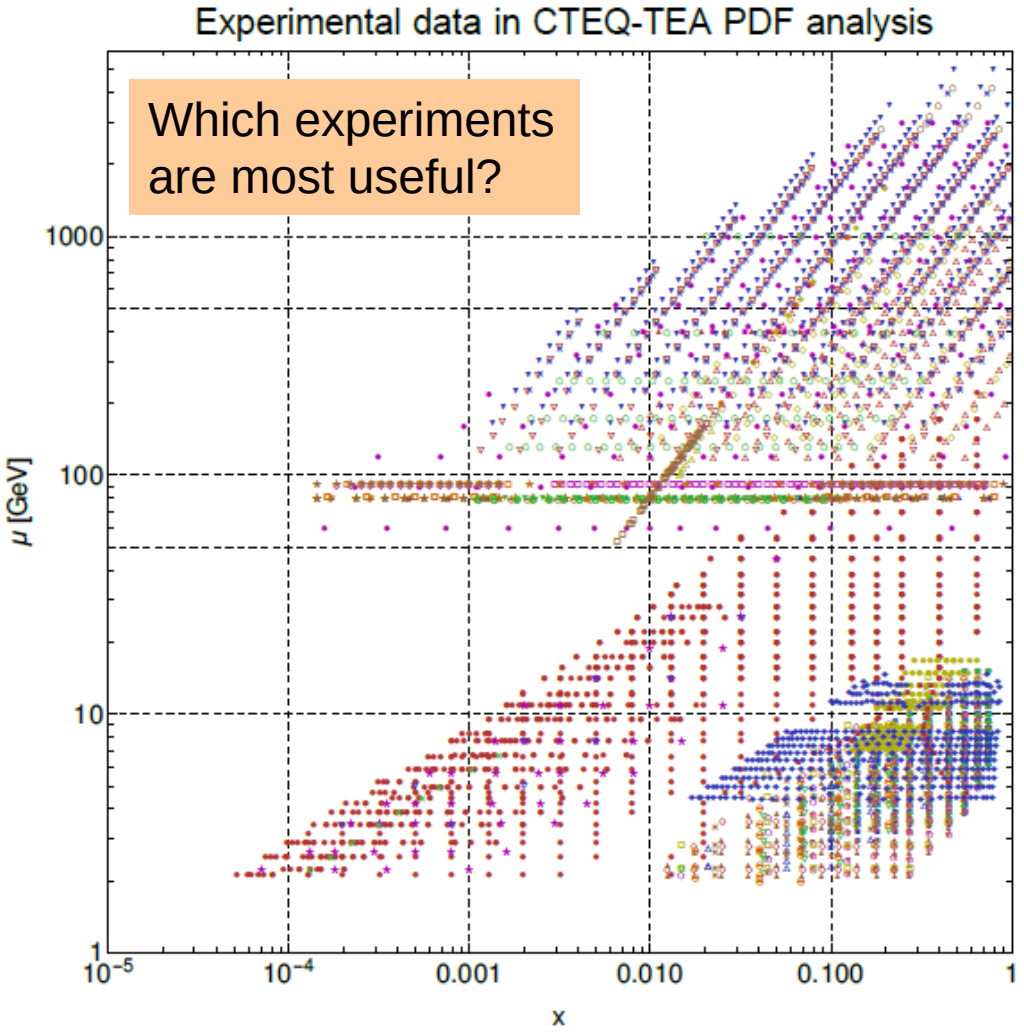
…ranking of strength of sensitivities of experimental data sets to PDF flavors without (re-)doing the full global fit

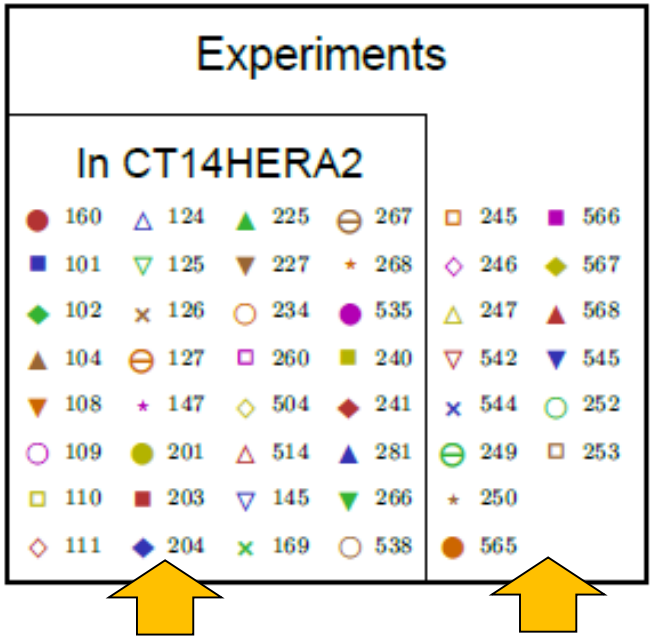
No.	Exp. ID	N_d	Rankings $\sum_f \lvert S_f^E \rvert$	$\langle \sum_f \lvert S_f^E \rvert \rangle$	$\lvert S_{\bar{d}}^E \rvert$	$\langle \lvert S_{\bar{d}}^E \rvert \rangle$	$\lvert S_{\bar{u}}^E \rvert$	$\langle \lvert S_{\bar{u}}^E \rvert \rangle$	$\lvert S_g^E \rvert$	$\langle \lvert S_g^E \rvert \rangle$	$\lvert S_u^E \rvert$	$\langle \lvert S_u^E \rvert \rangle$	$\lvert S_d^E \rvert$	$\langle \lvert S_d^E \rvert \rangle$	$\lvert S_s^E \rvert$	$\langle \lvert S_s^E \rvert \rangle$
1	160	1120.	620.	0.0922	B		A	3	A	3	A	3	B		C	
2	**545**	185	232.	0.209	C	3	C	3	B	2					C	3
3	111	86	218.	0.423	C	1	C	1		3	B	1	C	2		
4	**542**	158	194.	0.204	C	3	C	3	B	2					C	3
5	101	337	184.	0.0909			C		C		B	3	C			
6	104	123	169.	0.229	C	2					C	2	B	2		
7	102	250	141.	0.0938	C				C	3	C	3	C	3		
8	109	96	115.	0.199	C	2	C	2		3	C	2	C	3		
9	201	119	113.	0.158	C	2	C	2				3				
10	204	184	103.	0.0935		3	C	3			C	3				
11	110	69	89.3	0.216		3		3	C	2		3		2		3
12	108	85	82.4	0.161		3		3		3		3	C	3		
13	538	133	66.2	0.0829					C	3						
14	124	38	58.9	0.258		3		3				3		3	C	1
15	127	38	49.4	0.217		3		3				3		3	C	1
16	**544**	140	48.7	0.058						3						
17	126	40	48.	0.2		3		3				3		3	C	1
18	**250**	42	41.5	0.165		3		3		3		3		2		
19	268	41	39.6	0.161		3		3				3		3		3
20	**249**	33	39.2	0.198		2		3				3		2		3
21	514	110	36.8	0.0557						3						
22	125	33	36.7	0.185		3		3				3		3		2

…kinematical distributions of sensitivities to the PDFs in the $\{x, \mu\}$ plane

Sensitivity to the PDF error on $\sigma(pp \to H^0 X)$ at 14 TeV







in the CT14 HERA2 NNLO fit (N_{pt} =3250 points)

considered for the CTEQ-TEA fit (N_{pt}=734 points from ATLAS, CMS, LHCb)

Experiments are labeled according to the experimental ID's in the CT fit. The ID's are listed in the backup

1xx: DIS, **2xx**: vector boson production, **5xx**: jet and $t\bar{t}$ production.

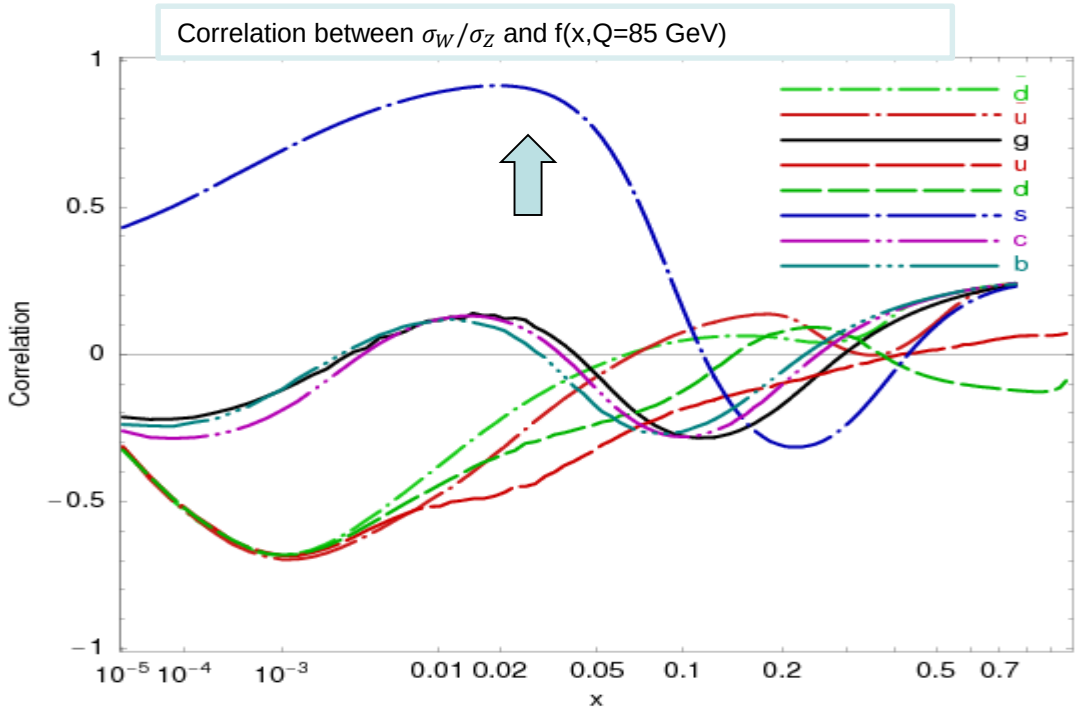
PDFSense: operating principles

- ~~PDF reweighting~~
- ~~Hessian profiling~~

} ⇒ mcgen (1607.06066), ePump (next talk)

