

# Do Structurally Similar InChIs Have Similar Hash Keys?

Dac-Trung Nguyen

National Center for Advancing Translational Sciences  
National Institutes of Health

InChI Virtual Workshop  
March 22–24, 2021

*InChI is the greatest thing ever since sliced bread.*

—Steve Heller

*InChIKey is perhaps the greatest thing ever since butter.*

—Dac-Trung Nguyen

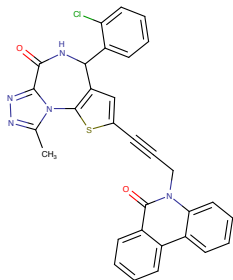
InChIKey is a hash key with...

- (i) Compact representation (27 characters)
- (ii) Structural “hints” (UHFFFAOYSA and charge suffix -N)
  - ▶ Approximately 74% of structures in PubChem contains UHFFFAOYSA
- (iii) Collision resistance (truncated SHA-2 with very low probability of collision)

InChIKey is well-suited for applications that require uniqueness (e.g., resolver). However, for many use cases such as registration and HTS analysis, we would like a more flexible hash key that can facilitate “meaningful” comparison while retaining relevant features of InChIKey. This is the story of *spectral hash key* for InChI.

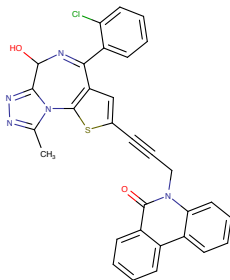
# Spectral hash key at a glance

CHEMBL279476



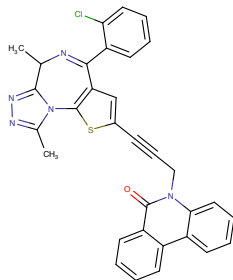
PVSI<sup>C</sup>LYUDGZPCO-UHFFFAOYSA-N  
111ZMSDKX4LX8LNL6ZR<sup>D</sup>728VM5HM7J

CHEMBL282495



GTTQCZ<sup>X</sup>VNFENV-UHFFFAOYSA-N  
111ZMSDKX4LX8LNL6ZR<sup>U</sup>L5DJK5J2XW

CHEMBL20178



ZIJDJRBEHTWHFS-UHFFFAOYSA-N  
111ZMSDKXW278SN92KHW12FPL5V1YL

- ▶ Spectral graph theory
- ▶ Graph spectrum
  - ▶ Adjacency
  - ▶ Laplacian
  - ▶ Normalized Laplacian
- ▶ Spectral properties
- ▶ Spectral hash key
- ▶ Do similar hash keys have similar biological activity?
- ▶ What is “similarity”?
- ▶ Code availability & Acknowledgements

- ▶ Graph  $G$  consists of a set of  $n$  vertices  $V = \{v_1, v_2, \dots, v_n\}$  and  $m$  edges  $E = \{e_1, e_2, \dots, e_m\}$  where  $e_k = v_i \sim v_j$
- ▶ Let  $M$  be a matrix that encodes  $G$  based on  $V$ ,  $E$ , or combinations thereof
- ▶ Spectral graph theory is about understanding the properties of  $G$  in terms of eigenvalues and eigenvectors of  $M$ , i.e.,

$$M\mathbf{v}_i = \lambda_i\mathbf{v}_i,$$

where  $\lambda_i$  is the  $i$ th eigenvalue and  $\mathbf{v}_i$  is the corresponding eigenvector.

- ▶ The eigenvalues  $\{\lambda_i\}$  define the *spectrum* of  $G$
- ▶ Outstanding problem: Which graphs are determined by their spectrum?
  - ▶ Under what conditions do non-isomorphic graphs have the same spectrum?

