

# Topological materials: from a catalogue to machine learning

N. Regnault

$\frac{1}{\sqrt{2}}$  (Ecole Normale Supérieure Paris and CNRS) +  $\frac{1}{\sqrt{2}}$  (Princeton University)

*Symposium on Condensed Matter Physics*  
KITP - November 2019



# Acknowledgements

- Maia Vergniory (Donostia International Physics Center)
- Luis Elcoro (University of the Basque Country)
- Benjamin Wieder (Princeton University)
- Zhijun Wang (Beijing National Laboratory for Condensed Matter Physics and IOP)
- Claudia Felser (MPI for Chemical Physics of Solids)
- Andrei Bernevig (Princeton University)
- Nikolas Claussen (ENS Paris → UCSB)

## References:

- M. Vergniory et al. Nature 556 (2019).
- N. Claussen et al., arXiv:1910.10161.

## A Catalogue of Topological Materials

# How to get the band structure?

A space group (SG) is a set of symmetries that defines a crystal structure in 3D.

- Unit lattice translations ( $\mathbb{Z}_3$ ).
- Point group operations (rotations, reflections).
- Non-symmorphic (screw, glide)

## Ingredients:

- one of the 230 SGs.
- Atoms at some lattice positions.
- Orbitals (s, p, d, ...).

How do we go from real space orbitals sitting at lattice sites to any electronic band structure (without a Hamiltonian)?

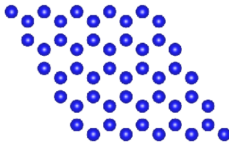
Elementary band representation (EBR)

# One SG, many options

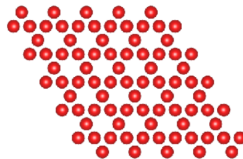
Several ways to arrange the atoms within the unit cell where all atoms are related by symmetry



1 atom/unit cell  
(triangular)



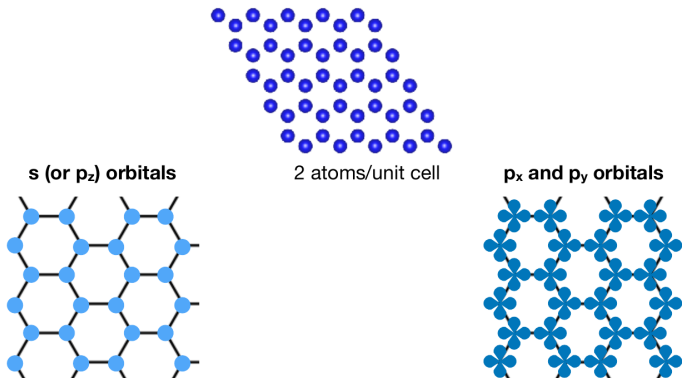
2 atoms/unit cell  
(honeycomb)



3 atoms/unit cell  
(kagome)

# One SG, many options

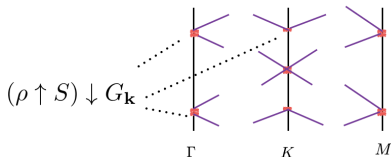
Several ways to possible choice for the electronic orbitals for a given arrangement (orbital arrangement must be consistent with SG symmetry)



# Elementary band representation (EBR)

Can we build all the possible band structures for these cases (i.e. all the atomic limits)?

- **Key Insight:** Think of bands as representations! (Zak, Bacry, Michelle).
- Then ask questions of representation reducibility (Elementary band)
- Find all the irreps  $\rightarrow$  EBR.
- Single/double group, w/wo time reversal and each rep, Wyckoff: 10398 such irreps, tabulated on the Bilbao Crystallographic server.
- The band representation also gives the Brillouin zone irreps at points and lines.
- By construction, a band representation has an atomic limit, and all atomic limits yield a band representation.