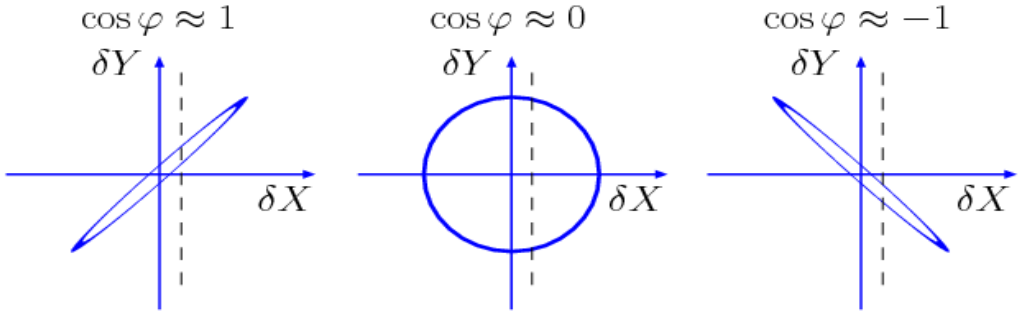
- Generalized correlations (**sensitivities** S_f) comparing experimental and PDF uncertainties for CT14HERA2 data points
- Based on the output from the CT14HERA2 fit in a new format (**shifted residuals for Hessian PDFs**) – available on the PDFSense website

Correlations carry useful, but limited information



CTEQ6.6 [arXiv:0802.0007]:
$\cos\varphi > 0.7$ shows that the ratio σ_W/σ_Z at the LHC must be sensitive to the strange PDF $s(x,Q)$

$\cos\varphi \approx \pm 1$ suggests that a measurement of X **may** impose tight constraints on Y



But, Corr[X,Y] between **theory** cross sections X and Y does not tell us about **experimental** uncertainties

Solution: choose X = a shifted residual r_i

$r_i(\vec{a}) = \frac{T_i(\vec{a}) - D_i^{sh}(\vec{a})}{s_i}$ are N_{pt} **shifted residuals** for point i, PDF parameters $\vec{a}$

$\bar{\lambda}_\alpha(\vec{a})$ are N_λ **optimized nuisance parameters** (dependent on $\vec{a}$)

The $\chi^2(\vec{a})$ for experiment E is

$$\chi^2(\vec{a}) = \sum_{i=1}^{N_{pt}} r_i^2(\vec{a}) + \sum_{\alpha=1}^{N_\lambda} \bar{\lambda}_\alpha^2(\vec{a}) \approx \sum_{i=1}^{N_{pt}} r_i^2(\vec{a})$$

$T_i(\vec{a})$ is the theory prediction for PDF parameters $\vec{a}$
D_i^{sh} is the data value **including the optimal systematic shift**

$$D_i^{sh}(\vec{a}) = D_i - \sum_{\alpha=1}^{N_\lambda} \beta_{i\alpha} \bar{\lambda}_\alpha(\vec{a})$$

s_i is the uncorrelated error

$r_i(\vec{a})$ and $\bar{\lambda}_\alpha(\vec{a})$ are tabulated or extracted from the cov. matrix ⇒ backup slides

Vectors of data residuals

For every data point i, construct a vector of residuals $r_i(\vec{a}_k^{\pm})$ for 2N Hessian eigenvectors. $\mathrm{k} = 1, \ldots, N$, with $N = 28$ for CT14 NNLO.

For example, **define**

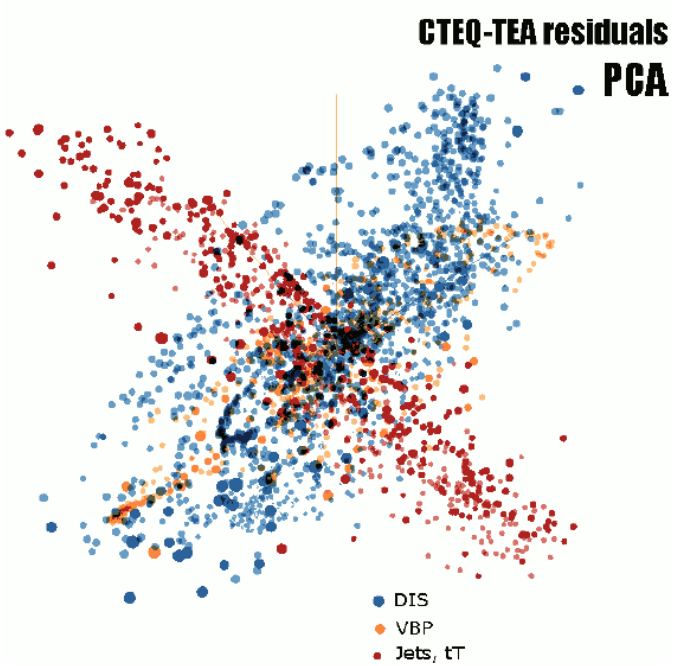
$$\vec{\delta}_i = \{\delta_{i,1}^+, \delta_{i,1}^-, \ldots, \delta_{i,N}^+, \delta_{i,N}^-\} \quad [\mathrm{N} = 28]$$

$$\delta_{i,k}^{\pm} \equiv \left(r_i(\vec{a}_k^{\pm}) - r_i(\vec{a}_0)\right) / \langle r_0 \rangle_E$$

-- a 56-dim vector normalized to $\langle r_0 \rangle_E$, the root-mean-squared residual for the experiment E for the central fit $\vec{a}_0$

$$\langle r_0 \rangle_E \equiv \sqrt{\frac{1}{N_{pt}} \sum_{i=1}^{N_{pt}} r_i^2(\vec{a}_0)} \approx \sqrt{\frac{\chi_E^2(\vec{a}_0)}{N_{pt}}}$$

$\langle r_0 \rangle_E \approx 1$ in a good fit to E



The TensorFlow Embedding Projector (http://projector.tensorflow.org) represents CT14HERA2 $\vec{\delta}_i$ vectors by their 10 principal components indicated by scatter points. A sample 3-dim. projection of the 56-dim. manifold is shown above. A symmetric 28-dim. representation can be alternatively used.

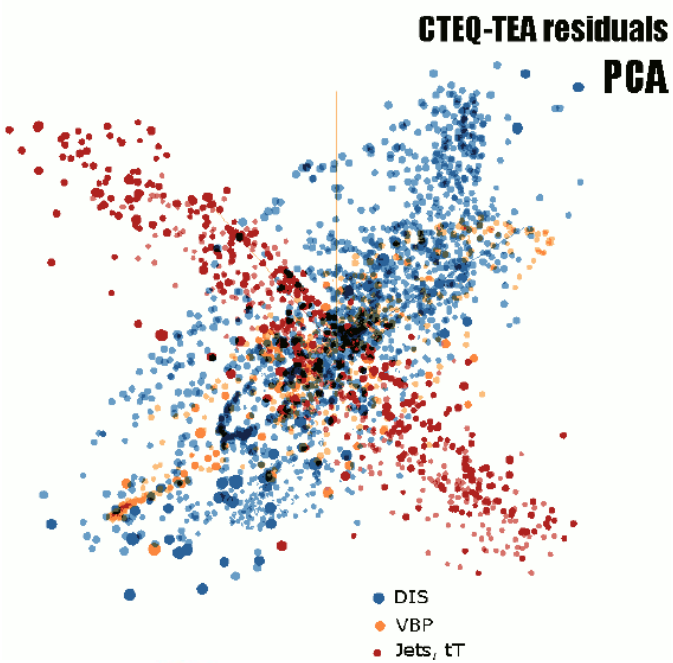
Manifolds of data residuals

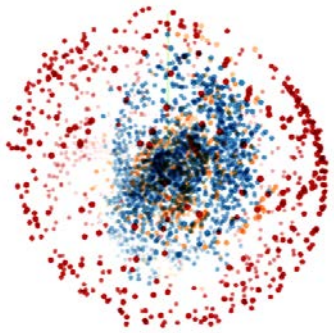
The 2N-dimensional distribution of $\vec{\delta}_i$ is easy to analyze with data-mining tools…

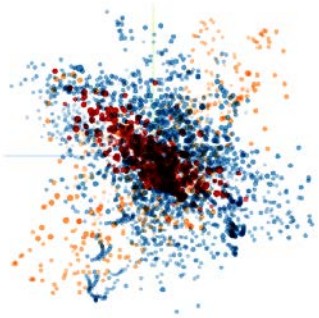
…to sort the fitted data points according to their PDF dependence (expressed by lengths and directions of $\vec{\delta}_i$);

…to identify high-value data points (having long $\vec{\delta}_i$ that point away from the rest of vectors).