# Graph spectrum

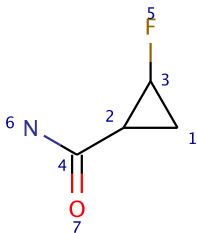
## Adjacency

The adjacency  $A$  representation of  $G$  is defined as

$$A_{ij} = \begin{cases} 1 & \text{if } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

## Foundation of Hückel theory

*The topology of a molecule, rather than its geometry, determines the form of the Hückel molecular orbitals.*



$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

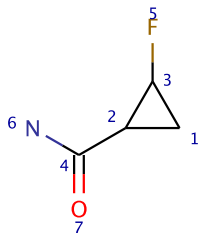
# Graph spectrum

## Laplacian

Let  $D$  be the degree matrix of  $G$ , i.e.,  $D_{ii} = \text{degree}(v_i)$  and 0 elsewhere, we have the Laplacian  $L$  defined as follows

$$L = D - A,$$

where  $A$  is the adjacency matrix.



$$L = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 3 & 0 & -1 & -1 \\ 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

# Graph spectrum

## Normalized Laplacian

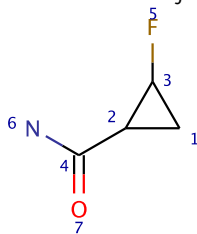
The normalized Laplacian is defined as

$$\tilde{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}},$$

or

$$\tilde{L}_{ij} = \begin{cases} 1 & i = j \\ -\frac{1}{\sqrt{d_i d_j}} & i \neq j \end{cases}$$

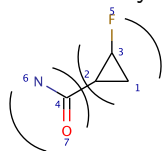
where  $d_i$  and  $d_j$  are the degrees of  $v_i$  and  $v_j$ , respectively.



$$\tilde{L} = \begin{pmatrix} 1.0 & -0.4 & -0.4 & 0.0 & 0.0 & 0.0 & 0.0 \\ -0.4 & 1.0 & -0.3 & -0.3 & 0.0 & 0.0 & 0.0 \\ -0.4 & -0.3 & 1.0 & 0.0 & -0.6 & 0.0 & 0.0 \\ 0.0 & -0.3 & 0.0 & 1.0 & 0.0 & -0.6 & -0.6 \\ 0.0 & 0.0 & -0.6 & 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.6 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & -0.6 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$



- ▶ The spectrum of  $A$  is bounded by the maximum degree in  $G$ , i.e.,  $|\lambda_i| \leq \max_k d(v_k)$  for  $k = 1, 2, \dots, n$ . For organic molecules,  $|\lambda_i| \leq 4$ .
- ▶  $L$  and  $\tilde{L}$ 's spectra are non-negative, i.e.,  $\lambda_i \geq 0$ .  $L$  and  $\tilde{L}$  are semidefinite.
- ▶ Multiplicity of  $\lambda_i = 0$  in  $L$  and  $\tilde{L}$  is the number of connected components in  $G$ .
- ▶ The spectrum of  $\tilde{L}$  is bounded by 2, i.e.,  $0 \leq \lambda_i \leq 2$ .
- ▶ Let  $\lambda_1 = 0 \leq \lambda_2 \leq \dots \leq \lambda_n$  for  $L$  and  $\tilde{L}$ . The first non-zero  $\lambda_i$  is the *algebraic connectivity* index with the corresponding eigenvector known as the *Fiedler* vector. This vector provides near-optimal 2-partition of  $G$ . The Fiedler vector is the foundation of many spectral clustering algorithms.



$$\mathbf{v}_2(\tilde{L}) = \begin{bmatrix} 0.29255 \\ 0.11507 \\ 0.44483 \\ -0.53342 \\ 0.32870 \\ -0.39416 \\ -0.39416 \end{bmatrix}$$

$|h_1|=9$   $|h_2|=10$   $|h_3|=11$   
111ZMSDKX 4LX8LNL6ZR D728VM5HM7J

## Algorithm

- (i) Let  $\{\lambda_i\}$  be the spectrum of  $\tilde{L}$  (largest component)
- (ii)  $h_1$  is the truncated (45 bits) SHA-1 digest of  $\{\lambda_i\}$
- (iii)  $h_2$  is the truncated (50 bits) SHA-1 digest of  $h_1$  and /c layer of the largest component
- (iv)  $h_3$  is the truncated (55 bits) SHA-1 digest of  $h_2$  and full InChI string
- (v) Spectral hash key is an 150-bit string  $h_1h_2h_3$