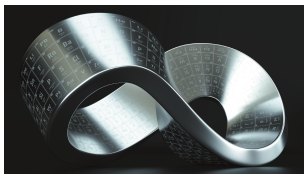
# Topological Quantum Chemistry (TQC)

TQC provides a unified framework for the treatment of *all topological phases arising from crystalline symmetries*.

It relies on:

- EBRs which enumerate a basis for all electronic bands induced from atomic orbital (atomic limits)
- compatibility relations, constraining how bands can connect across the Brillouin zone.

*“All sets of bands not induced from symmetric, localized orbitals, are topologically non-trivial by design.”*





# Trivial and Topological Insulators (TIs)

Topological “insulating” classes  $\left\{ \begin{array}{l} \text{EBR1} \\ \text{EBR2} \end{array} \right.$

$$\mathbf{b} = \sum_i n_i(\mathbf{k}) = \sum_m a_m \times \text{EBR}$$

LCEBR

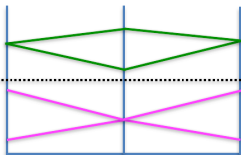
SEBR

NLC

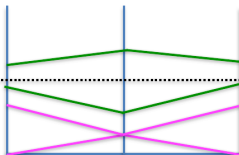
Linear Combination EBR

Split EBR

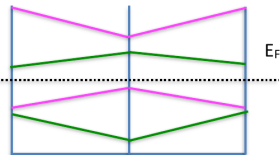
Non Linear Combination



$$\mathbf{b} = \text{EBR1}$$



$$\mathbf{b} = \text{EBR1} + 1/2 \times \text{EBR2}$$



$$\mathbf{b} \neq \text{LCEBR}$$

\* Fragile:  $\text{EBR}_F = \text{EBR1} - \text{EBR2}$

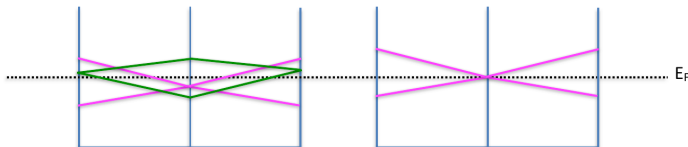
$\text{EBR}_F + \text{EBR2} = \text{EBR1}$ : trivial

# Topological Semi-Metals (TSM)

Topological “metallic” classes {  $\text{EBR1}$   
 $\text{EBR2}$

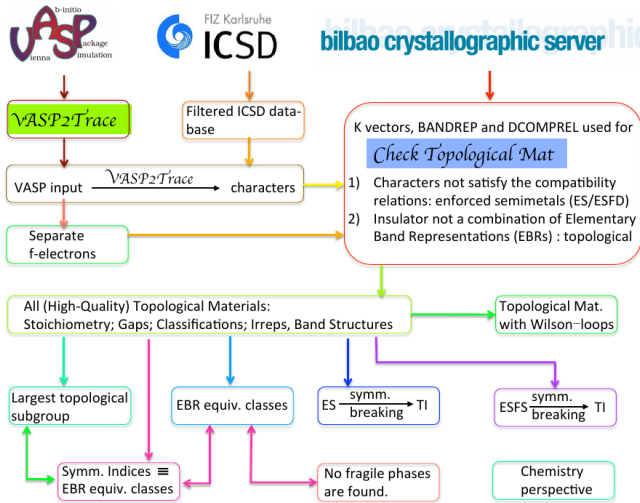
ES  
Enforced semimetals

ESFD  
ES Fermi Degeneracy



- *ESFD*: high-symmetry point degeneracy at the Fermi level.
- *ES*: the degeneracy is away from the high-symmetry points.