Some projections separate DIS, DY, jet and $t\bar{t}$ data residuals according to their PDF dependence.





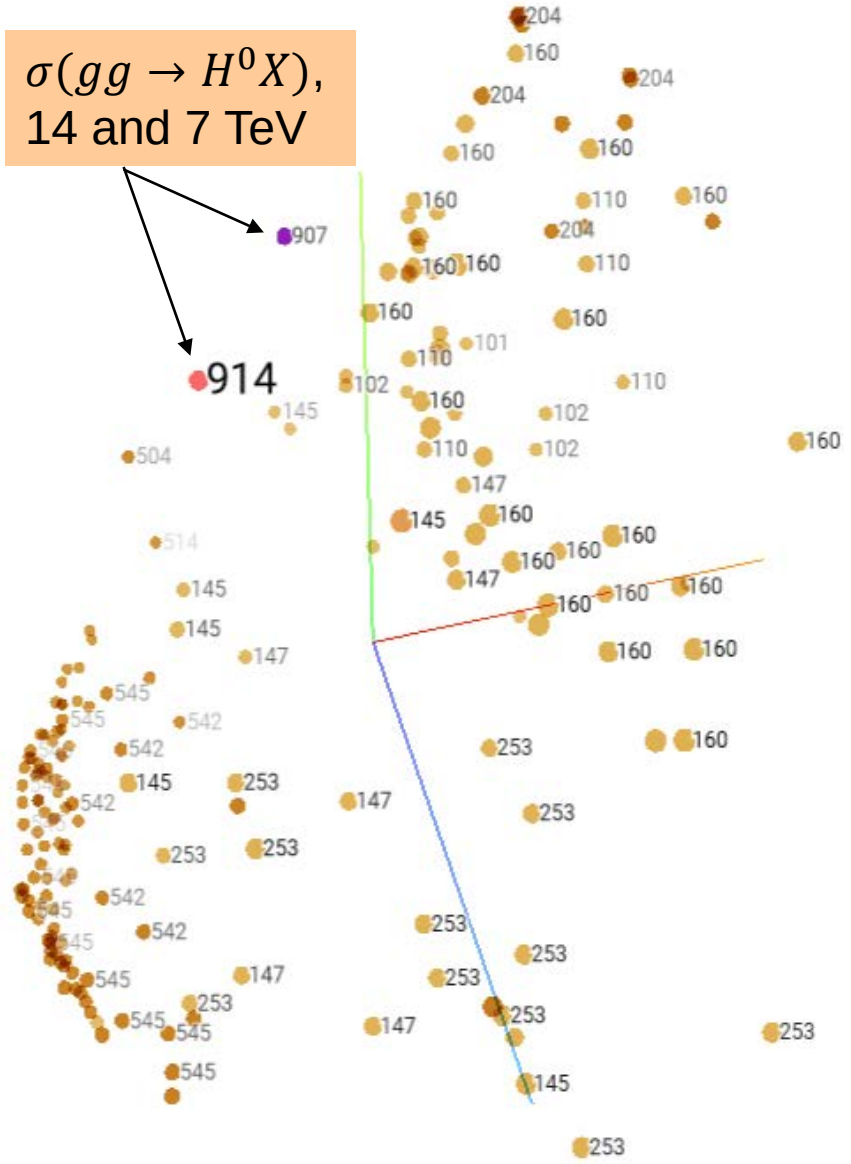


A PDF-dependent quantity f, such as the Higgs cross section at 7 or 14 TeV (ID=907, 914), defines a direction $\vec{\delta}_f$ in the (2)N-dim space.

The 3-dim projection on the right shows 300 vectors $\vec{\delta}_i$ of the CT14HERA2 global set whose directions are closest to $\vec{\delta}_f(\sigma(H^0))$. **These vectors are given by the experiments:**

160=HERA I+II; 101, 102=BCDMS; 110=CCFR F2p; 147, 145=HERA I+II c, b; 204=E866 σ_{pp}; 253=Z p_T 8 TeV; 542, 545=CMS jets 7, 8 TeV; 504, 514=Tevatron jets

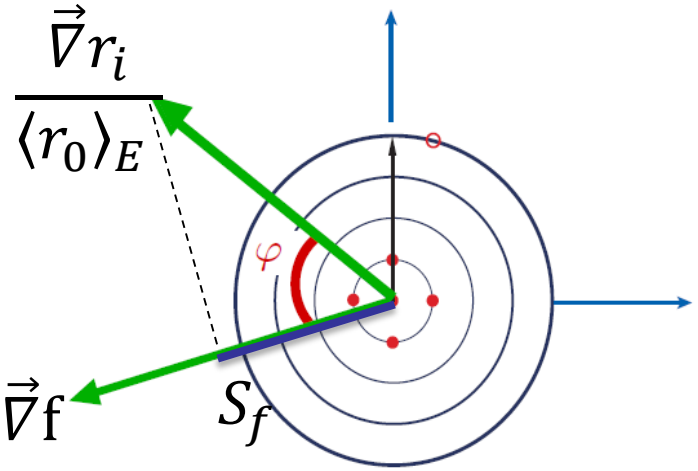
The net constraint of the i-th point on $\sigma(H)$, including systematic errors, is quantified by the projection of $\vec{\delta}_i$ on $\vec{\boldsymbol{\delta}}_{\boldsymbol{f}}[\sigma(H)]$, called the sensitivity $S_{f,i}$.



Sensitivity of expt E = sum of $S_{f,i}$ over data points in E

Correlation C_f and sensitivity S_f

The relation of data point i on the PDF dependence of f can be estimated by:



- $C_f \equiv \mathrm{Corr}[\rho_i(\vec{a})), f(\vec{a})] = cos\varphi$

$\vec{\rho}_i \equiv \vec{\nabla} r_i / \langle r_0 \rangle_E$ -- gradient of r_i normalized to the r.m.s. average residual in expt E;

$$\left(\vec{\nabla} r_i\right)_k = \left(r_i(\vec{a}_k^+) - r_i(\vec{a}_k^-)\right)/2$$

C_f is **independent** of the experimental and PDF uncertainties. In the figures, take $|C_f| \gtrsim 0.7$ to indicate a large correlation.

- $S_f \equiv |\vec{\rho}_i| cos\varphi = C_f \dfrac{\Delta r_i}{\langle r_0 \rangle_E}$ -- projection of $\vec{\rho}_i(\vec{a})$ on $\vec{\nabla} f$

S_f is proportional to $\cos\varphi$ and the ratio of the PDF uncertainty to the experimental uncertainty. We can sum $|S_f|$.

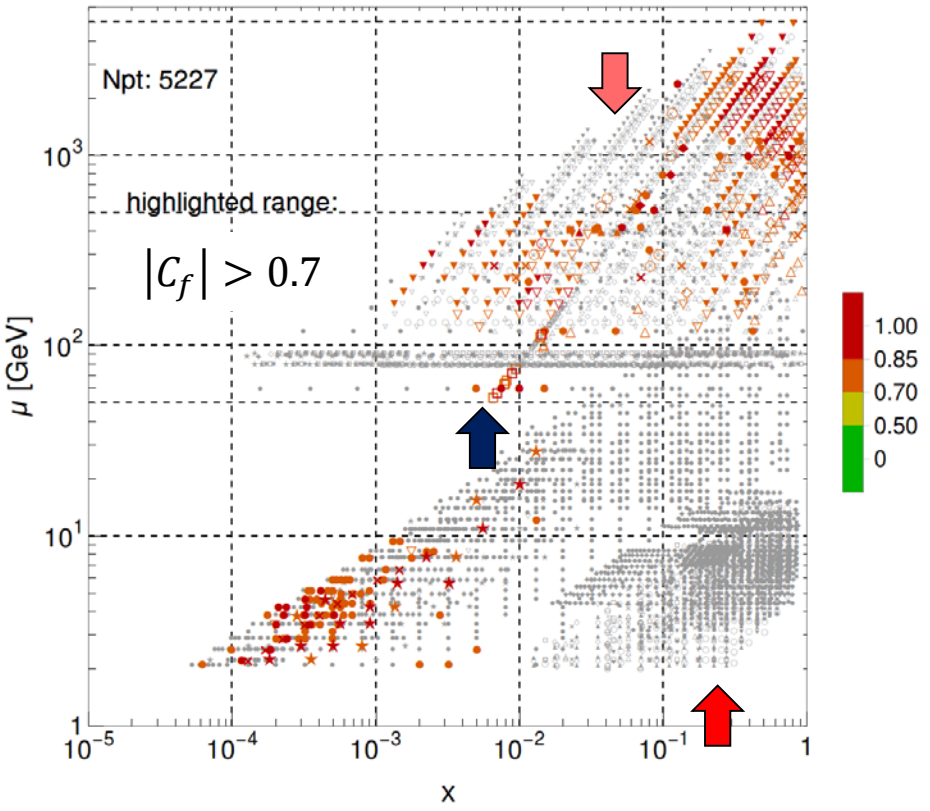
In the figures, take $|S_f| > 0.25$ to be significant.