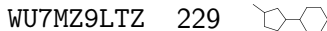
## Properties

- ▶ Hash key is a base-32 encoded string with the alphabet  $\{A, \dots, Z\} \cup \{1, \dots, 9\} \setminus \{E, I, O\}$
- ▶ Three logical blocks  $h_1$ ,  $h_2$ , and  $h_3$  with progressively increased resolution
- ▶ Hash chaining allows the individual blocks to be used independently
- ▶ Structure grouping with sort

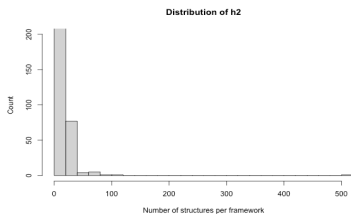
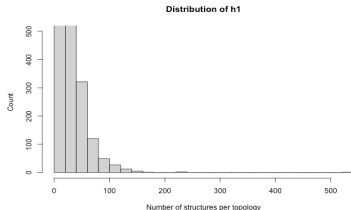
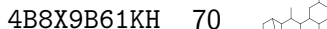
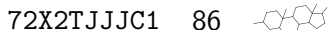
# Spectral hash key by the numbers

ChEMBL 28

- ▶ 1,268,784 unique values for  $h_1$  with  $h_1 = \text{VXL4K9UW2}$  comprising of 537 structures (peptide)



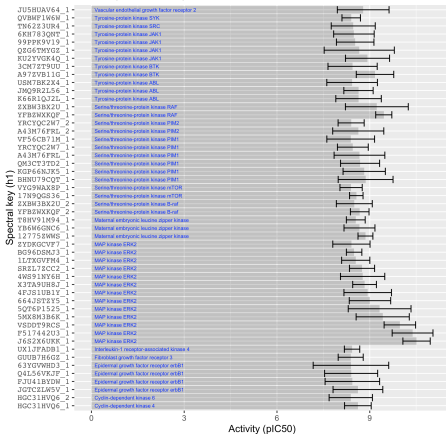
- ▶ 1,727,459 unique values for  $h_2$  with  $h_2 = 9154K6U6NH$  comprising of 515 structures



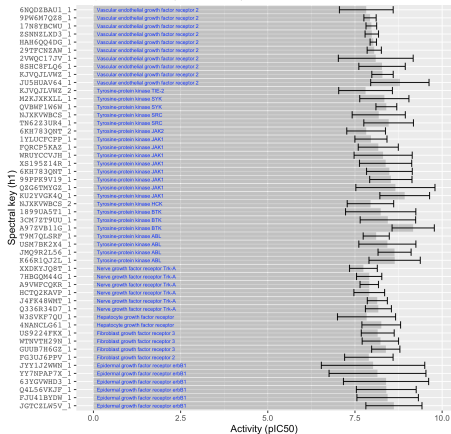
# Do similar hash keys have similar biological activity?

Kinase ( $h_1$ )

Kinase (ChEMBL 28)