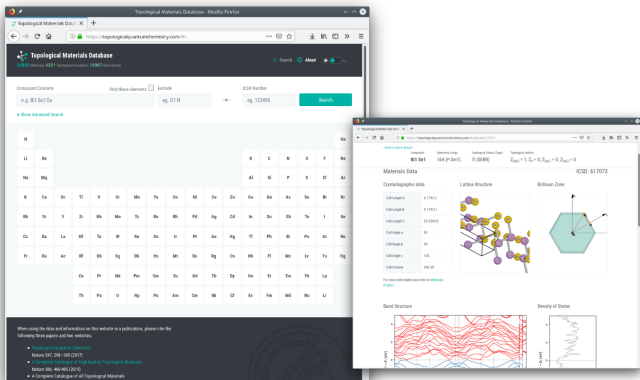
# The material search procedure



A massive search: more than 70k valid ICSDs, 20M CPU hours.  
Topology is not rare! Trivial (47%) TIs (16%) TSMs (37%)

# The Topological Material Database

[www.topologicalquantumchemistry.com](http://www.topologicalquantumchemistry.com)




Links from and to materialsproject.org

Vergniory et al. Nature 556 (2019)

**See also** T. Zang et al. Nature 566 (2019), [materiae.iphy.ac.cn](http://materiae.iphy.ac.cn) and F. Tang et al. Nature 566 (2019)

# Sneak peek of the new website

 **Topological Materials Database**

Total Materials

38423

Topological Insulators

6171

Semi-Metals

14111

NAVIGATION

Search

Predict

About

Wiki

SETTINGS

UI Mode

Compound Contains

Only these elements ☐ Exclude

ICSD Number

Bi Te

eg. 01 N

- or -

eg. 123456

Search


▼ Show Advanced Search

16 TIs, 9 SMs and 30 Trivialities found

1																	2				
H																	He				
3	4															5	6	7	8	9	10
Li	Be															B	C	N	O	F	Ne
11	12															13	14	15	16	17	18
Na	Mg															Al	Si	P	S	Cl	Ar
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36				
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr				
37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54				
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe				
55	56	57	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86				
Cs	Ba	Ac	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn				
87	88	89	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118				
Fr	Ra		Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og				
			58	59	60	61	62	63	64	65	66	67	68	69	70	71					
			Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu					
			90	91	92	93	94	95	96	97	98	99	100	101	102	103					
			Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr					

When using the data and information on this website in a publication, please cite the following three papers and two websites:

- [Topological Quantum Chemistry](#)  
Nature 547, 298–305 (2017)
- [A Complete Catalogue of High-Quality Topological Materials](#)  
Nature 566, 480-485 (2019)
- [A Complete Catalogue of All Topological Materials](#)  
In Preparation
- [Topological Material Database](#)
- [Bilbao Crystallographic Server](#)



Imprint

Privacy Policy

# Sneak peek of the new website

The screenshot shows the Topological Materials Database website. On the left is a dark sidebar with navigation links: Search, Predict, About, and Wiki. Below these are settings for UI Mode. The main content area features a search bar with the text 'Compound Contains' and a filter for 'Only these elements' (eg. Ti N) or 'Exclude'. A search button is present. Below the search bar, a periodic table is displayed with elements colored by category. A search result for 'Ti' is highlighted in green. At the bottom, there is a disclaimer about citing the website in publications and a list of references.

Topological Materials Database

Total Materials: 34423  
Topological Insulators: 6171  
Semimetals: 14311

NAVIGATION

- Search
- Predict
- About
- Wiki

SETTINGS

UI Mode

Compound Contains: [Ti N] Only these elements ☐ Exclude ☒ eg. O1 N -or- ICSD Number: eg. 123456 Search

16 Ti's, 9 Si's and 31 Ti's found

When using the data and information on this website in a publication, please cite the following three papers and two websites:

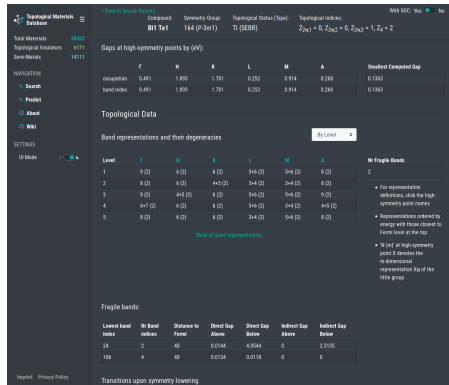
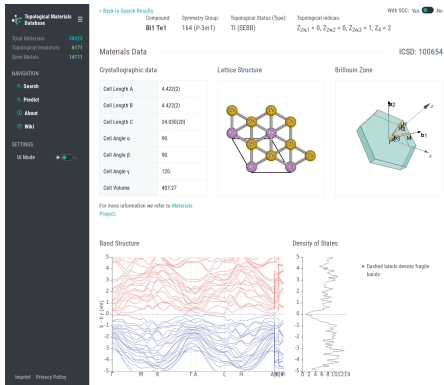
- Topological Quantum Chemistry  
Nature 547, 298–305 (2017)
- A Complete Catalogue of High-Quality Topological Materials  
Nature 566, 480–485 (2019)
- A Complete Catalogue of All Topological Materials  
In Preparation
- Topological Material Database
- Bilbao Crystallographic Server

Imprint Privacy Policy

- Improved U.I..
- *More than 38k unique materials* (71k unique ICSDs).
- With and without SOC.
- Dynamical plots.
- Fragile phases.
- Wiki and more.



# Sneak peek of the new website



## Machine Learning and Topological Materials

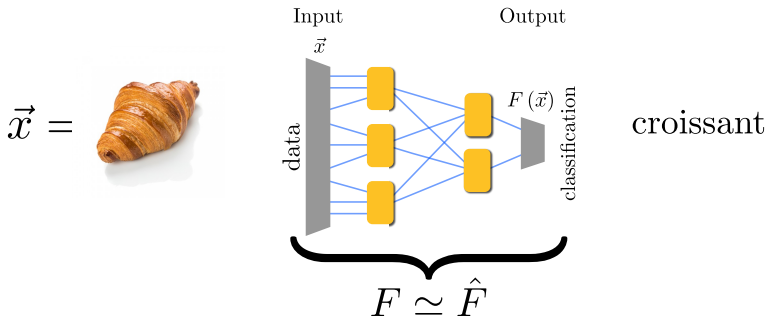


# Machine Learning: what it does

The only useful "French" mathematical function:  $\hat{F}$

$$\hat{F} \left[ \begin{array}{c} \text{image of croissants and coffee} \end{array} \right] = \text{croissant} \quad \hat{F} \left[ \begin{array}{c} \text{image of pain au chocolat} \end{array} \right] = \text{pain au chocolat}$$

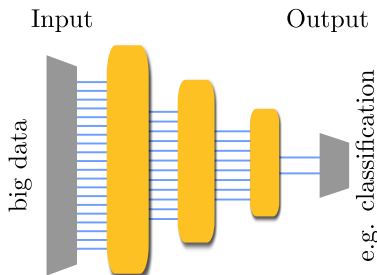
In a physicist language : find a variational ansatz  $F$  capturing  $\hat{F}$



# Machine Learning : how to train your network

## Supervised learning:

- Train the network with a large amount of labeled data (input-output pairs): Reduce a cost function (distance measure between network output and labels) via e.g. gradient descent.
- Verify the network performance on a distinct test data set.



## Unsupervised learning:

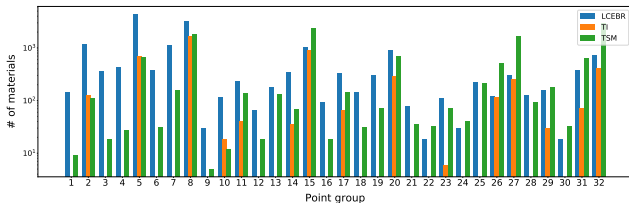
Use unlabeled data, the network learns to cluster data/find structure/learn probability distribution of features.