Points with $|C_f| > 0.7$, $|S_f| > 0.25$ for $g(x, \mu)$

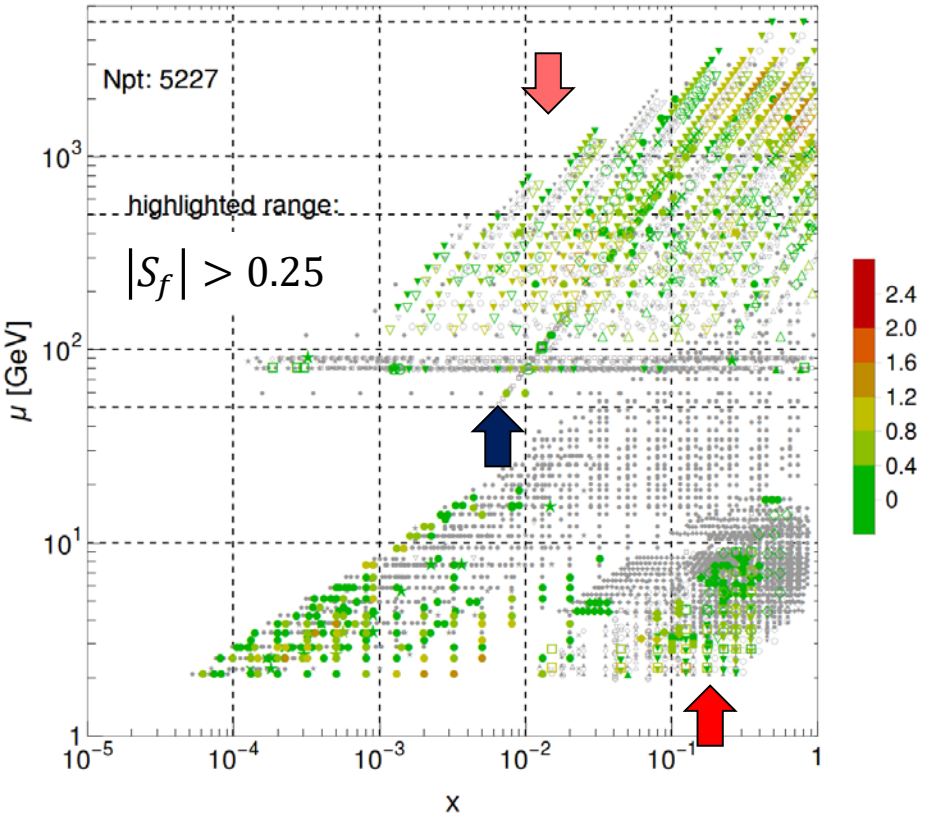
C_f does not identify fixed-target DIS points sensitive to $g(x_i, \mu_i)$. But S_f does.

Corr. syst. errors smear S_f over many data points for jet production, etc. Z p_T data (blue arrow) have high $|C_f|$ and $|S_f|$, a small number of points.