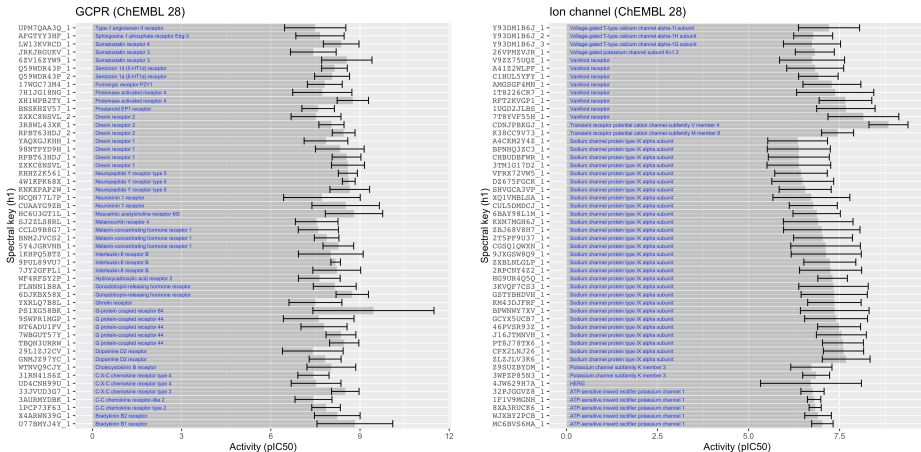


TK protein kinase (ChEMBL 28)



# Do similar hash keys have similar biological activity?

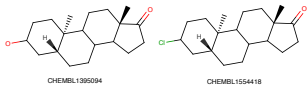
## GPCR & Ion channel ( $h_1$ )



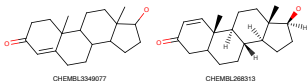
# What is “similarity”?

The problem

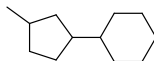
What should the similarity be between the following structures?



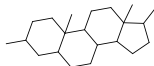
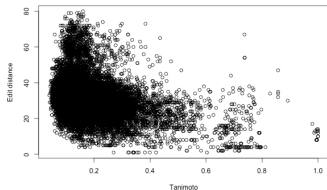
Tanimoto = 0.55, InChI edit distance = 8  
What about this pair?



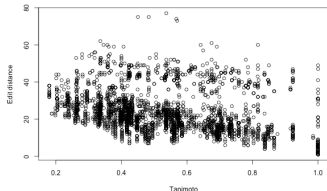
Tanimoto = 1, InChI edit distance = 47  
*Limitations of the Tanimoto metric are well-recognized within the cheminformatics community, but what about the InChI edit distance?*



Pairwise for h1=WU7MZ9LTZ










Pairwise for h2=3U51T8HJB72X2TJJJC1



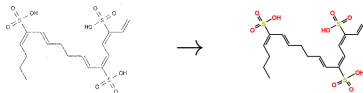
# What is “similarity”?

What's wrong with InChI edit distance? (a diatribe)

- ▶ *Can you translate chemical images to text?*  
<https://www.kaggle.com/c/bms-molecular-translation/>
- ▶ Use InChI edit distance as an evaluation metric
- ▶ Despite our best efforts so far, we're being destroyed by the GPU-bound Python notebook crowds

42	copypaste		12.50	1	5h
43	John Robinson		12.58	1	6h
44	<a href="https://molvec.ncats.io">https://molvec.ncats.io</a>		12.96	5	33m
<b>Your Best Entry</b>					
Your submission scored 12.96, which is an improvement of your previous score of 18.40. Great job! <a href="#">Tweet this!</a>					
45	George #2		13.09	2	3d
46	Binh Nguyen		13.44	8	3d
47	miralisa loval		13.48	7	2d
48	Thomas SELECK		14.00	11	1d

- ▶ Ignorance is bliss (or less is more)



Perfect reconstruction, but the extra E/Zs implied by the drawing cost us 34 edit distance!

# What is “similarity”?

Next step

## Challenge to the InChI community

*Develop a robust distance or similarity metric for InChI that reflects the chemist's intuitions*

- ▶ Graph edit distance based on MCS (e.g., NextMove's smallworld)
- ▶ Edit distance based on graph spectrum



## Code availability

- ▶ Self-contained source code in C at  
`https://github.com/ncats/spectral\_hk`
- ▶ Spectral hash keys generated for ChEMBL 28 are available at  
`mysql -u chembl -h chembl.ncats.io chembl28_ncats`
- ▶ Welcome questions and feedback: `nguyenda@mail.nih.gov`

## Acknowledgements

- ▶ Alexey Zakharov
- ▶ Tyler Peryea (FDA)
- ▶ Lu Chen
- ▶ Ewy Mathé
- ▶ Noel Southall

## Funding

NCATS intramural