## Reinforced learning:

Agents, reward : direct the action of software agents in an environment to maximize some cumulative reward (e.g. videogame score).

# A large database? Overfitting?

- **Crystallographic data:** chemical elements, symmetry group, atom positions.
- Large wrt to chemistry (35k unique materials), small for machine learning.
- Unbalanced samples: Trivial (47%) vs TIs (16%) or TSMs (37%)

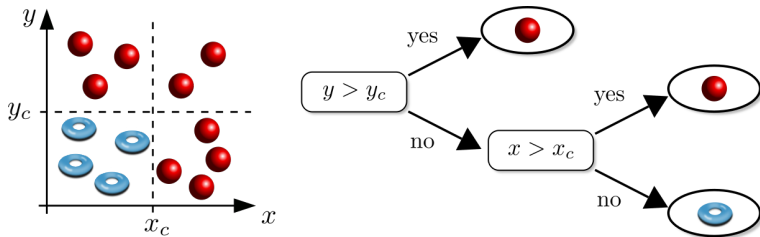


- Lots of information per compound (*curse of high dimension*).
- Not easy to encode: how to encode chemical structure, (basis-free) atom positions?

Using a NN architecture: more (parameters) is not better, how to prepare the data?

# Which network architecture?

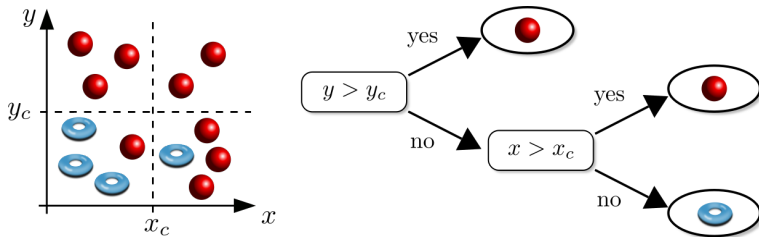
- Pick the right architecture depending on the data structure.
- Unstructured data (i.e., not pictures): decision tree classifiers



- Improved version: gradient boosted trees (GBT) - training an ensemble of simple (weak) decision trees instead of unique but complex one (to avoid overfitting).
- Two libraries: *sklearn* and *xgboost*.

# Which network architecture?

- Pick the right architecture depending on the data structure.
- Unstructured data (i.e., not pictures): decision tree classifiers



- Improved version: gradient boosted trees (GBT) - training an ensemble of simple (weak) decision trees instead of unique but complex one (to avoid overfitting).
- Two libraries: *sklearn* and *xgboost*.

# Our situation

- **Input:**

- Nbr of electrons: encoded in binary (easy to detect parity effects, like ESFDs in some specific SGs).
- Symmetry group: number + frequencies of each class (fingerprint of each SG)
- Chemical structure: mean number of  $s, p, d, f$  valence electrons, number of atoms per column/row in the periodic table (to encode chemical similarity).
- (optional): atom position encoded as average and variance of distances between atoms and their nearest neighbors, coulomb matrix, ...

- **Output:**

- *Coarse grained label*: Trivial/TI/TSM.
- Full label: Trivial/NLC/SEBR/ES/ESFD.

- **Training and testing:**

- 32k materials for the training, 2.5k for the testing.
- Cross-validation (to estimate the error).

## Results: coarse grained label

Model	$d$	Acc. [%]	$F_1$ Triv. [%]	$F_1$ TI [%]	$F_1$ TSM [%]
Full model (FM)	49	89.7(5)	94.0(3)	70(1)	92.0(5)
FM + Non-SOC	50	92.0(3)	96.5(2)	77(1)	93.3(4)
Baseline model	94	86.0(5)	92.5(5)	67(1)	91.0(5)
<i>spdf</i> + model	10	87.7(5)	93.0(5)	69(1)	92.0(5)
FM + nearest-neighbor	184	89.0(5)	94.0(3)	69(3)	92.0(5)
FM without SG	48	84.0(5)	91.5(3)	57(2)	86(1)

- $d$ : size of the input vector.
- Full model: SG,  $N_e$ , *spdf*+, number of atoms from each periodic table row and column.
- Baseline model: SG,  $N_e$ , baseline descriptor (nbr of atoms from each element in stoichiometric formula).