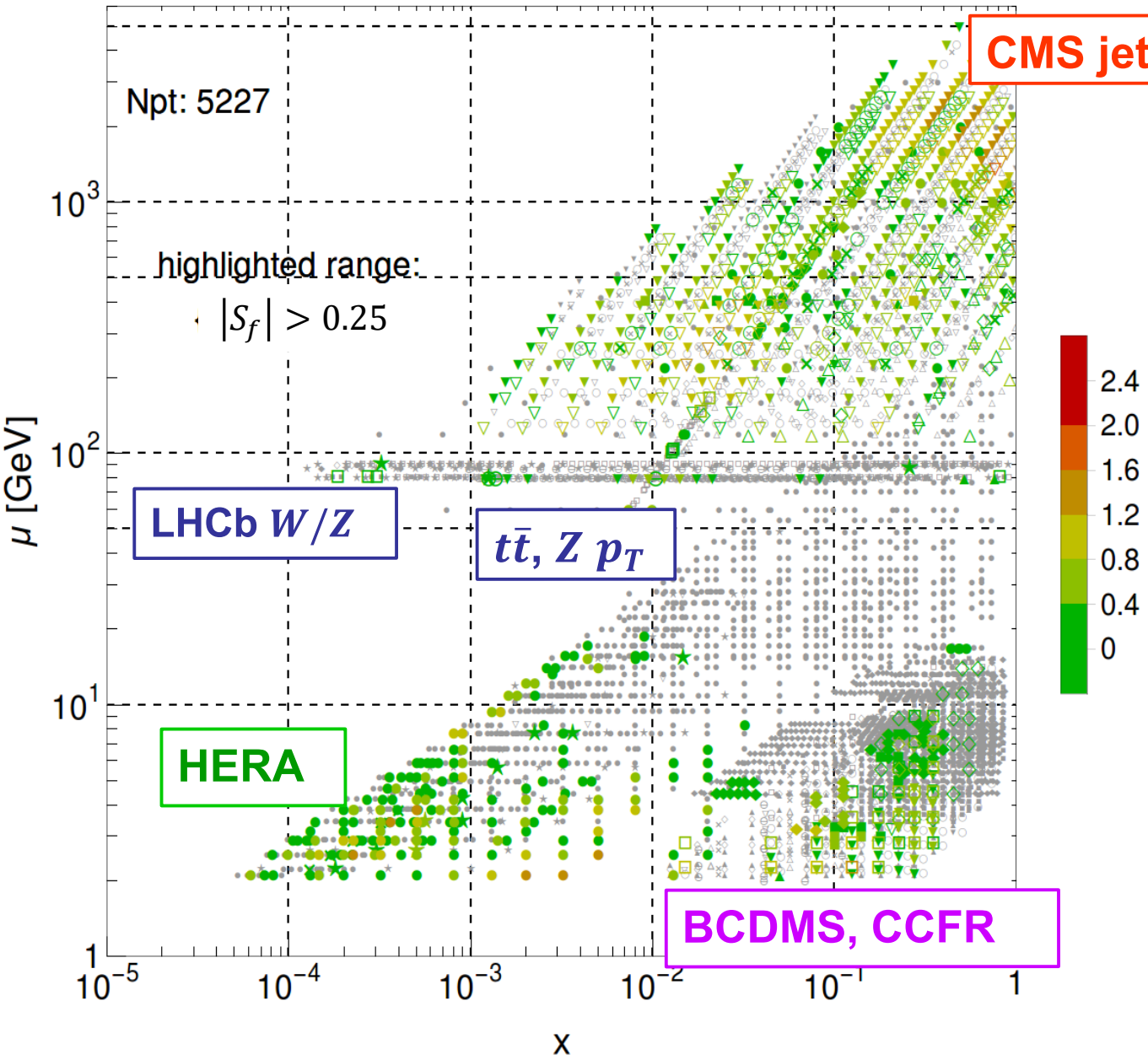
| C_f | for g(x,μ), CT14HERA2NNLO



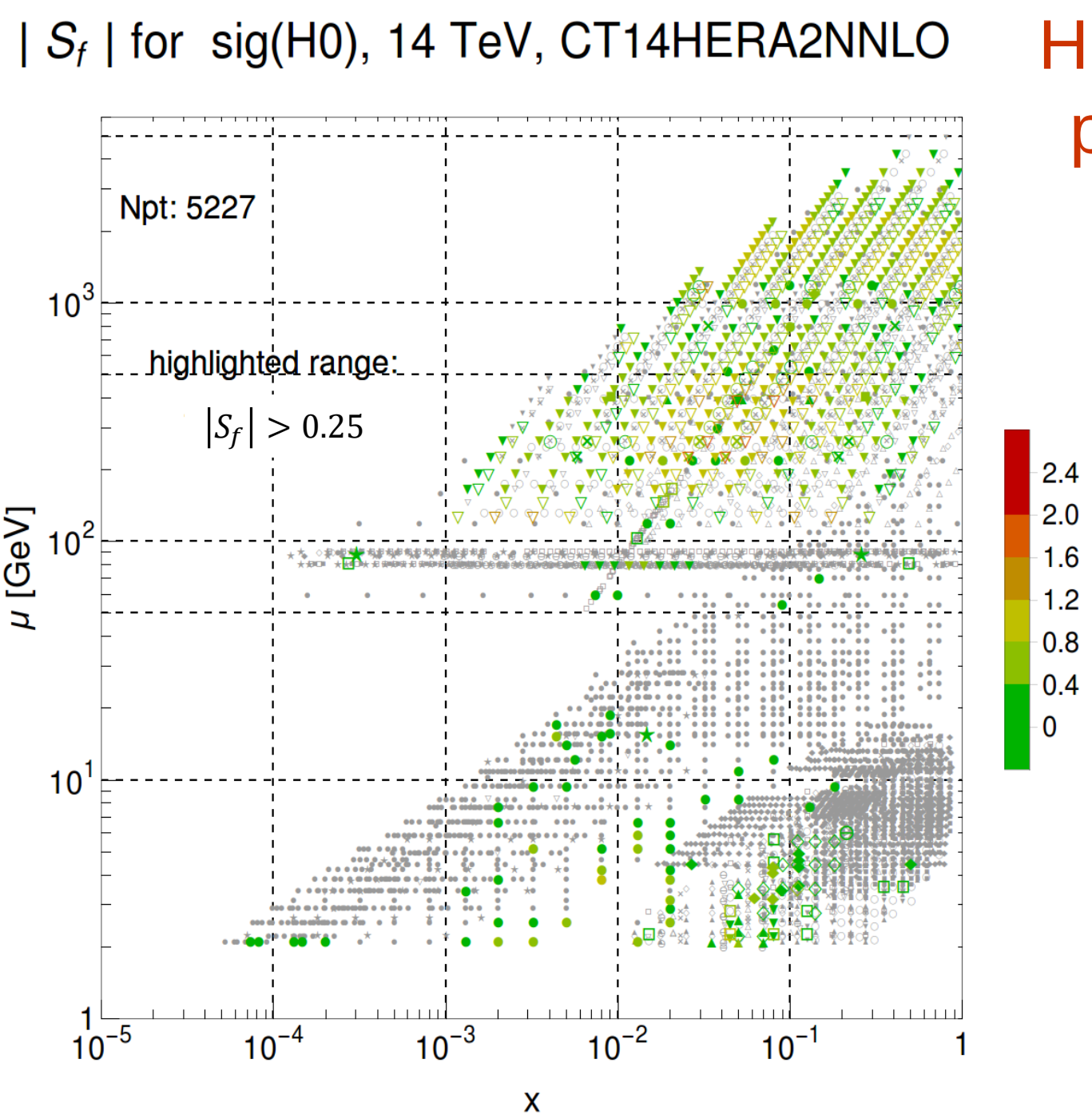
| S_f | for g(x,μ), CT14HERA2NNLO



$|S_f|$ for g(x,μ), CT14HERA2NNLO



Experiments with the highest sensitivity to $g(x_i, \mu_i)$

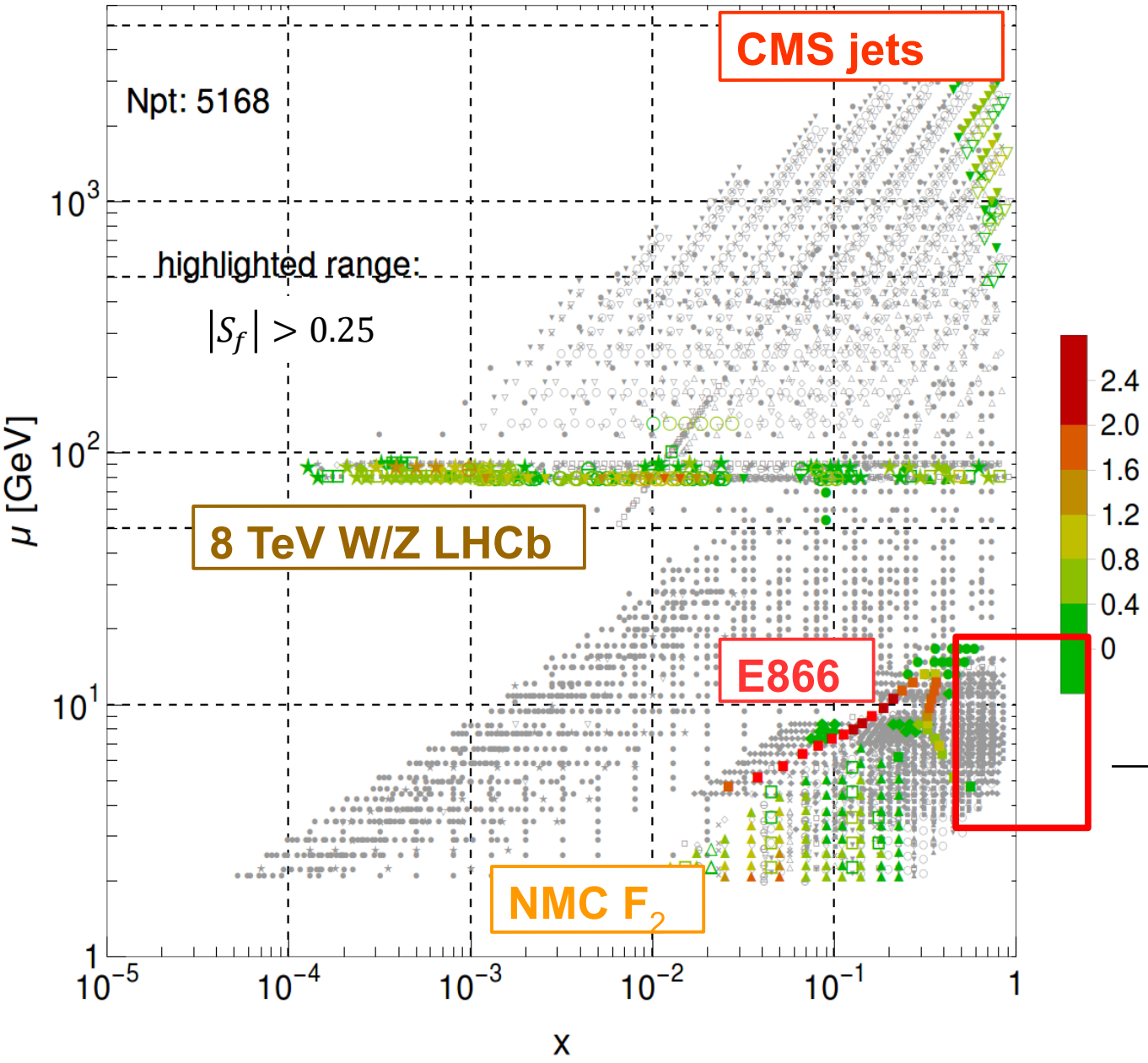


Higgs boson production

We find CMS/ATLAS **inclusive jet production (all bins)** and HERA DIS to have the dominant sensitivity!

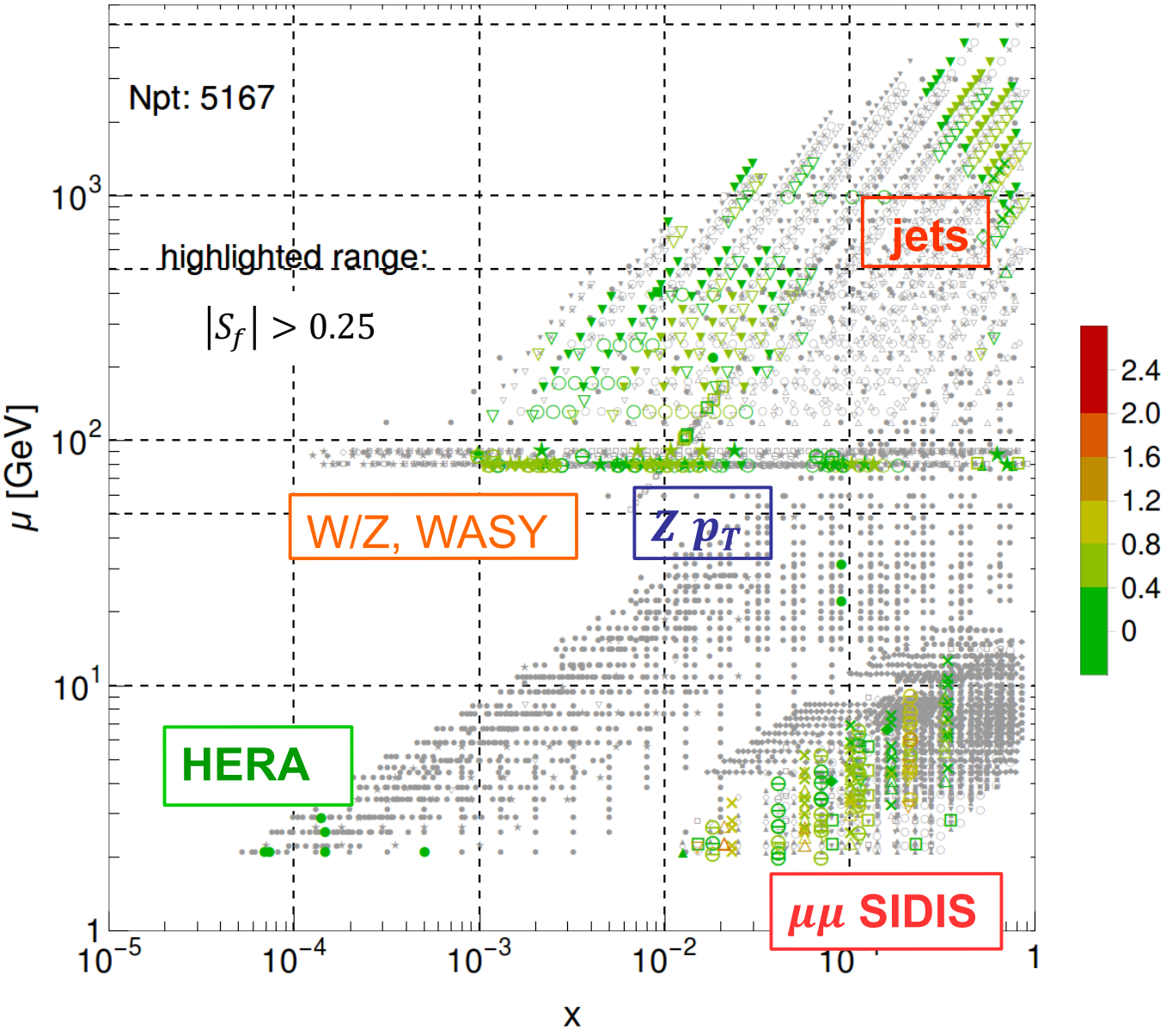
Good correlations with some points in E866, BCDMS, CCFR, CMS WASY, Z p_T and $t\bar{t}$ production; but fewer sensitive points in these processes

$|S_f|$ for $\bar{d}/\bar{u}(x,\mu)$, CT14HERA2NNLO



- PDF ratio is sensitive to flavor symmetry breaking in the light quark sea
- the large E866 pd/pp sensitivity *degrades* at larger x
- this is a prime motivation for higher x DY measurements at **E906** (SeaQuest)
- some contribution at high x from CMS inclusive jet production

$|S_f|$ for s(x,μ), CT14HERA2NNLO



- Constraints on $s(x,\mu)$ are weaker than on the other flavors

NuTeV, CCFR dimuon SIDIS most sensitive

- Sensitivities of vector boson production, jets are comparable

Ranking tables

No.	Exp. ID	N_{pt}	Rankings													
			$\sum_f \lvert S_f^E \rvert$	$\langle \sum_f \lvert S_f^E \rvert \rangle$	$\lvert S_{\bar{d}}^E \rvert$	$\langle \lvert S_{\bar{d}}^E \rvert \rangle$	$\lvert S_{\bar{u}}^E \rvert$	$\langle \lvert S_{\bar{u}}^E \rvert \rangle$	$\lvert S_g^E \rvert$	$\langle \lvert S_g^E \rvert \rangle$	$\lvert S_u^E \rvert$	$\langle \lvert S_u^E \rvert \rangle$	$\lvert S_d^E \rvert$	$\langle \lvert S_d^E \rvert \rangle$	$\lvert S_s^E \rvert$	$\langle \lvert S_s^E \rvert \rangle$
1	160	1120.	620.	**HERA**			**A**	3	**A**	3	**A**	3	B		C	
2	**545**	185	232.	**CMS jets 8**	3		C	3	B	2					C	3
3	111	86	218.	CCF3 F3p	**1**		C	**1**		3	B	**1**	C	2		
4	**542**	158	194.	**CMS jets 7**	3		C	3	B	2					C	3
5	101	337	184.	BCDMS F2p			C		C		B	3	C			
6	104	123	169.	NMC	2						C	2	B	2		
7	102	250	141.	BCDMS F2d					C	3	C	3	C	3		
8	109	96	115.	CDHSW	2		C	2		3	C	2	C	3		
9	201	119	113.	E605	2		C	2				3				

Experiments are listed in the descending order of the summed sensitivities to $\bar{d}, \bar{u}, g, u, d, s$

For each flavor, A and 1 indicate the strongest total sensitivity and strongest sensitivity per point

C and 3 indicate marginal sensitivities; low sensitivities are not shown

41	**247**	8	5.84	Z pT 7 TeV	3		3		3		3		3			
42	169	9	3.99	HERA F_L					2							
43	**567**	7	3.9	$t\bar{t}$					2							
44	227	11	3.7	CDF WASY (2005)									3			
45	**568**	5	3.4						2							
46	**566**	5	3.19	$t\bar{t}$					2							
47	145	10	1.14	HERA b												

Good per-point S_f , small N_{pt}

Results on **http://tinyurl.com/PDFSense**

- Normalized shifted residual vectors $\vec{\delta}_i$ from the CT14HERA2 NNLO analysis (.tsv files for visualization in The Embedding Projector, Mathematica, etc.)
- Inline Mathematica code to plot $\{x,\mu\}$ distributions of global data, residuals, sensitivities, correlations,…
- Sample plots of sensitivities to PDFs, Mellin moments, cross sections; projected impact of LHeC data; sensitivities to physical cross sections

arXiv:1803.02777 summarizes the key formulas (sensitivities, reciprocated distances,…); identifies the high-value experiments for the CT17 analysis (LHC jets, …); compares S_f values of various experiments to the PDFs and their combinations

Extra details

Experiments in the CT14 HERA2 fit

ID#	Experimental dataset		N_d
101	BCDMS F_2^p	[47]	337
102	BCDMS F_2^d	[48]	250
104	NMC F_2^d/F_2^p	[49]	123
108	CDHSW F_2^p	[50]	85
109	CDHSW F_3^p	[50]	96
110	CCFR F_2^p	[51]	69
111	CCFR xF_3^p	[52]	86
124	NuTeV $\nu\mu\mu$ SIDIS	[40]	38
125	NuTeV $\bar{\nu}\mu\mu$ SIDIS	[40]	33
126	CCFR $\nu\mu\mu$ SIDIS	[41]	40
127	CCFR $\bar{\nu}\mu\mu$ SIDIS	[41]	38
145	H1 σ_r^b (57.4 pb^{-1})	[53][54]	10
147	Combined HERA charm production (1.504 fb^{-1})	[39]	47
160	HERA1+2 Combined NC and CC DIS (1 fb^{-1})	[6]	1120
169	H1 F_L (121.6 pb^{-1})	[55]	9

ID#	Experimental dataset		N_d
201	E605 DY	[56]	119
203	E866 DY, $\sigma_{pd}/(2\sigma_{pp})$	[57]	15
204	E866 DY, $Q^3 d^2\sigma_{pp}/(dQdx_F)$	[58]	184
225	CDF Run-1 $A_e(\eta^e)$ (110 pb^{-1})	[59]	11
227	CDF Run-2 $A_e(\eta^e)$ (170 pb^{-1})	[60]	11
234	DØ Run-2 $A_\mu(\eta^\mu)$ (0.3 fb^{-1})	[61]	9
240	LHCb 7 TeV W/Z muon forward-η Xsec (35 pb^{-1})	[62]	14
241	LHCb 7 TeV W $A_\mu(\eta^\mu)$ (35 pb^{-1})	[62]	5
260	DØ Run-2 Z $d\sigma/dy_Z$ (0.4 fb^{-1})	[63]	28
266	CMS 7 TeV $A_\mu(\eta)$ (4.7 fb^{-1})	[64]	11
267	CMS 7 TeV $A_e(\eta)$ (0.840 fb^{-1})	[65]	11
268	ATLAS 7 TeV W/Z Xsec, $A_\mu(\eta)$ (35 pb^{-1})	[66]	41
281	DØ Run-2 $A_e(\eta)$ (9.7 fb^{-1})	[67]	13
504	CDF Run-2 incl. jet ($d^2\sigma/dp_T^j dy_j$) (1.13 fb^{-1})	[36]	72
514	DØ Run-2 incl. jet ($d^2\sigma/dp_T^j dy_j$) (0.7 fb^{-1})	[37]	110
535	ATLAS 7 TeV incl. jet ($d^2\sigma/dp_T^j dy_j$) (35 pb^{-1})	[68]	90
538	CMS 7 TeV incl. jet ($d^2\sigma/dp_T^j dy_j$) (5 fb^{-1})	[69]	133

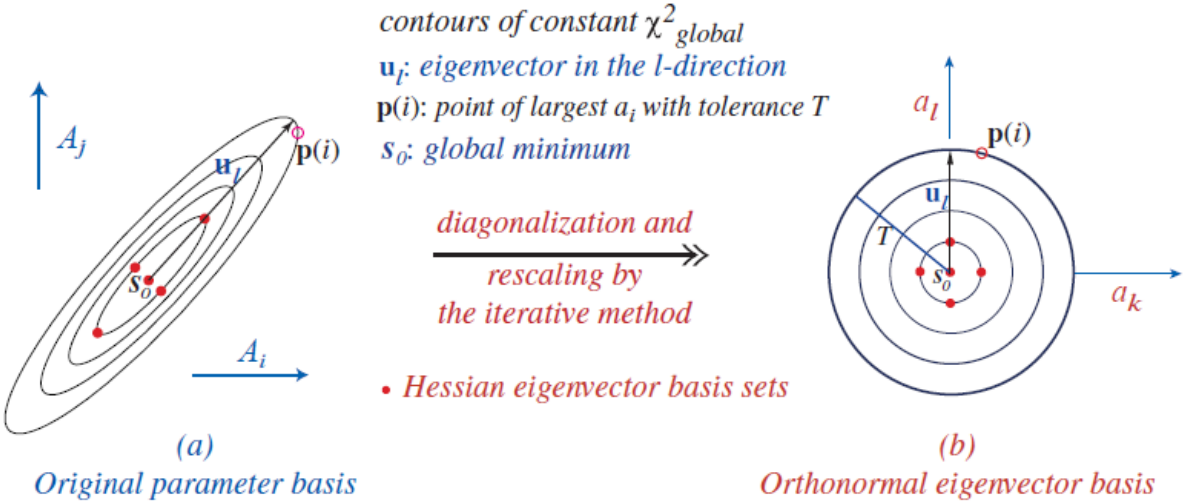
Candidate experiments in the CTEQ-TEA fit

ID#	Experimental dataset		N_d
245	LHCb 7 TeV Z/W muon forward-η Xsec (1.0 fb^{-1})	[70]	33
246	LHCb 8 TeV Z electron forward-η $d\sigma/dy_Z$ (2.0 fb^{-1})	[71]	17
247	ATLAS 7 TeV $d\sigma/dp_T^Z$ (4.7 fb^{-1})	[72]	8
249	CMS 8 TeV W muon, Xsec, $A_\mu(\eta^\mu)$ (18.8 fb^{-1})	[73]	33
250	LHCb 8 TeV W/Z muon, Xsec, $A_\mu(\eta^\mu)$ (2.0 fb^{-1})	[74]	42
252	ATLAS 8 TeV Z ($d^2\sigma/d|y|_{ll}dm_{ll}$) (20.3 fb^{-1})	[75]	48
253	ATLAS 8 TeV ($d^2\sigma/dp_T^{ll}dm_{ll}$) (20.3 fb^{-1})	[76]	45
542	CMS 7 TeV incl. jet, R=0.7, ($d^2\sigma/dp_T^j dy_j$) (5 fb^{-1})	[34]	158
544	ATLAS 7 TeV incl. jet, R=0.6, ($d^2\sigma/dp_T^j dy_j$) (4.5 fb^{-1})	[33]	140
545	CMS 8 TeV incl. jet, R=0.7, ($d^2\sigma/dp_T^j dy_j$) (19.7 fb^{-1})	[35]	185
565	ATLAS 8 TeV $t\bar{t}$ $d\sigma/dp_T^t$ (20.3 fb^{-1})	[38]	8
566	ATLAS 8 TeV $t\bar{t}$ $d\sigma/dy_{<t/\bar{t}>}$ (20.3 fb^{-1})	[38]	5
567	ATLAS 8 TeV $t\bar{t}$ $d\sigma/dm_{t\bar{t}}$ (20.3 fb^{-1})	[38]	7
568	ATLAS 8 TeV $t\bar{t}$ $d\sigma/dy_{t\bar{t}}$ (20.3 fb^{-1})	[38]	5