# Results: coarse grained label

$F_1$ -score: choose your poison

- **Precision:** reliability of a binary classifier's positive predictions (i.e. how many *False Positive*).
- **Recall:** Ability to find all the true positive sample points (i.e. how many *False Negative*).

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- You can always trade Precision for Recall (and vice versa).
- Much better than a random classifier only using the statistics per SG ( $F_1$ -score around 1.2% vs 70% for TIs).
- Given the SG, the positions of the atoms within the crystal lattice are of limited importance for the material's topology. Rather, **it is the “average orbital character”** (*spdf*+).

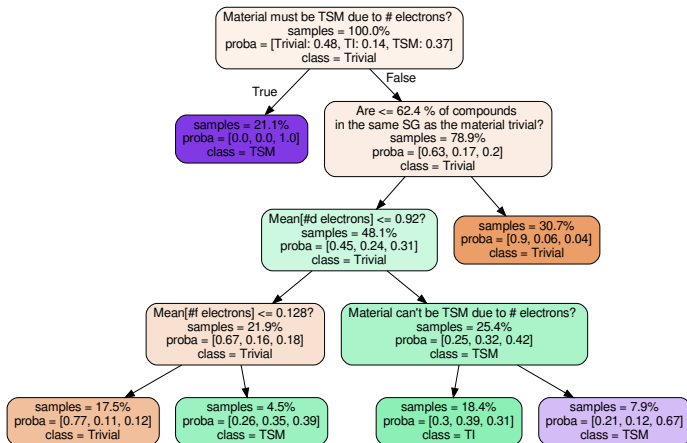


## Results: full label

Acc. [%]	$F_1$ Triv. [%]	$F_1$ NLC [%]	$F_1$ SEBR [%]	$F_1$ ES [%]	$F_1$ ESFD [%]
87.0(3)	94.0(4)	66(2)	59(3)	73(2)	95.5(3)
89.7(5)	94.0(3)	70(1)		92.0(5)	

- The sample size per label starts being small.
- Swapping e.g. NLC and SEBR for one material would only lead to a wrong prediction in the full label classification.
- ES are harder to detect than ESFD: this latter can in many cases be detected from  $N_e$  and the SG alone.
- Our model correctly identifies the groups in which TIs are allowed according to TQC.
- NLCs are easier to predict than SEBRs. SEBR type depends on energetics, not fully captured by our model.


# A simplified version



Overall accuracy (70%) does not match the GBT one but it already capture some of the important classification rules and relevant features.

# Test your own material

<https://www.topologicalquantumchemistry.com/mltqc>

 Topological Materials Database

NAVIGATION

[Search](#)


[About](#)

[ML](#)


## Detection of Topological Materials with Machine Learning

This online tool predicts the topological classification of materials. It is based on gradient boosted trees trained with the ab-initio results from the [topological quantum chemistry database](#). A full description of this method is available in [arxiv:1910.10161](#).

Provide your material information

1. Upload your VASP input file (POSCAR) 

No file selected.

2. Or provide the chemical composition of the primitive unit cell 

3. Choose your symmetry group:

4. Add the topological classification without spin-orbit coupling:

When using the information on this website in a publication, please cite the following three papers:

[Topological Quantum Chemistry](#)  
Nature 298, 547-555 (2017)

[A Complete Catalogue of High-Quality Topological Materials](#)  
Nature 566, 480-485 (2019)

[Detection of Topological Materials with Machine Learning](#)  
arxiv:1910.10161