N_d is the number of data points

Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (26) PDF parameter space

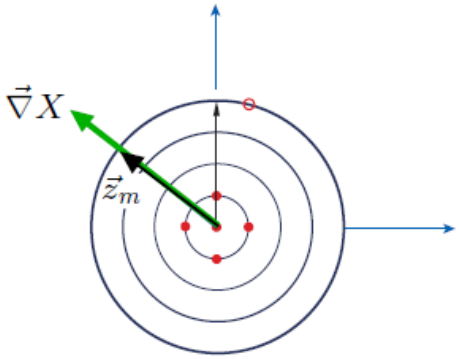


A hyperellipse $\Delta\chi^2 \leq T^2$ in space of N physical PDF parameters $\{\mathrm{A}_i\}$ is mapped onto a filled hypersphere of radius T in space of N orthonormal PDF parameters $\{\mathrm{a}_i\}$

Hessian method: Pumplin et al., 2001

Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (26) PDF parameter space



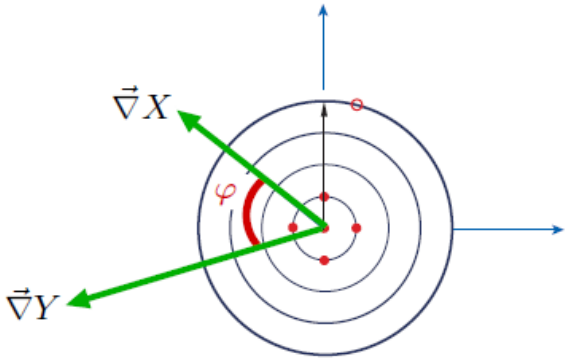
(b)
Orthonormal eigenvector basis

A symmetric PDF error for a physical observable X is given by

$$\Delta X = \vec{\nabla} X \cdot \vec{z}_m = \left|\vec{\nabla} X\right| = \frac{1}{2}\sqrt{\sum_{i=1}^{N}\left(X_i^{(+)} - X_i^{(-)}\right)^2}$$

Tolerance hypersphere in the PDF space

2-dim (i,j) rendition of N-dim (26) PDF parameter space



(b)
Orthonormal eigenvector basis

Correlation cosine for observables X and Y:

$$\cos\varphi = \frac{\vec{\nabla}X \cdot \vec{\nabla}Y}{\Delta X \Delta Y} = \frac{1}{4\Delta X\, \Delta Y} \sum_{i=1}^{N} \left(X_i^{(+)} - X_i^{(-)}\right)\left(Y_i^{(+)} - Y_i^{(-)}\right)$$

$cos\varphi \equiv Corr[X,Y]$ -- realization of the Pearson correlation coefficient in the Hessian method

Finding shifted residuals r_i from the covariance matrix

The CTEQ-TEA fit returns tables of $r_i(\vec{a})$ and $\bar{\lambda}_\alpha(\vec{a})$ for every i and α

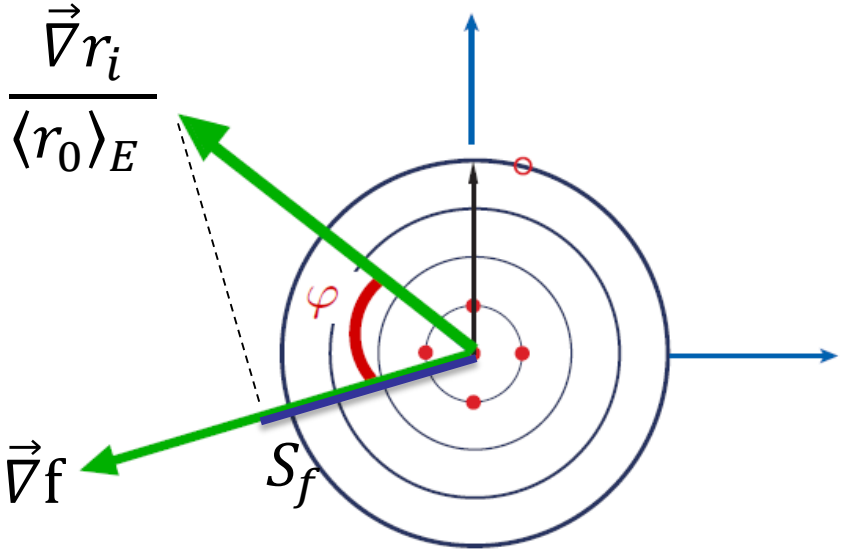
Alternatively, they can be found from the covariance matrix:

$$r_i(\vec{a}) = s_i \sum_{j=1}^{N_{pt}} (\mathrm{cov}^{-1})_{ij} \, (T_j(\vec{a}) - D_j) \,, \qquad \overline{\lambda}_\alpha(\vec{a}) = \sum_{i,j=1}^{N_{pt}} (\mathrm{cov}^{-1})_{ij} \frac{\beta_{i\alpha}}{s_i} \frac{(T_j(\vec{a}) - D_j)}{s_j}$$

The underlying geometrical picture

S_f is a projection of a residual gradient $\vec{\nabla} r_i$ onto the direction for $\vec{\nabla} f$.

S_f captures a small part of information about the 28- (or 56)-dimensional population of $\vec{\nabla} r_i$ for 4000 data points

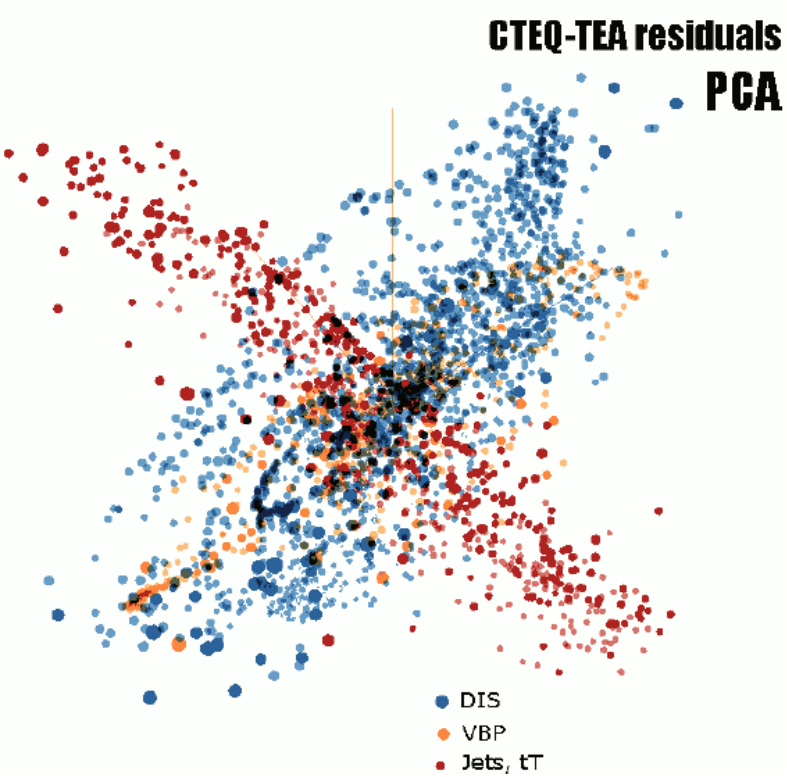


PDFSense can export the normalized $\vec{\nabla} r_i$ vectors (specifically, $(r_{i,L} - r_{i,0})/\langle r_0 \rangle_E$ for $\mathrm{i} = 1,4000$; $L = 1,56$) in the .TSV format to analyze the whole population of $\vec{\nabla} r_i$ using machine learning tools

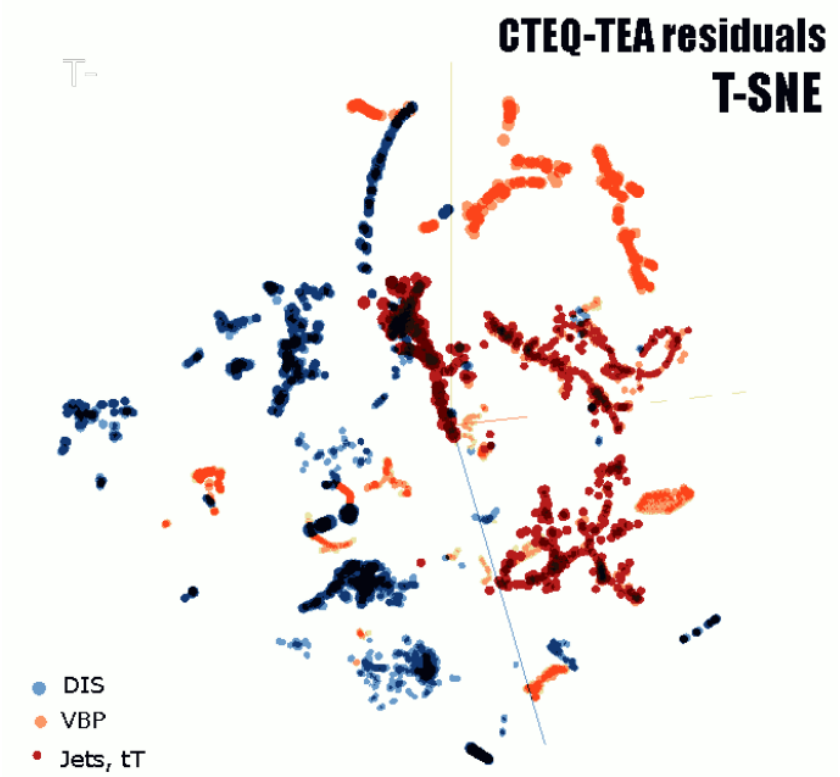
TensorFlow Embedding Projector

http://projector.tensorflow.org

Reads 2 .tsv files with $\vec{\nabla} r_i / \langle r_0 \rangle_E$ vectors and metadata (descriptions of data points)

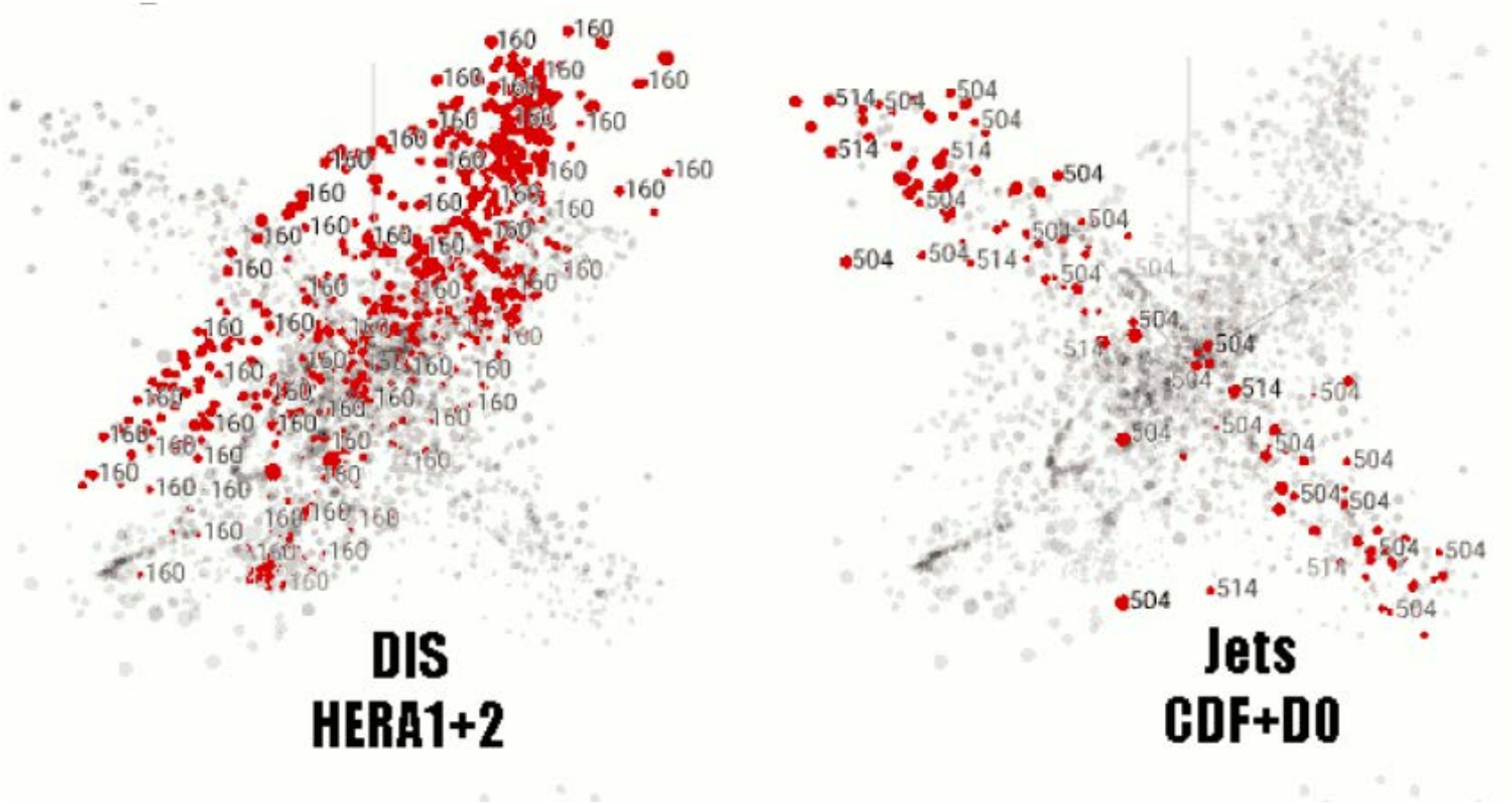


Principal Component Analysis (PCA) visualizes the 56-dim. manifold by reducing it to 10 dimensions (à la META PDFs)

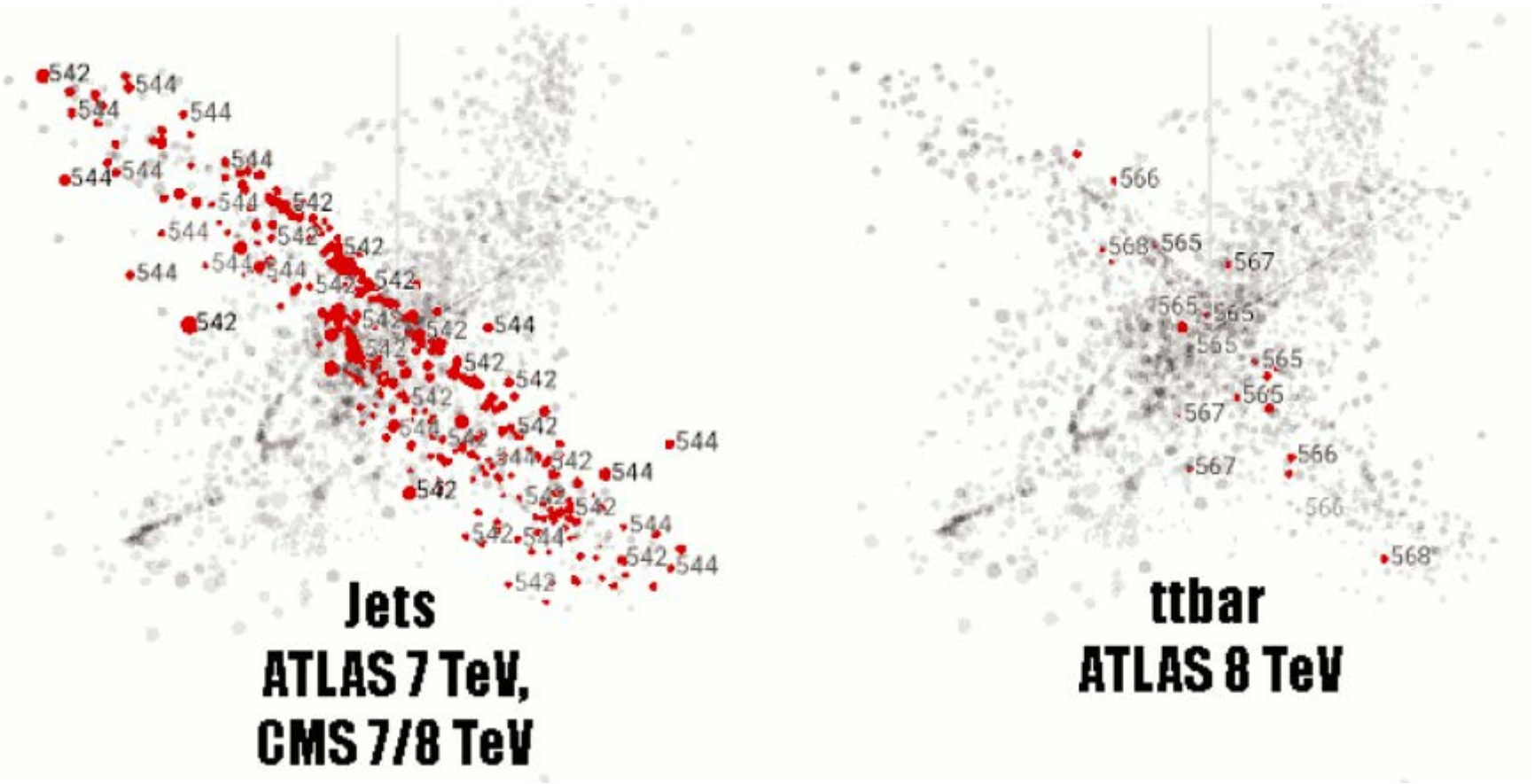


t-distributed stochastic neighbor embedding (**t-SNE**) sorts $\vec{\nabla} r_i / \langle r_0 \rangle_E$ vectors according to their similarity

CTEQ-TEA residuals PCA



CTEQ-TEA residuals PCA



CTEQ-TEA residuals PCA